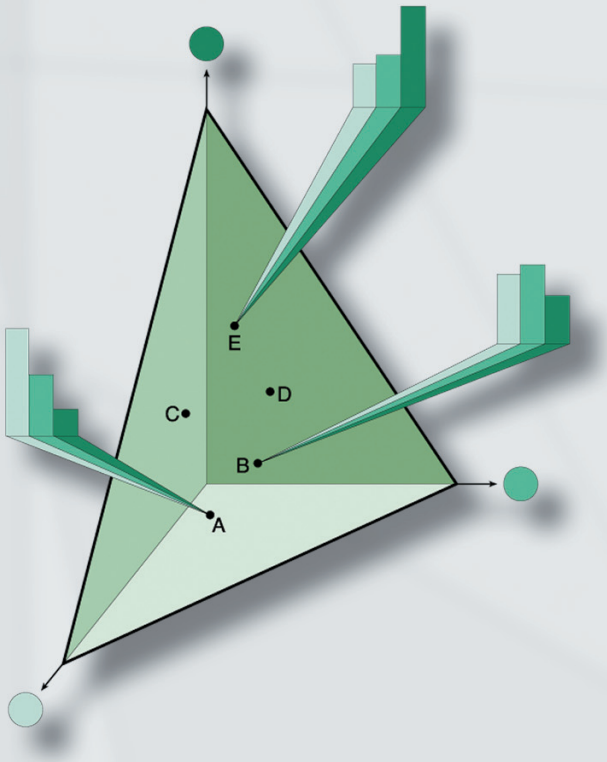
Michael Greenacre

La práctica del análisis de correspondencias



Fundación BBVA

La práctica del análisis de correspondencias

La práctica del análisis de correspondencias

Michael Greenacre

Fundación BBVA

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

Primera edición: julio 2008

Edición y producción: Rubes Editorial

ISBN: 978-84-96515-71-0
Depósito legal: B-35631-2008

Impreso en España - *Printed in Spain*

Impreso por Valant 2003
sobre papel elaborado según las más exigentes normas ambientales europeas.

A Françoise, Karolien y Gloudina

Índice

Apéndices

Prólogo

Este libro es una edición revisada y ampliada de *Correspondence Analysis in Practice*, publicado por primera vez en 1993. Creo que hoy sigue siendo válido lo que escribí en el prólogo de la primera edición; era lo siguiente:

Extracto del prólogo de la primera edición de *Correspondence Analysis in Practice*

> El análisis de correspondencias es una técnica estadística útil para estudiantes, investigadores y profesionales que trabajan con datos categóricos, por ejemplo, datos obtenidos en encuestas sociales. El método es especialmente eficaz para analizar las tablas de contingencia con datos de frecuencias numéricas, ya que nos proporciona una representación gráfica elegante y simple que permite una rápida interpretación y comprensión de los datos. A pesar de que los orígenes teóricos de esta técnica tienen más de 50 años, fue el matemático y lingüista francés Jean-Paul Benzécri quien, junto con sus colegas y estudiantes, dio un impulso real a las aplicaciones modernas del análisis de correspondencias, a principios de los años sesenta en la Universidad de Rennes, y posteriormente, en el campus Jussieu de la Universidad de París. En los Países Bajos y Japón, Jan de Leeuw y Chikio Hayashi también fueron pioneros en desarrollos paralelos del análisis de correspondencias. Mi implicación con el análisis de correspondencias empezó en 1973 cuando inicié mi doctorado en el laboratorio de análisis de datos de Benzécri en París. La publicación en 1984 de mi primer libro, *Theory and Applications of Correspondence Analysis,* coincidió con el inicio de una amplia difusión del análisis de correspondencias fuera de Francia. En aquel momento tenía la esperanza de que mi libro pudiera servir para que en el futuro se extendiera la aplicación práctica del análisis de correspondencias. Las evoluciones posteriores y la reciente popularidad del método no pudieron haber sido más gratificantes: centenares de investigadores se introdujeron en el mismo y se familiarizaron con la capacidad de éste para hacer comprender a los no especialistas, mediante representaciones gráficas, la información contenida en complejas tablas de datos numéricos. Los investigadores con los que entré en contacto procedían de disciplinas tan diversas como sociología, ecología, paleontología, arqueología, geología, pedagogía, medicina, bioquímica, microbiología, lingüística, investigación de mercados, publicidad, religión, filosofía, arte y música. En 1989, Jay Magidson de Statistical Innovations Inc., me invitó a colaborar con Leo Goodman y Clifford Clogg en la presentación en un curso de dos días en Nueva York titulado *«Correspondence Analysis and Association Models: Geometric Representation and Beyond»*. La mayor parte de los participantes eran profesionales del *marketing* de las principales compañías norteamericanas. Preparé unas notas para este curso que reforzaban la aproximación práctica del análisis de corres-

pondencias, la orientación al usuario. La reacción positiva de la audiencia fue contagiosa y me animó posteriormente a impartir cursos sobre análisis de correspondencias en Sudáfrica, Inglaterra y Alemania. Este libro es fruto de las notas preparadas para esos cursos.

Las conferencias de Colonia y Barcelona

En 1991, el profesor Walter Kristof de la Universidad de Hamburgo me propuso que organizáramos, con el apoyo de Jörg Blasius del Archivo Central para Investigación Social Aplicada *(Zentralarchiv für Empirische Sozialforschung)*, de la Universidad de Colonia, una conferencia sobre análisis de correspondencias. Este evento fue la primera conferencia internacional de este tipo. Atrajo a Colonia a una gran audiencia procedente de Alemania y países europeos vecinos. Este primer encuentro evolucionó hacia una serie de conferencias cuatrianuales, que tuvieron lugar en 1995 y 1999 en Colonia, en 2003 en la Universidad Pompeu Fabra, en Barcelona, y en 2007 en la Universidad Erasmus de Rotterdam. De la conferencia de 1991 surgió la publicación del libro *Correspondence Analysis in the Social Sciences*, mientras que la de 1995 dio lugar a otro libro, *Visualisation of Categorical Data*. Ambas publicaciones recibieron excelentes críticas. En la conferencia de 1999 sobre análisis de datos a gran escala, los participantes tuvieron que presentar análisis de datos del Programa Internacional de Encuestas Sociales *(Internacional Social Survey Programme*, ISSP). Esta reunión interdisciplinaria incluyó presentaciones no sólo sobre los últimos desarrollos metodológicos del análisis de encuestas, sino también sobre temas tan diversos como la religión, el medio ambiente y la desigualdad social. En 2003, en la conferencia de Barcelona, se retomó el tema original —bautizado con nombre propio de mujer (en catalán), CARME, siglas de *Correspondence Analysis and Related, Methods*—, que dio lugar a la creación de la red CARME (www.carme-n.org). Esta conferencia nos llevó a Jörg Blasius y a mí a editar un tercer libro, *Multiple Correspondence Analysis and Related Methods,* publicado por Chapman & Hall en junio de 2006. Como ocurrió con los dos libros anteriores, nuestra idea era editar una obra colectiva con participación de diversos autores. Para ello pedimos la contribución de expertos en el tema y, en nuestra tarea de editores, completamos la obra con una introducción; se relacionaron las distintas contribuciones, se unificó la notación y se compiló una lista común de referencias y el índice. De algún modo, estos libros marcan el desarrollo del tema, al menos en lo que se refiere a las ciencias sociales. Son muy recomendables para cualquier persona interesada en profundizar en el conocimiento del análisis de correspondencias y en los métodos relacionados.

Una segunda edición ampliada

Para mí fue muy gratificante reescribir *Correspondence Analysis in Practice*, 13 años después de la primera edición y después de haber acumulado mucha más experiencia en la investigación social y medioambiental. He impartido cursos en todo el mundo, incluyendo España, Italia, Bélgica, Brasil, Canadá y Australia, al igual que cursos de análisis multivariante para biólogos medioambientales en Noruega, Islandia y España, en los que el análisis de correspondencias era uno de los temas

principales. La experiencia acumulada me ha ido sugiriendo nuevas aplicaciones y renovadas ideas, imprescindibles para actualizar el contenido de la primera edición. Aparte de revisar completamente los capítulos originales y renovar algunos ejemplos, se han añadido cinco nuevos capítulos, sobre las relaciones de transición y de regresión (en los que se exponen los resultados del análisis de correspondencias y donde los datos originales están relacionados mediante funciones lineales), sobre las tablas concatenadas (acerca de cómo podemos analizar conjuntamente varias tablas de contingencia), el análisis de correspondencias de subgrupos (una simple pero efectiva variante del algoritmo del análisis de correspondencias que permite analizar partes seleccionadas de un conjunto de datos), el análisis de tablas cuadradas (en el que se aprende a descomponer las tablas cuadradas —por ejemplo las tablas de movilidad social y las matrices de *brand switching*— en partes que podemos visualizar de forma separada utilizando el análisis de correspondencias) y, por último, el análisis de correspondencias canónico (sobre cómo tener en cuenta las relaciones con variables explicativas externas, una ampliación del análisis de correspondencias muy utilizado en ecología). Después de 33 años de trabajo en este ámbito del conocimiento, puedo aventurar que esta segunda edición contiene casi todo mi saber práctico sobre el tema.

Comparación entre la primera y la segunda edición

En una conferencia a la que asistí en los años ochenta, me obsequiaron con una insignia de solapa, cuya foto se reproduce en esta página. Dicha insignia contiene una bonita y ambigua máxima, que bien podría ser, en todo el mundo, el lema del análisis de correspondencias.

Para ilustrar el significado más obvio de este lema, y para dar un simple ejemplo de análisis de correspondencias, hice un recuento del número de tablas y figuras que contenía cada uno de los capítulos de la primera edición y lo comparé con los capítulos homólogos de la segunda. Para que la comparación entre ambas ediciones fuera válida, expresé estos resultados con relación al número total de páginas de cada capítulo. A continuación llevé a cabo una variante de la técnica llamada

análisis de correspondencias de subgrupos (descrita en el capítulo 21), que me permitió dibujar el mapa que mostramos en la imagen 1 —estas ideas se comprenderán mejor una vez leído el libro, sin embargo, por el momento consideremos este ejemplo como una especie de diagrama de dispersión—. Los dos vectores, *figuras* y *tablas*, apuntan hacia la derecha, por tanto, los capítulos situados a la derecha tienen porcentajes de figuras y tablas superiores a la media. Por ejemplo, los capítulos 9 y 11 de la segunda edición tienen los mayores porcentajes de figuras y tablas, respectivamente. Unas tres cuartas partes de los capítulos de la segunda edición que hemos considerado para el análisis, se hallan situados a la derecha del centro del mapa (el centro representa la media), mientras que solamente un cuarto de los capítulos de la primera edición se hallan situados a la derecha. Este resultado demuestra un incremento sustancial de las figuras y tablas en la segunda edición. En realidad, en ésta, han aumentado sustancialmente la información y las aplicaciones prácticas del libro. La primera edición tenía un cuerpo central (sin incluir los apéndices) de 177 páginas —en la segunda edición en inglés éste ha aumentado a 200; un incremento del 13% que ha consistido en un incremento del 25% del número de capítulos y de temas, un incremento del 18% en el número de conjuntos de datos, un incremento del 44% en el número de figuras y del 63% en el de tablas.

Formato de la segunda edición

Este libro presenta algunas innovaciones que merecen mencionarse. Es nuestra intención que esta edición tenga un enfoque todavía más didáctico que la primera: así cada capítulo representa una cantidad limitada de información con el objetivo

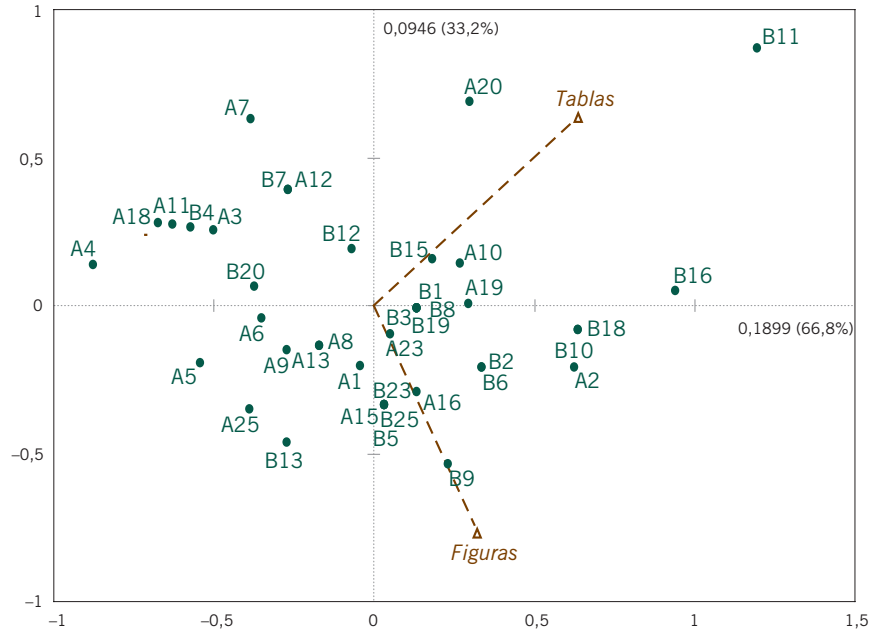


Imagen 1: *Mapa del análisis de correspondencias de subgrupos de los porcentajes de tablas y figuras de los 20 capítulos de la primera edición (simbolizado por* A*) y de sus homólogos de la segunda edición (simbolizados por* B*), que hemos presentado como un* biplot *estándar. Los números de los capítulos de la segunda edición corresponden a los de sus homólogos de la primera edición*

de facilitar la lectura y su uso en la enseñanza. No había una mejor manera de conseguirlo que limitar la extensión de los capítulos, por lo que cada uno de ellos [en la versión en inglés] tiene exactamente ocho páginas. [Dicha extensión se ha visto ampliada a nueve páginas en la versión en castellano.] Era uno de los aspectos más interesantes del proyecto, que además fue evolucionando de forma natural a medida que escribía el libro. Uno de mis colegas comparó la labor didáctica con la poética: el símil de escribir sonetos de 14 versos con una rima estricta resultó ser válido en mi caso: definitivamente, el formato contribuyó al proceso creativo. Otra innovación ha sido el amplio uso de apostillas o notas en los márgenes, cuya función de encabezamiento de párrafo se suma a la de resumir al inicio de cada capítulo, los principales contenidos del mismo. También se ha optado por ubicar las leyendas de figuras y tablas en los márgenes, de manera que la información que aporten pueda ser más extensa y visual que con una clásica disposición al pie. Por último, cada capítulo finaliza con un resumen, en el que se esquematizan las ideas clave.

Apéndice teórico (A)

Igual que en la primera edición, el libro se orienta hacia la práctica del análisis de correspondencias, por lo que hemos reunido en un apéndice final los temas más técnicos, así como los aspectos más matemáticos. El apéndice teórico (A) es más extenso que el de la primera edición. Incluye teoría adicional sobre los nuevos temas del libro como, por ejemplo, el análisis de correspondencias canónico.

Apéndice de cálculo (B)

Una de las principales características de esta edición, que claramente la distingue de la original, y que está en total sintonía con la presente era digital, es la extensión del apéndice de cálculo (B), en el que utilizamos el programa R de dominio público, que se ha convertido de facto en un procedimiento de cálculo estadístico estándar. Casi todos los análisis contenidos en el libro están reflejados en este apéndice. Para ello, se presentan las instrucciones en R necesarias para obtener los correspondientes resultados. Además, se describen tres tecnologías distintas utilizadas para crear las representaciones gráficas del libro. Aunque los mapas puedan parecer simples, obtenerlos no ha sido en absoluto un ejercicio trivial.

Apéndice bibliográfico, glosario y epílogo

En los 25 capítulos del libro no se facilitan referencias bibliográficas. Sin embargo, se incluye un capítulo de bibliografía, relativamente breve, para orientar al lector hacia lecturas adicionales que contienen reseñas bibliográficas mucho más extensas. El libro acaba con un glosario de los términos más importantes y con un epílogo con reflexiones finales.

Agradecimientos

La primera edición de este libro la escribí en Sudáfrica, la presente edición la he escrito en Cataluña, España. Mucha gente, y muchas instituciones, han contribuido, de una manera u otra, a este proyecto. En primer lugar, y de forma principal, quiero expresar mi agradecimiento a Rafael Pardo, director de la Fundación BBVA, sin cuyo apoyo, tanto desde el punto de vista humano como económico,

no hubiera podido dejar mis tareas como profesor durante seis meses y, honestamente, puedo afirmar que la segunda edición en inglés de este libro no hubiera visto la luz. La Fundación BBVA, además, hace posible la publicación de la versión en español del libro. Asimismo deseo expresar mi agradecimiento a la Universidad Pompeu Fabra, en Barcelona, en la que trabajo desde 1994, y muy especialmente a Xavier Calsamiglia, director del Departamento de Economía y Empresa, y a todo el equipo humano del departamento, por haberme dado la libertad de dedicar tanto tiempo a este proyecto.

Quiero agradecer a todos mis amigos y colegas de todo el mundo su apoyo moral e intelectual, especialmente a Zerrin Asan, Jörg Blasius, John Gower, Carles Cuadras, Trevor Hastie, Michael Browne, Victor Thiessen, Karl Jöreskog, Lesley Andres, John Aitchison, Paul Lewi, Patrick Groenen, Pieter Kroonenberg, Ludovic Lebart, Michael Friendly, Antoine de Falguerolles, Salve Dahle, Stig Falk-Petersen, Raul Primicerio, Johs Hjellbrekke, Tom Backer Johnsen, Tor Korneliussen, Ümit Senesen, Brian Monteith, Ken Reed, Gillian Heller, Antonella Curci, Gianna Mastrorilli, Paola Bordandini, Walter Zucchini, Oleg Nenadić, Thierry Fahmy, Tamara Djermanovic, Volker Hooyberg, Gurdeep Stephens, Rita Lugli y Danilo Guaitoli, así como a toda la comunidad de Gréixer —¡todos habéis tenido un papel en esta historia!—. Un agradecimiento especial para Jörg Blasius por una minuciosa corrección del manuscrito original [en inglés] y a Oleg Nenadić por su colaboración en la preparación del paquete **ca** de R. Igual que en la primera edición, he dedicado este libro a mis tres hijas, que nunca dejan de asombrarme por su alegría, sentido del humor y creatividad. También quiero agradecer al editor Rob Calver de Chapman & Hall/CRC Press su confianza y su constante cooperación, lo que ha permitido que la segunda edición en inglés sea una realidad.

Agradecimientos a esta edición

Con respecto a la edición que el lector tiene en sus manos, deseo agradecer la colaboración de Jordi Comas en la traducción de la obra y la experta revisión realizada por Carles Cuadras. Asimismo quiero dar las gracias al equipo humano de Rubes Editorial, en especial a Imma Rullo, Núria Gibert y Jaume Estruch. Las aportaciones de todos ellos han contribuido a dar cuerpo y alma a mi libro.

Por último, deseo expresar, una vez más, mi gratitud al profesor Rafael Pardo, y a Cathrin Scupin, directora editorial de la Fundación BBVA, por el inestimable respaldo que dicha Fundación me ha brindado. Su apoyo convierte este proyecto en una realidad al alcance del público hispanoparlante.

Michael Greenacre
Barcelona, junio de 2008

CAPÍTULO 1

Diagramas de dispersión y mapas

El análisis de correspondencias es un método de análisis de datos que representa gráficamente tablas de datos. El análisis de correspondencias es una generalización de una representación gráfica con la que todos estamos familiarizados, el *diagrama de dispersión*. Un diagrama de dispersión representa los datos en forma de puntos con relación a dos ejes de coordenadas perpendiculares: el eje horizontal, eje de las *x*, y el eje vertical, eje de las *y*. Para introducirnos poco a poco en el análisis de correspondencias, es conveniente que reflexionemos sobre lo que entendemos por diagrama de dispersión y sobre cómo interpretamos los datos que éste representa gráficamente. Haremos énfasis en cómo interpretar las distancias entre puntos y en averiguar cuándo podemos considerar que los diagramas de dispersión son *mapas de datos*.

Contenido

Conjunto de datos 1: mis viajes en 2005

A finales de 2005, cuando empecé a escribir este libro, reflexioné sobre los viajes que durante ese año había hecho a tres de mis países favoritos: Noruega, Canadá y Grecia. Según mi diario pasé 18 días en Noruega, 15 días en Canadá y 29 días en Grecia. Aparte de estas visitas, también hice algunos viajes cortos a Francia y a Alemania, en total 24 días. Podemos representar esta descripción numérica del tiempo que estuve de viaje en gráficos como los de la imagen 1.1. Este ejemplo, aparentemente trivial, esconde algunos conceptos importantes para la interpretación de gráficos en los que representamos los datos con relación a dos ejes de coordenadas, y que eventualmente nos pueden ayudar a comprender el análisis de correspondencias. Vamos a revisar estos conceptos uno a uno.

Variables continuas

El eje vertical situado a la izquierda, que hemos etiquetado como Días, es una escala con información numérica de una variable *continua*. La escala de este eje indica claramente el número de días que pasé en algunos países extranjeros. Hemos ordenado los valores numéricos desde 0 días, en la parte inferior de la escala, hasta 30 días en la parte superior de la misma. En el diagrama de barras situado a la derecha de la imagen 1.1, mostramos una representación gráfica muy habitual de datos, en la cual la longitud de las barras es proporcional a los valores de la variable. Hemos redondeado el tiempo que pasé en cada país a número de días, sin embargo, seguimos considerando esta variable como continua, ya que el tiempo es esencialmente una variable continua.

Expresión de los datos en valores relativos

El eje vertical situado a la derecha de los dos gráficos de la imagen 1.1 expresa el número de días de viaje en cada país, como porcentaje, con relación al total de mis 86 días de viaje. Por ejemplo, 18 días en Noruega corresponde al 21% del tiempo total. El total de 86 días es la *base* con relación a la cual expresamos los valores relativos de los datos. En este caso tenemos un solo conjunto de datos, y en consecuencia sólo una base. En estos dos gráficos podemos representar, en el mismo gráfico, la escala absoluta original de la izquierda y la escala de valores relativos de la derecha.

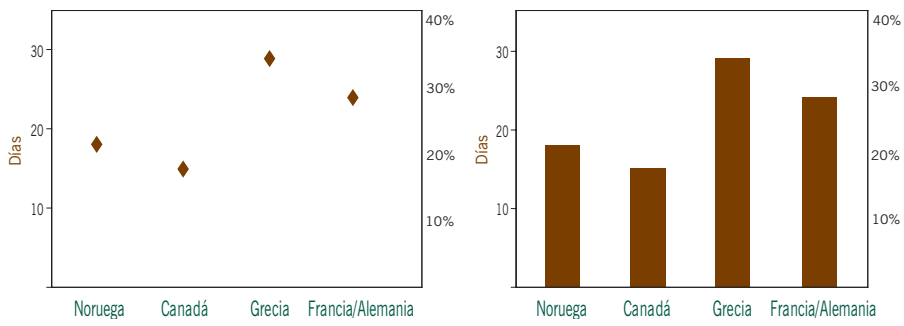


Imagen 1.1:
Gráficos sobre el número de días que pasé en países extranjeros en 2005, en forma de diagrama de dispersión y de diagrama de barras. A la derecha de cada gráfico, el eje vertical expresa el número de días en porcentaje con relación al total de 86 días de viaje

Variables categóricas

A diferencia del eje vertical, eje *y*, el eje horizontal, eje *x*, corresponde claramente a una variable no numérica. En este eje, los cuatro puntos son sólo posiciones en las que hemos situado las etiquetas que indican el país visitado. La escala horizontal representa una variable *categórica*. Hay dos características de este eje horizontal que no tienen significado sustantivo alguno en el gráfico: la ordenación de las categorías y la distancia entre ellas.

Ordenación de las categorías

En primer lugar, no hay ninguna razón de peso por la cual hayamos situado a Noruega en primer lugar, a Canadá en segundo y a Grecia en tercer lugar; quizás el hecho de que visité estos países por este orden. Como la etiqueta Francia/Alemania indica un conjunto de viajes cortos que realicé en distintos momentos del año, hemos situado esta etiqueta después de las otras. Sin embargo, en este tipo de representaciones gráficas en las que el orden es irrelevante, siempre es bueno reordenar las categorías de manera que tengan algún significado sustantivo, por ejemplo, los valores de la variable. Así, podríamos ordenar los países en orden descendiente de acuerdo con el tiempo que pasé en cada país. En tal caso habríamos situado los países en el siguiente orden: Grecia, Francia/Alemania, Noruega y Canadá. Esta sencilla reordenación facilita la interpretación de los datos, especialmente cuando tenemos muchos. Por ejemplo, si hubiera visitado 20 países distintos, la ordenación contendría información relevante que no obtendríamos de forma rápida a partir de la ordenación original.

Distancias entre las categorías

En segundo lugar, no existe razón alguna por la cual hayamos situado los cuatro puntos a intervalos iguales en el eje de las *y*. Asimismo, no existe tampoco razón por la cual hayamos de situarlos a intervalos distintos; en realidad los hemos situado a intervalos iguales por conveniencia y estética. Cuando utilicemos el análisis de correspondencias, veremos que existen distintas maneras de definir intervalos entre las categorías de las variables como la que acabamos de comentar. Es más, presentaremos el análisis de correspondencias como un procedimiento para la cuantificación de las categorías de una variable y, así, tanto las distancias entre categorías como su ordenación tendrán un significado importante.

Interpretación de las distancias en los diagramas de dispersión

En el eje horizontal del gráfico de la izquierda de la imagen 1.1, tanto la ordenación de los países como la separación entre éstos son arbitrarias, por tanto, no tiene ningún sentido que midamos e interpretemos las distancias entre los puntos mostrados en el gráfico de la izquierda. Dada la naturaleza numérica del eje vertical que indica frecuencia (o frecuencia relativa), las únicas medidas de distancia que tienen sentido son estrictamente las distancias en dirección vertical.

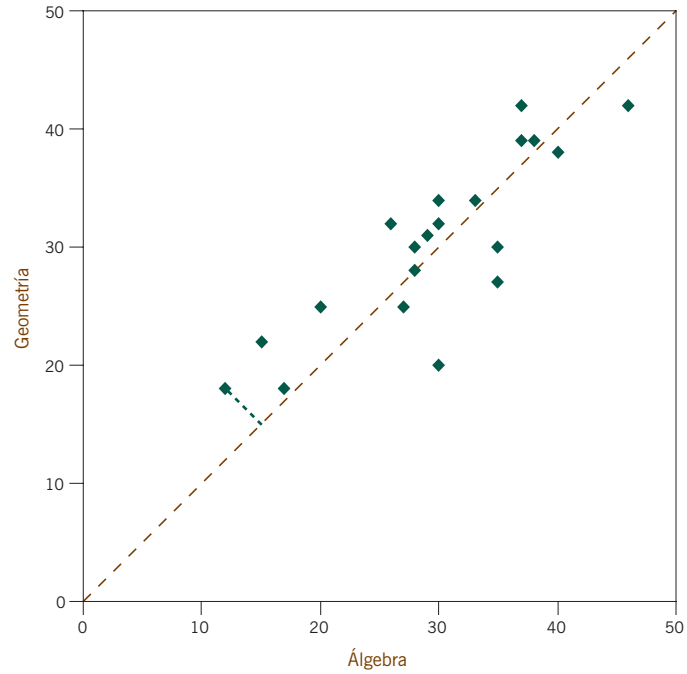


Imagen 1.2:
Diagrama de dispersión de las calificaciones de 20 estudiantes en dos materias (álgebra y geometría) en un examen de matemáticas. Los puntos tienen propiedades especiales. Así podemos obtener la calificación total de los estudiantes proyectando los puntos perpendicularmente sobre la bisectriz que hemos calibrado de 0 (abajo a la izquierda) a 100 (arriba a la derecha)

Los diagramas de dispersión como mapas

En algunos casos especiales, las dos variables que definen a los ejes de los diagramas de dispersión tienen la misma naturaleza numérica y escalas similares. Por ejemplo, supongamos que 20 estudiantes han realizado un examen de matemáticas que consta de dos partes, álgebra y geometría. Supongamos que cada parte representa el 50% de la nota final. En la imagen 1.2, hemos representado gráficamente los pares de calificaciones de los estudiantes. Es importante que los dos ejes, que representan las respectivas calificaciones, tengan escalas con unidades de la misma longitud. Dada la naturaleza similar de las dos variables y de sus dos escalas, en esta representación gráfica podemos medir distancias en cualquier dirección; no solamente horizontal o verticalmente —igual que en un mapa en el que podemos medir distancias entre poblaciones—. Dos puntos que se hallen cerca tendrán calificaciones similares. Por tanto, tiene sentido que nos fijemos en la forma de la distribución de los puntos y, en particular, remarcar que hay un pequeño grupo de cuatro estudiantes con calificaciones elevadas y sólo un estudiante con calificaciones muy elevadas. Podemos considerar la imagen 1.2 un *mapa*, ya que las posiciones de los estudiantes vienen definidas por posiciones bidimensionales, de la misma manera que, en una región, las localizaciones geográficas vienen definidas por la longitud y la latitud.

Calibración de una dirección en un mapa

Los mapas tienen interesantes propiedades geométricas. Por ejemplo, en la imagen 1.2, la bisectriz, que hemos representado como una línea discontinua, define un eje que expresa las calificaciones finales de los estudiantes, combi-

nando las calificaciones de álgebra y de geometría. Si calibramos este eje de 0 (abajo izquierda) hasta 100 (arriba a la derecha), podemos leer las calificaciones finales de los estudiantes en el mapa, proyectando de forma perpendicular sobre el mencionado eje los puntos que representan sus calificaciones. En la representación gráfica podemos ver un ejemplo para un estudiante que obtuvo 12 puntos sobre 50 en álgebra y 18 sobre 50 en geometría. A la proyección de este punto sobre la bisectriz, de coordenadas 15 y 15, le corresponde una calificación final de 30.

Transformación de la información en la representación gráfica

Los diagramas de dispersión de las imágenes 1.1 y 1.2 son dos maneras distintas de expresar, de forma gráfica, la información numérica contenida en dos tablas que contienen datos sobre viajes y calificaciones, respectivamente. En ambos casos, no hay pérdida de información entre los datos y las representaciones gráficas. Dados los gráficos, es fácil recuperar exactamente la información numérica. Decimos que los diagramas de dispersión o los mapas son «instrumentos de transformación de la información» en los que, en absoluto, se produce un procesado de los datos; simplemente expresamos los datos de forma visual, es decir, se trata de una manifestación alternativa de la misma información.

Variables nominales y variables ordinales

En el ejemplo sobre mis viajes, la variable categórica «país» tiene cuatro categorías, y dado que no existe una ordenación intrínseca de las categorías, llamamos a esta variable *nominal*. En cambio, si podemos ordenar de forma natural las categorías de una variable categórica, llamamos a la variable *ordinal*. Por ejemplo, podemos clasificar los días en tres categorías de acuerdo con el tiempo que dediqué cada día a trabajar: a) menos de una hora («festivos»), b) más de una pero menos de seis horas («medias jornadas») y c) más de seis horas («jornadas completas»). Por tanto, hemos ordenado estas categorías de acuerdo con una variable continua «tiempo diario de trabajo» que hemos dividido en intervalos. Tendremos en cuenta esta ordenación en cualquier representación gráfica de las variables. En muchas encuestas sociales, se dan las respuestas en una escala ordinal. Por ejemplo, una escala ordinal sobre valoración de la importancia: nada importante/algo importante/muy importante. Otro ejemplo típico es la escala de acuerdo/desacuerdo: muy de acuerdo/algo de acuerdo/ni de acuerdo ni en desacuerdo/algo en desacuerdo/muy en desacuerdo. Aquí la posición ordinal de la categoría «ni de acuerdo ni en desacuerdo» puede no estar situada entre «algo de acuerdo» y «algo en desacuerdo», podría ser, por ejemplo, una categoría utilizada por algunos encuestados para expresar que «no sabe» cuando éstos o bien no comprenden la pregunta o bien no tienen una respuesta clara. Veremos este tema más adelante (cap. 21), una vez hayamos desarrollado las herramientas que nos permitan estudiar las asociaciones entre las respuestas en cuestionarios de datos multivariantes.

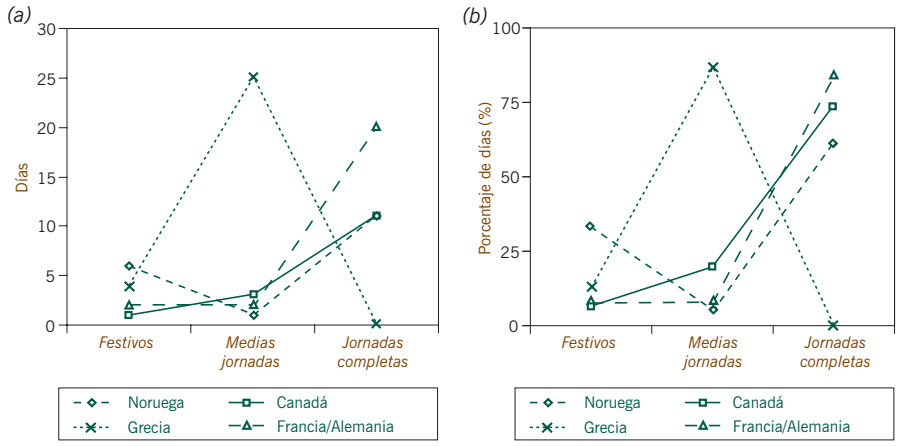
Imagen 1.3: *Frecuencias de los tipos de día en los cuatro viajes*

País	*Festivos*	*Medias jornadas*	*Jornadas completas*	*Total*
Noruega	6	1	11	*18*
Canadá	1	3	11	*15*
Grecia	4	25	0	*29*
Francia/Alemania	2	2	20	*24*
Total	*13*	*31*	*42*	*86*

Representación gráfica de más de un conjunto de datos

Supongamos que clasificamos mis 86 días de viaje en el extranjero de acuerdo con las categorías, *festivos*, *medias jornadas* y *jornadas completas*. En la imagen 1.3 se muestra una tabla que corresponde a la *clasificación cruzada* de país por tipo de día. Podemos ver esta tabla de dos formas distintas: como un conjunto de filas o como un conjunto de columnas. En este caso, las columnas son conjuntos de frecuencias que caracterizan a los respectivos tipos de día; mientras que las filas caracterizan a los respectivos países. En la figura (a) de la imagen 1.4, se muestra un diagrama de frecuencias de los distintos países (filas), en el que hemos situado el tipo de día (las columnas) en el eje horizontal. Dado que hemos ordenado las categorías de la variable «tipo de día», tiene sentido unir los valores de las categorías de esta variable mediante líneas. Sin embargo, si queremos comparar los países entre sí, hemos de tener en cuenta que el número de días que pasé en cada país no fue el mismo. El número total de días que pasé en cada país nos proporciona una base sobre la que podemos reexpresar los valores de las filas de la imagen 1.3, como porcentajes con relación a estos totales (imagen 1.5). En la representación gráfica de la imagen 1.4(b), hemos visualizado estos porcentajes, y ahora sí podemos comparar los tipos de día de los distintos viajes.

Imagen 1.4: *Diagramas de frecuencias absolutas (a) y de frecuencias relativas (b), expresadas como porcentajes de las filas de la imagen 1.3*



País	*Festivos*	*Medias jornadas*	*Jornadas completas*
Noruega	33%	6%	61%
Canadá	7%	20%	73%
Grecia	14%	86%	0%
Francia/Alemania	8%	8%	83%
Global	*15%*	*36%*	*49%*

Imagen 1.5: *Porcentajes correspondientes a los tipos de día en cada país, así como los porcentajes globales de los países, donde la suma de los valores de las filas es el 100%*

Interpretación de las frecuencias absolutas y de las frecuencias relativas

De estas representaciones gráficas tenemos que extraer una lección fundamental para el análisis de frecuencias de datos. Cada viaje ha implicado un diferente número de días y, por tanto, corresponde a una base distinta sobre la que expresar la frecuencia de los tipos de día. Sólo podemos comparar los 6 *festivos* en Noruega, con los 4 en Grecia, con relación al número total de días que pasé en cada uno de estos países. Como porcentajes, estos valores se transforman en valores muy distintos; 6 de 18 es el 33%, mientras que 4 de 29 es el 14%. La visualización de las frecuencias relativas de la imagen 1.4(b) nos permite una comparación más precisa de cómo pasé mi tiempo en los diferentes países. También podemos expresar las frecuencias «marginales» (18, 15, 29 y 24, de los países y 13, 31, 42 del tipo de día) con relación a sus respectivos totales (por ejemplo, en la última fila de la imagen 1.5 mostramos los porcentajes correspondientes al tipo de día para la combinación de todos los países). Estas frecuencias marginales relativas, también las podíamos haber representado en la imagen 1.4 (b).

Descripción e interpretación de los datos versus modelización e inferencia estadística

Cualquier conclusión que hayamos sacado sobre la posición de los puntos de la imagen 1.4(b) es sólo una interpretación de los datos, no es una afirmación sobre la significación estadística de lo que hemos observado. Estos aspectos estadísticos de las representaciones gráficas, los veremos solamente al final del libro (cap. 25). Por tanto, en la mayor parte del libro nos concentraremos en la descripción y en la interpretación de los datos. La deducción de que, en proporción, pasé más días festivos en Noruega que en ningún otro país es ciertamente verdadera, lo podemos ver en la imagen 1.4(b). Sin embargo, analizar si este fenómeno es estadísticamente comparable con un modelo o con una hipótesis sobre mi comportamiento que, por ejemplo, postule que la proporción de festivos fue la misma en todos mis viajes, es un tema completamente distinto. Gran parte de la metodología estadística existente se concentra en saber si los datos se ajustan, o se pueden comparar, con un determinado modelo teórico o con una hipótesis preconizada. Se dedica poca atención a desarrollar procedimientos para describir datos, para interpretarlos o para generalizar hipótesis. Un ejemplo típico, en ciencias sociales, es la utilización omnipresente del estadístico ji-cuadrado para contrastar asociaciones en tablas de contingencia. A menudo se hallan asociaciones estadísticamente significativas, pero en cambio no existen herramientas

sencillas para detectar qué partes de la tabla son las responsables de esta asociación. El análisis de correspondencias es una herramienta que puede contribuir a rellenar este vacío. Permite al analista visualizar las asociaciones existentes en los datos, y en consecuencia le permite formular hipótesis que éste puede contrastar en una etapa más avanzada de su investigación. En la mayor parte de las situaciones, podemos describir, interpretar y modelizar los datos. De todas formas, existen situaciones en las que la descripción y la interpretación de los datos tiene, por sí misma, una importancia capital, por ejemplo, cuando los datos representan a la totalidad de la población de interés.

Conjuntos de datos grandes

A medida que las tablas de datos aumentan de tamaño, debido al excesivo número de puntos, se hace difícil representar éstos de forma simple, como hemos hecho, por ejemplo, en la imagen 1.4. Supongamos que durante un año hubiera visitado 20 países, al clasificar el tiempo pasado en cada uno de ellos, hubiese obtenido una tabla de contingencia con muchas más filas. También podría haber registrado otros datos, como por ejemplo la meteorología de cada día («buen tiempo», «parcialmente nublado» o «lluvioso»), con el objetivo de estudiar posibles relaciones con el tipo de día. Tendría, pues, una tabla de datos con muchas más columnas y muchas más filas. Representar, de la misma manera como hemos hecho en la imagen 1.4, a los 20 conjuntos de puntos clasificados en muchas más categorías podría llevarnos a una gran confusión entre puntos y etiquetas. Resultaría absolutamente imposible identificar pauta alguna. Por tanto, en estas situaciones para resaltar las características esenciales de esos datos, tendríamos que buscar una alternativa a los diagramas de dispersión, el instrumento para la descripción de datos que hemos utilizado hasta ahora. Tal como veremos en el libro, el análisis de correspondencias, un método de representación gráfica de datos igual que los diagramas de dispersión, nos permitirá trabajar fácilmente con conjuntos de datos grandes.

RESUMEN: Diagramas de dispersión y mapas

1. Los diagramas de dispersión representan gráficamente dos variables con relación a un eje horizontal y un eje vertical, el eje *x* y el eje *y*, respectivamente.
2. A menudo, la naturaleza de la variable *x* es completamente distinta a la de la variable *y*, de manera que solamente podemos interpretar distancias en la dirección de unos de los dos ejes, de acuerdo con una determinada escala de medida con la que hayamos calibrado el eje. En estas situaciones, no tiene sentido medir o interpretar distancias en cualquier otra dirección del gráfico.
3. El algunos casos, las variables *x* e *y* son de naturaleza similar con escalas de medida comparables. En estas situaciones, podemos interpretar las distancias entre los puntos como una medida de la diferencia, o de la disimilitud, entre los puntos representados. En estos casos especiales consideramos que los diagramas de dispersión son *mapas*.

4. Cuando representamos valores positivos (en general, en nuestro contexto, frecuencias), estamos interesados tanto en los valores relativos como en los absolutos.
5. Cuanto más complejos sean los datos, menos conveniente será representarlos en forma de diagramas de dispersión.
6. Este libro, más que sobre la modelización de información compleja, trata sobre la descripción y la interpretación de la información.

CAPÍTULO 2

Perfiles y espacio de perfiles

El concepto de *perfil*, un conjunto de frecuencias relativas, es fundamental para el análisis de correspondencias (AC de aquí en adelante). Estos conjuntos de frecuencias relativas, o *vectores*, tienen características geométricas especiales debido a que la suma de sus elementos es 1 (o el 100%). Cuando analicemos una tabla de frecuencias nos podemos fijar en las frecuencias relativas de las filas o en las frecuencias relativas de las columnas, las llamaremos *perfiles fila* y *perfiles columna*, respectivamente. En este capítulo representaremos los perfiles como puntos en un espacio de perfiles. En particular, lo estudiaremos para el caso especial de perfiles con sólo tres elementos.

Contenido

Perfiles

Veamos una vez más los datos de la imagen 1.3, una tabla de frecuencias con cuatro filas (los países) y tres columnas (los tipos de día). En el AC, el concepto de *perfil*, un conjunto de frecuencias divididas por su total, es el primer concepto y más fundamental. En la imagen 2.1 se muestran los perfiles de las filas, que llamaremos perfiles fila, de estos datos. Por ejemplo, el perfil de Noruega es

Imagen 2.1:
Perfiles fila (país): frecuencias relativas de los tipos de día de cada viaje, y perfil fila medio que muestra las frecuencias relativas de todos los viajes conjuntamente

País	*Festivos*	*Medias jornadas*	*Jornadas completas*
Noruega	0,33	0,06	0,61
Canadá	0,07	0,20	0,73
Grecia	0,14	0,86	0,00
Francia/Alemania	0,08	0,08	0,83
Media	*0,15*	*0,36*	*0,49*

[0,33 0,06 0,61], donde 0,33 = 6/18, 0,06 = 1/18, 0,61 = 11/18. Decimos que se trata del «perfil de Noruega respecto al tipo de día». También podemos expresar el perfil como porcentaje; por ejemplo, en este caso es [33% 6% 61%], como vimos en la imagen 1.5. De forma similar, el perfil de Canadá respecto al tipo de día es [0,07 0,20 0,73], igual que ocurre con Noruega, la mayor concentración se produce en la categoría *jornadas completas*. En cambio, Grecia tiene un perfil de [0,14 0,86 0,00], la mayor concentración se produce en la categoría *medias jornadas*, y así sucesivamente. Estos son los valores que hemos representado en la imagen 1.4(b) de la página 20.

Perfil medio

En la imagen 2.1, además de los perfiles de los cuatro países, tenemos una fila adicional que hemos llamado *media*. Se trata del perfil de la fila final [13 31 42] de la imagen 1.3, que contiene la suma de las columnas de la tabla; es el perfil de todos los viajes considerados conjuntamente. En el capítulo 3 se explica más específicamente porqué lo llamamos *perfil medio*. Por el momento, basta con darnos cuenta de que de los 86 días de viaje, sin tener en cuenta el país visitado, el 15% fueron *festivos*, el 36% *medias jornadas* y el 49% *jornadas completas* de trabajo. Cuando comparemos perfiles, podemos comparar el perfil de un país con el de otro, pero también el perfil de un país con el perfil medio. Por ejemplo, en los datos de la imagen 2.1, vemos que los países con perfiles más parecidos son los de Canadá y de Francia/Alemania. Podemos ver que, comparados con el perfil medio, ambos tienen un mayor porcentaje de *jornadas completas* y, en cambio, están por debajo de la media por lo que respecta a *festivos* y *medias jornadas*.

Perfiles fila y perfiles columna

Hasta ahora nos hemos fijado en los perfiles fila con el objetivo de comparar los diferentes países entre sí. Sin embargo, también podemos considerar la imagen 1.3 como un conjunto de columnas y comparar cómo se distribuyen los diferentes tipos de día con relación a los países. En la imagen 2.2 mostramos los perfiles de las columnas, que llamaremos perfiles columna, así como el perfil columna medio. Por ejemplo, de los 13 *festivos*, el 46% fueron en Noruega, el 8% en Canadá, el 31% en Grecia y el 15% en Francia/Alemania. Podemos comparar los valores

País	*Festivos*	*Medias jornadas*	*Jornadas completas*	*Media*
Noruega	0,46	0,03	0,26	*0,21*
Canadá	0,08	0,10	0,26	*0,17*
Grecia	0,31	0,81	0,00	*0,34*
Francia/Alemania	0,15	0,06	0,48	*0,28*

Imagen 2.2:
Los perfiles de los tipos de día con relación a los países, y el perfil columna medio

de los perfiles de los tipos de día con los valores del perfil columna medio, para ver si sus valores están por encima o por debajo de los de la media. Así por ejemplo, el 46% de los *festivos* los pasé en Noruega, no obstante, el número de días que pasé en Noruega fue sólo el 21% del total de los 86 días de viaje, un gran número de *festivos* comparado con la media.

Tratamiento simétrico de las filas y de las columnas

Veamos otra vez la proporción de 0,46 (6/13) de *festivos* que pasé en Noruega (imagen 2.2) y comparémosla con la proporción 0,21 (18/86) de todos los días que pasé en este país. Podríamos calcular el cociente 0,46/0,21 = 2,2, y llegar a la conclusión de que *festivos* en Noruega está ligeramente por encima de dos veces la media. Llegamos exactamente a la misma conclusión si hacemos un cálculo similar con los perfiles fila. Así, en la imagen 2.1, podemos ver que la proporción de *festivos* en *Noruega* era de 0,33 (6/18) mientras que, considerando todos los países conjuntamente, la proporción de *festivos* era de 0,15 (13/86). Vemos que 0,33/0,15 es 2,2 veces mayor que la media. La misma proporción que hemos obtenido a partir de los perfiles columna (llamamos a esta proporción *cociente de contingencia*, lo veremos de nuevo en los próximos capítulos). Tanto si razonamos a partir de los perfiles fila como a partir de los perfiles columna, llegamos a la misma conclusión. En el capítulo 8 mostraremos que el AC analiza de forma similar las filas y las columnas de una tabla de contingencia; también podríamos decir que trata filas y columnas de forma *simétrica*.

Consideración asimétrica de una tabla de datos

A pesar de ello, en la práctica, a menudo vemos y consideramos las tablas de datos de forma no simétrica, o *asimétrica*, ya sea como un conjunto de filas o un conjunto de columnas. Por ejemplo, dado que cada fila de la imagen 1.3 constituye un viaje distinto, podría ser más natural ver la tabla como formada por filas, como en la imagen 2.1. La decisión sobre cuál es la forma más adecuada de analizar una tabla depende de la naturaleza de los datos y de los objetivos de investigación; a menudo no es una decisión consciente. Para conocer la decisión adoptada por un determinado investigador, tenemos que fijarnos en cómo éste interpreta los datos, si se refiere a porcentajes en filas o a porcentajes en columnas. Sin embargo, cualquiera que sea la decisión, los resultados del AC no dependen de esta elección.

Representación de los perfiles en el espacio de perfiles

Consideremos ahora una forma completamente distinta de representar los cuatro perfiles fila y el perfil fila medio de la imagen 2.1. A diferencia de la imagen 1.4(b), en la que el eje horizontal identificaba el tipo de día y el eje vertical representaba los porcentajes de días, ahora proponemos utilizar tres ejes, que correspondan a los tres tipos de día, a modo de diagrama de dispersión tridimensional. Imaginar tres ejes perpendiculares no es difícil: basta con mirar el suelo de la habitación en la que trabajamos y buscar una esquina despejada, veremos tres ejes como los que mostramos en la imagen 2.3. En cada uno de estos tres ejes de la habitación podríamos situar a uno de los tres elementos del perfil. Podemos considerar estos tres elementos como las coordenadas de un punto que represente todo el perfil —situación bastante distinta de la del gráfico de la imagen 1.4(b), en la que teníamos un punto distinto para cada uno de los elementos del perfil—. Etiquetamos los tres ejes como *festivos*, *medias jornadas* y *jornadas completas* y los calibramos de 0 a 1. Ahora, representar los cuatro perfiles es un ejercicio sencillo. El perfil de Noruega [0,33 0,06 0,61] (imagen 2.1), se halla a 0,33 unidades en el eje *festivos*, a 0,06 unidades en el eje *medias jornadas* y a 0,61 unidades en el eje *jornadas completas*. Otro ejemplo es el perfil de Grecia [0,14 0,86 0,00], que toma el valor cero en la dirección *jornadas completas*, por tanto su posición se halla en la «pared» de la izquierda, definida por las coordenadas 0,14 y 0,86 en los ejes *festivos* y *medias jornadas*, respectivamente, que delimitan la «pared». A los restantes perfiles los podemos representar de la misma manera en el espacio tridimensional, incluyendo el perfil fila medio [0,15 0,36 0,49].

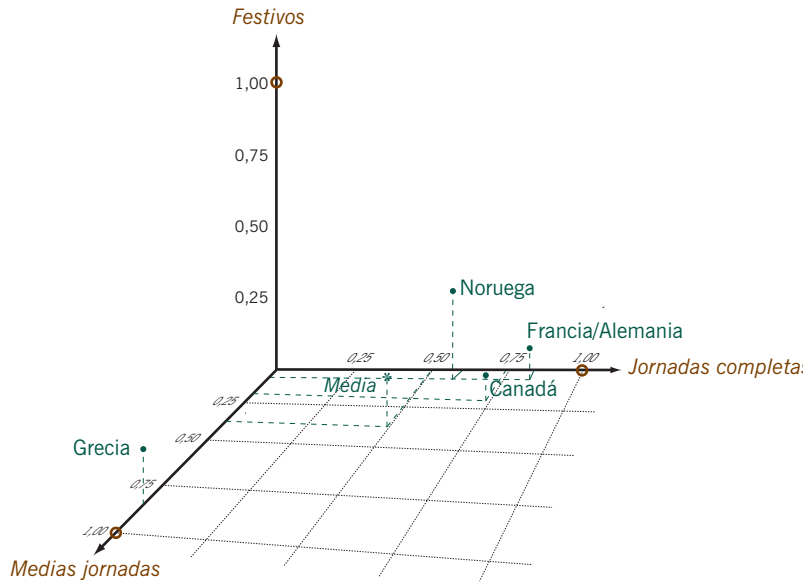


Imagen 2.3:
Posiciones de los cuatro perfiles fila (•) de la imagen 2.1 así como la del perfil fila medio (*) en un espacio tridimensional, que hemos representado como la esquina de una habitación con baldosas en el suelo. Así, por ejemplo, Noruega *toma el valor 0,06 en el eje medias jornadas, 0,61 en el de jornadas completas y 0,33 en la dirección vertical correspondiente al eje festivos. En cada eje hemos representado los puntos correspondientes a la unidad (vértices) por círculos huecos (○)*

Los vértices definen los extremos del espacio de perfiles

Con un poco de imaginación podríamos ver, como hemos representado gráficamente en la imagen 2.4, que los perfiles de la imagen 2.3 se hallan todos exactamente en un plano definido por un triángulo equilátero que une los tres *puntos unidad* [1 0 0], [0 1 0] y [0 0 1], situados en sus respectivos ejes. Llamaremos vértices a los tres extremos de este triángulo equilátero. Los vértices coinciden con los perfiles extremos, es decir perfiles totalmente concentrados en un solo tipo de día. Por ejemplo, el vértice [1 0 0] corresponde a un viaje con sólo *festivos* (desafortunadamente, ficticio en mi caso). De la misma manera, el vértice [0 0 1] corresponde a un viaje con sólo *jornadas completas* de trabajo.

El sistema de coordenadas triangular (o ternario)

Visto que, en realidad, en el espacio tridimensional, todos los perfiles se hallan en un triángulo (bidimensional), podemos situar este triángulo en un plano, como en la imagen 2.5. ¡Mirar los perfiles en un plano es más cómodo que intentar imaginar sus posiciones tridimensionales en la esquina de una habitación! A este tipo especial de representaciones las llamaremos *sistemas de coordenadas triangular* (o *ternario*). Las podemos utilizar siempre que tengamos datos compuestos de tres elementos que sumados den 1, como es el caso de los perfiles fila de nuestro ejemplo. Este tipo de datos es frecuente, por ejemplo, en geología y en química, disciplinas en las que se obtienen muestras que se caracterizan por la pro-

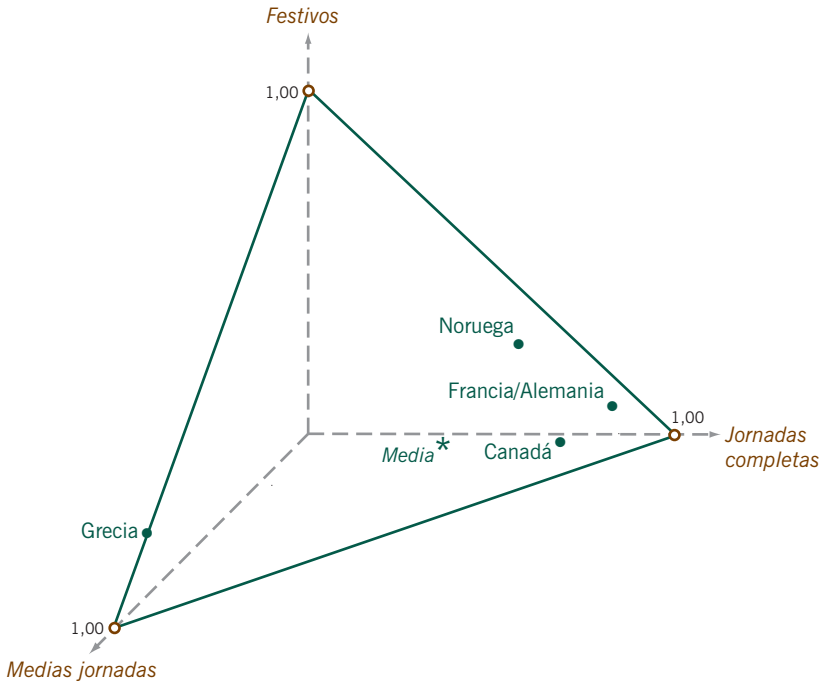
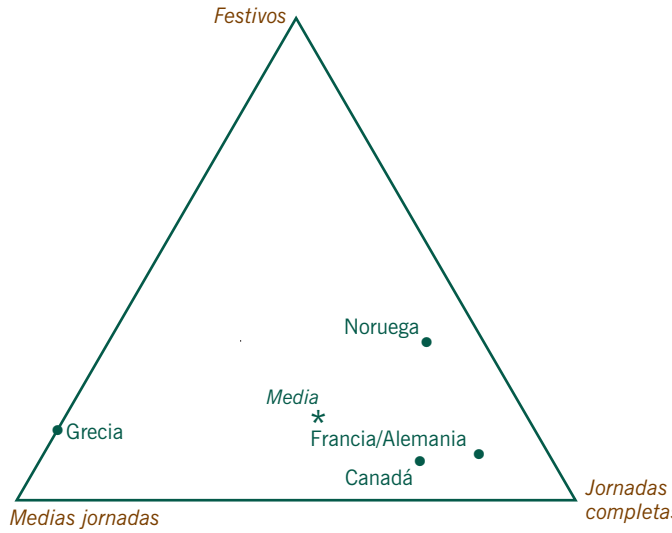


Imagen 2.4:
Los perfiles de la imagen 2.3 se hallan en un triángulo equilátero formado uniendo los vértices del espacio de perfiles. Por tanto, los perfiles tridimensionales son, en realidad, bidimensionales. El perfil de Grecia *se halla en el borde del triángulo debido que su valor para jornadas completas es cero*

Imagen 2.5:
El triángulo de la imagen 2.4 con los perfiles fila (países). Las tres esquinas, o vértices, del triángulo representan las columnas (tipo de día)



porción de tres elementos, en peso o volumen, y que por tanto podemos representar gráficamente como un solo punto en un sistema de coordenadas triangular (o ternario).

Situación de los puntos en un sistema de coordenadas triangular

Supongamos que tenemos un triángulo equilátero en blanco y los valores de los perfiles, ¿cómo podemos situar los perfiles sin tener que pasar por el espacio tridimensional que mostramos en las imágenes 2.3 y 2.4? En el sistema de coordenadas triangular, los lados del triángulo definen tres ejes. Suponemos que cada lado tiene una longitud igual a 1 y que los calibramos de forma lineal de 0 a 1. A continuación situamos los valores de los perfiles en sus respectivos ejes. Por ejemplo, como vemos en la imagen 2.5, para situar a Noruega tomamos el valor 0,33 en el eje *festivos*, 0,06 en el eje *medias jornadas* y 0,61 en el eje *jornadas completas*. A partir de estos valores de los perfiles, podemos trazar líneas paralelas a los lados del triángulo. Estas líneas paralelas confluyen en un punto que define la posición de Noruega (imagen 2.6). En realidad, es suficiente cualquier combinación de dos de las tres coordenadas de los perfiles, para situar los perfiles de la manera indicada. La tercera coordenada es, en realidad, innecesaria. Es otra manera de demostrar que los perfiles son intrínsecamente bidimensionales.

Geometría de los perfiles con más de tres elementos

Sólo podemos utilizar el sistema de coordenadas triangular para perfiles con tres elementos. Sin embargo, como veremos en el próximo capítulo, podemos generalizar fácilmente esta idea a perfiles con cualquier número de elementos. En tal caso llamamos al sistema de coordenadas *sistema de coordenadas baricéntri-*

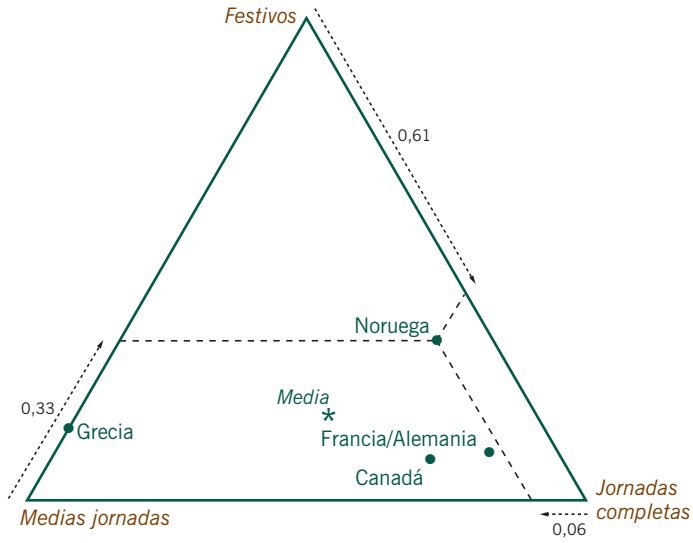


Imagen 2.6:
Noruega
[0,33 0,06 0,61]

co («baricentro» es sinónimo de «media ponderada»). La dimensionalidad de este sistema de coordenadas es siempre igual al número de elementos de los perfiles menos uno. Así, acabamos de ver que los perfiles con tres elementos se hallan exactamente en un espacio de perfiles triangular de dos dimensiones. La dimensionalidad de los perfiles con cuatro elementos es tres: los perfiles se hallan en un tetraedro de cuatro extremos situado en un espacio tridimensional. El triángulo bidimensional y el tetraedro tridimensional son ejemplos de lo que en matemáticas se llama un *símplex regular*. Con el fin de poder percibir un espacio de perfiles tridimensional, en el apéndice de cálculo, indicamos la codificación R que permite visualizar un ejemplo tridimensional. Para perfiles de dimensión superior a tres, necesitaríamos para ser capaces de «ver» el espacio de perfiles con mayores dosis de imaginación. Afortunadamente, como veremos más adelante, el AC nos ayudará mucho a visualizar este tipo de perfiles multidimensionales.

Datos en una escala de razón

Hasta ahora, hemos visto el concepto de perfil en un contexto de datos de frecuencias, el principal tipo de datos que utilizamos en el AC. Sin embargo, el AC lo podemos aplicar a un abanico mucho más amplio de tipos de datos. De hecho lo podemos aplicar siempre que tenga sentido expresar los datos como valores relativos, es decir, que podamos expresar los datos en una *escala de razón*. Por ejemplo, supongamos que tenemos datos sobre los importes de dinero invertidos por diferentes países en distintas áreas de investigación —los valores relativos de interés podrían ser, por ejemplo, los porcentajes invertidos en medio ambiente, en biomedicina, etc.—. Otro ejemplo podrían ser las medidas

morfométricas de organismos vivos, por ejemplo peces, como la longitud total, la anchura, la longitud de las aletas, etc., en centímetros, que podríamos expresar con relación a la suma total. Dicha suma total sería una manera de expresar el tamaño del pez, de modo que podríamos analizar y comparar a partir de los perfiles de los diferentes peces, y no de las medidas originales que constituyen su morfología.

Datos en una escala común

Una condición necesaria para los datos en el AC es que expresemos todas las observaciones en la misma escala. Por ejemplo, los recuentos de individuos en una tabla de frecuencias, una determinada unidad monetaria en una tabla sobre investigación de inversiones, o las medidas en centímetros en un estudio morfométrico. En el AC no tendría sentido que analizáramos datos expresados en distintas escalas de medida. A no ser que hagamos una transformación previa que nos permita homogeneizar las distintas escalas de la tabla. La mayor parte de los datos de este libro son datos de frecuencias, sin embargo, en el capítulo 23 se analiza una amplia variedad de otro tipo de datos. Veremos también cómo recodificarlos para que sean apropiados para el AC.

RESUMEN: Perfiles y espacio de perfiles

1. Un *perfil* es un conjunto de frecuencias (u otros valores positivos o ceros) divididas por su total, es decir es un conjunto de frecuencias relativas.
2. En las tablas de contingencia, las filas y las columnas definen conjuntos de frecuencias, que podemos reexpresar con relación a sus respectivos totales para obtener así perfiles fila o perfiles columna.
3. También podemos expresar las frecuencias marginales de las tablas de contingencia con relación a sus totales (el total de la tabla) para obtener así el perfil fila medio y el perfil columna medio.
4. Comparando los perfiles fila con su media, llegamos a las mismas conclusiones que comparando los perfiles columna con su media.
5. Podemos representar los perfiles con m elementos como puntos en un espacio m-dimensional. Sin embargo, debido a que la suma de estos m elementos es 1, los perfiles ocupan, en realidad, una región restringida de este espacio de dimensionalidad $(m - 1)$ que llamaremos *símplex*. Las líneas de unión de todos los pares de los m puntos unidad de los m ejes perpendiculares delimitan el símplex. A estos puntos unidad les llamamos *vértices* del símplex o del espacio de perfiles. Llamamos *sistema de coordenadas baricéntrico* al sistema de coordenadas definido por el símplex.
6. Es fácil visualizar los perfiles con tres elementos. El símplex formado por la unión de los tres vértices es simplemente un triángulo equilátero. A este caso particular de sistema de coordenadas baricéntrico lo llamamos *sistema de coordenadas triangular* (o *ternario).*

7. Podemos ampliar la idea de perfil a datos expresados en *escalas de razón*, es decir, a datos expresados como valores relativos. En dicho caso, es necesario haber obtenido todos los datos originales en la misma escala de medida.

CAPÍTULO 3

Masas y centroides

Existe una forma equivalente de ver las posiciones de los perfiles en el espacio de los perfiles, que nos será útil para nuestra eventual comprensión e interpretación del AC. Se basa en el concepto de media ponderada o centroide de un conjunto de puntos. En el cálculo usual de la media (no ponderada), todos los puntos tienen la misma masa. Sin embargo, una media ponderada permite asociar diferentes masas a los distintos puntos. Cuando ponderamos los puntos de distinta manera, el centroide no se sitúa exactamente en el centro «geográfico» de la nube de puntos, sino que tiende a situarse cerca de los puntos con mayor masa.

Contenido

Conjunto de datos 2: tipos de lectura y nivel de educación

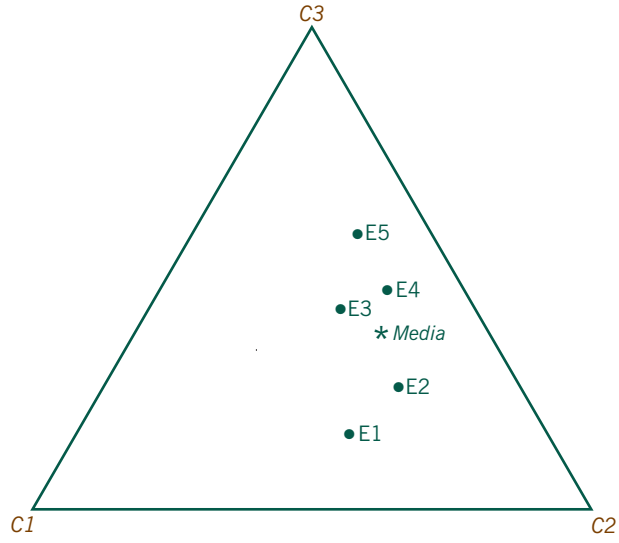
Utilicemos ahora un conjunto de datos habitual en investigación en ciencias sociales, una tabla de contingencia (o «clasificación cruzada») derivada de dos variables obtenidas en una encuesta. La tabla de la imagen 3.1 hace referencia a 312 lectores de un determinado periódico; en particular contiene datos sobre la minuciosidad de los encuestados en la lectura del periódico. Hemos clasificado a los encuestados en tres grupos de lectores: *rápidos*, *minuciosos* y *muy minuciosos*. Hemos cruzado estas categorías de lectura con el nivel de educación de los encuestados: una variable ordinal con cinco categorías que van desde algo de educación

Imagen 3.1:
Tabla del cruce del nivel de educación por tipo de lector, que muestra los perfiles fila y el perfil fila medio entre paréntesis, así como las masas de las filas (derivadas de los totales de las filas)

	Tipo de Lector				
Nivel de educación	*Rápidos C1*	*Minuciosos C2*	*Muy minuciosos C3*	*Total*	*Masas de las filas*
Educación primaria incompleta	5	7	2	*14*	*0,045*
E1	(0,357)	(0,500)	(0,143)		
Educación primaria	18	46	20	*84*	*0,269*
E2	(0,214)	(0,548)	(0,238)		
Educación secundaria incompleta	19	29	39	*87*	*0,279*
E3	(0,218)	(0,333)	(0,448)		
Educación secundaria	12	40	49	*101*	*0,324*
E4	(0,119)	(0,396)	(0,485)		
Educación universitaria incompleta	3	7	16	*26*	*0,083*
E5	(0,115)	(0,269)	(0,615)		
Total	*57*	*129*	*126*	*312*	
Perfil fila medio	*(0,183)*	*(0,413)*	*(0,404)*		

primaria, hasta algo de educación universitaria. En la imagen 3.1 mostramos los recuentos originales, así como los perfiles de los niveles de educación entre paréntesis, es decir, los perfiles fila. En la imagen 3.2 hemos representado gráficamente el diagrama de coordenadas triangular de los perfiles fila, similar al que vimos en el capítulo 2. En esta imagen, los puntos de las esquinas, o los vértices, del triángulo equilátero, representan los tres tipos de lectores (recordemos que cada vértice ocupa la posición de un perfil fila «puro», es decir un perfil completamente

Imagen 3.2:
Representación gráfica de los perfiles fila (nivel de educación) de la imagen 3.1 en coordenadas triangulares, que también indica la posición del perfil fila medio (última fila de la imagen 3.1)



concentrado en una categoría). Por ejemplo, el vértice *muy minuciosos, C3*, representa un perfil fila ficticio [0 0 1], que supuestamente contiene un 100% de lectores *muy minuciosos*.

Los puntos como medias ponderadas

Las posiciones de los niveles de educación en el triángulo las podemos ver también como medias ponderadas. Asignar pesos a los valores de una variable es un concepto bien conocido en estadística. Por ejemplo, supongamos que, en una clase de 26 estudiantes, la media de sus calificaciones calculada sumando las calificaciones de los 26 estudiantes y dividiendo por 26 es 7,5. En realidad, tres estudiantes obtuvieron un 9, siete un 8, y 16 un 7, de manera que podemos calcular de forma equivalente la calificación media asignando un peso de 3/26 a la calificación de 9, un peso de 7/26 a la de 8 y un peso de 16/26 a la de 7, siendo los pesos las frecuencias relativas de cada calificación. Dado que la calificación de 7 tiene más peso que las restantes, el valor de la media ponderada, 7,5, se halla «más cerca» de esta calificación. La media aritmética usual de los valores 7, 8 y 9 es de 8.

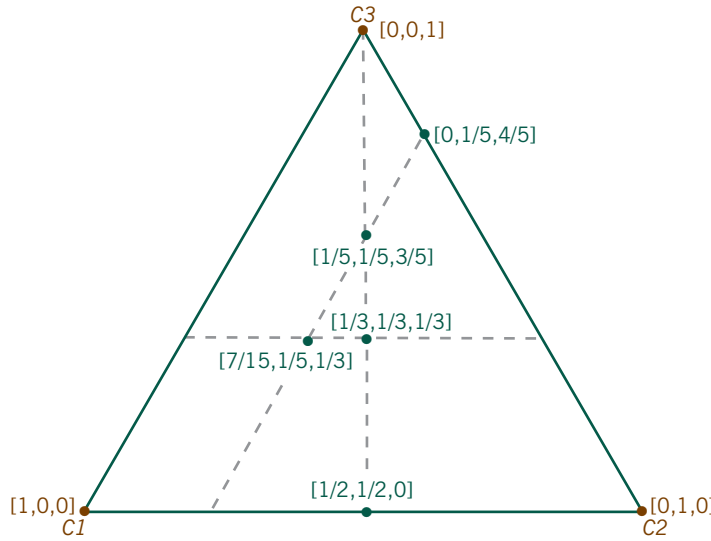
Los valores de los perfiles son los pesos asignados a los vértices

En la última fila de los datos de la imagen 3.1, vemos que a los encuestados de nivel de educación E5 (algo de educación universitaria) les corresponden las frecuencias 3, 7 y 16, es decir, las frecuencias relativas 0,115, 0,269 y 0,615, respectivamente. Imaginemos ahora cuál sería la *posición* media de estos 26 encuestados, si tres casos estuvieran situados en el vértice *rápidos, C1*, del triángulo; siete casos en el vértice *minuciosos, C2*, y 16 casos en el vértice *muy minuciosos, C3*. Es decir, en el espacio de perfiles, más que asociar los pesos con los valores de una variable, asociamos los pesos con las posiciones de los vértices. Hay más casos en la esquina de los *muy minuciosos*, por tanto, cabe esperar que la posición media de E5 se halle más cerca de este vértice, como ocurre en realidad. Por la misma razón, el perfil fila E1 se halla lejos de la esquina *muy minuciosos*, *C3*, ya que tiene muy poco peso (2 de 14, el 0,143) en esta categoría. Es decir, dentro del triángulo, situamos el punto que representa cada perfil fila como un punto medio de los vértices, en el que los valores del perfil —es decir, las frecuencias relativas— son los pesos asignados a los vértices. En consecuencia, podemos considerar los valores de los perfiles no sólo como coordenadas en un espacio multidimensional, sino también como los pesos asignados a los vértices de un símplex. Podemos extender este concepto a perfiles de más dimensiones. Por ejemplo, un perfil con cuatro elementos es también una posición media con relación a los cuatro vértices de un tetraedro tridimensional, que hemos ponderado con los elementos de este perfil.

Cada perfil es una media ponderada, o centroide, de los vértices

Términos alternativos a media ponderada son *centroide* o *baricentro*. En la imagen 3.3, hemos representado gráficamente algunos ejemplos de medias ponderadas en un espacio de perfiles. Por ejemplo, el perfil [1/3 1/3 1/3], que da el mismo peso a los tres vértices, se halla exactamente en el centro del triángulo, equidis-

Imagen 3.3:
Ejemplos de algunos centroides (medias ponderadas) de los vértices de un espacio de coordenadas triangular: los tres valores son los pesos asignados a los vértices (C1, C2, C3)



tante de los vértices, es decir en la posición de la media usual de los tres vértices. El perfil [1/2 1/2 0] se halla a medio camino entre el primer y el segundo vértices, ya que da igual peso a estos dos vértices y un peso igual a cero al tercero. En general podemos expresar la posición de un perfil [*a b c*], para el que se cumple que $a + b + c = 1$, como la de un centroide de los tres vértices, de la manera siguiente:

$$\textit{posición del centroide} = (a \times \text{vértice 1}) + (b \times \text{vértice 2}) + (c \times \text{vértice 3})$$

Por ejemplo, en la imagen 3.2 obtenemos la posición del nivel de educación E5 de la siguiente manera:

$$\text{E5} = (0{,}115 \times \textit{rápidos}) + (0{,}269 \times \textit{minuciosos}) + (0{,}615 \times \textit{muy minuciosos})$$

De forma similar, la posición del perfil medio es también una media ponderada de los vértices:

$$\textit{media} = (0{,}183 \times \textit{rápidos}) + (0{,}413 \times \textit{minuciosos}) + (0{,}404 \times \textit{muy minuciosos})$$

La media se halla más lejos del vértice *rápidos*, ya que su peso en este vértice es menor que el de los otros dos, que tienen aproximadamente los mismos pesos (imagen 3.2).

El perfil medio es también una media ponderada de los perfiles

El perfil medio es un punto especial; como acabamos de ver, no es solamente un centroide de los tres vértices, como cualquier otro perfil, sino que también es un centroide de los cinco perfiles fila, a los que hemos asignado diferentes pesos. Si volvemos de nuevo a la imagen 3.1, vemos que los totales de las filas son distintos: el primer nivel de educación E1 (algo de educación primaria) tiene sólo 14 indivi-

duos, mientras que el nivel de educación E4 (educación secundaria) tiene 101 individuos. En la última columna, titulada «masas de las filas», aparecen estas frecuencias marginales de las filas expresadas con relación al total de la muestra, 312. De la misma manera que contemplamos los perfiles como medias ponderadas de los vértices, podemos ver el perfil fila medio de la imagen 3.2 como una media ponderada de los perfiles, a los que hemos asignado pesos de acuerdo con sus frecuencias marginales; como si hubiera 14 individuos (una proporción de 0,045 de la muestra) en la posición E1, 84 individuos (una proporción de 0,269 de la muestra) en la posición E2, y así sucesivamente. Asignando estos pesos a los cinco perfiles, obtenemos exactamente la posición del perfil fila medio:

$$\textit{Perfil fila medio} = (0{,}045 \times \text{E1}) + (0{,}269 \times \text{E2}) + (0{,}279 \times \text{E3}) + (0{,}324 \times \text{E4}) + (0{,}083 \times \text{E5})$$

Este perfil fila medio se halla en una posición central entre los perfiles fila, pero más cerca de los perfiles con mayor frecuencia.

Las masas de las filas y las masas de las columnas

En el AC, los pesos asignados a los perfiles son tan importantes que les damos un nombre específico: *masas*. En la última columna de la imagen 3.1, se muestran las masas de las filas: 0,045, 0,269, 0,279, 0,324 y 0,083. En el AC preferimos el término «masa», sin embargo, es completamente equivalente a «peso». Preferimos un término alternativo a peso, para diferenciar la ponderación geométrica, de otros tipos de ponderación que podemos encontrar en el AC, como, por ejemplo, los pesos que asignamos a los subgrupos poblacionales en una encuesta. Todo lo referido a perfiles fila y a masas de las filas se puede aplicar de la misma manera a las columnas. En la imagen 3.4 mostramos la misma tabla de contingencia de la

	Tipo de Lector				
Nivel de educación	*Rápidos C1*	*Minuciosos C2*	*Muy minuciosos C3*	*Total*	*Perfil columna medio*
Educación primaria incompleta	5	7	2	*14*	*0,045*
E1	(0,088)	(0,054)	(0,016)		
Educación primaria	18	46	20	*84*	*0,269*
E2	(0,316)	(0,357)	(0,159)		
Educación secundaria incompleta	19	29	39	*87*	*0,279*
E3	(0,333)	(0,225)	(0,310)		
Educación secundaria	12	40	49	*101*	*0,324*
E4	(0,211)	(0,310)	(0,389)		
Educación universitaria incompleta	3	7	16	*26*	*0,083*
E5	(0,053)	(0,054)	(0,127)		
Total	*57*	*129*	*126*	*312*	
Masas de las columnas	*(0,183)*	*(0,413)*	*(0,404)*		

Imagen 3.4:
Tabla del cruce de nivel de educación por tipo de lector, que muestra los perfiles columna y el perfil columna medio entre paréntesis, así como las masas de las filas (obtenidas de los totales de las filas)

imagen 3.1, pero desde la óptica de las columnas. Es decir, se expresan las tres columnas como frecuencias relativas, con respecto al total de las columnas. Así, hemos obtenido tres perfiles con cinco elementos cada uno de ellos. Ahora, los totales de las columnas, con relación al total de la tabla, son las masas de las columnas que asignaremos a los perfiles de las columnas. El perfil columna medio está constituido por los totales de las filas dividido por el total de la tabla. Igual que antes para las filas, podemos expresar el perfil columna medio como una media ponderada de los tres perfiles columna *C1*, *C2* y *C3*:

$$\textit{Perfil columna medio} = (0{,}183 \times C1) + (0{,}413 \times C2) + (0{,}404 \times C3)$$

Fijémonos en que las masas de las filas y las masas de las columnas ejercen dos papeles distintos: como pesos y como elementos de los perfiles medios. En la imagen 3.4, el perfil columna medio está formado por las masas de las filas de la imagen 3.1. Sin embargo, aquí, las masas de las columnas son los elementos de lo que anteriormente era el perfil fila medio.

Interpretación del espacio de perfiles

En este momento, a pesar de que todavía no hemos visto algunos de los conceptos clave del AC, podemos empezar a interpretar la imagen 3.2. Los vértices del triángulo representan «perfiles puros» de tipos de lectores *C1*, *C2* y *C3*, mientras que los niveles de educación están constituidos por «mezclas» de tipos de lectores. Sus posiciones dentro del triángulo dependen de las proporciones de cada una de las categorías anteriores. Fijémonos en los siguientes aspectos de la representación gráfica:

- Dentro del triángulo, el grado de dispersión de los perfiles nos da una idea sobre la variabilidad existente en la tabla de contingencia. Cuanto más cerca se hallen los perfiles del centroide, menor será la variabilidad. En cambio, cuanto más se alejen del centroide, mayor será la variabilidad. El espacio de los perfiles se halla delimitado, de manera que los perfiles más extremos se situarán cerca de los lados del triángulo, o en el caso más extremo en uno de los vértices (por ejemplo, individuos con poca formación se situarían cerca del vértice *C1*). En las tablas de datos sobre ciencias sociales, como las que estamos considerando, dado que, en general, la variabilidad de los valores de los perfiles es relativamente pequeña, los perfiles ocupan sólo una pequeña región del espacio de perfiles, cerca de la media. Por ejemplo, el recorrido de los perfiles del primer elemento (es decir, la categoría de lector *C1*) va solamente de 0,115 a 0,357 (imagen 3.1), mientras que su recorrido potencial va de 0 a 1. En cambio, en datos sobre la investigación en ecología, como veremos más adelante, el recorrido de los valores de los perfiles es mucho mayor, a menudo debido a la presencia de muchos ceros en la tabla; en consecuencia, los perfiles se dispersan mucho más dentro del espacio de perfiles (segundo ejemplo del capítulo 10).

- En la representación, los perfiles se esparcen, en lo que se llamamos «dirección de dispersión», aproximadamente de abajo hacia arriba. Efectivamente, vemos que los cinco perfiles de los niveles de educación se sitúan de abajo hacia arriba en su orden natural (de E1 a E5), de menos a más formación. Arriba, el grupo E5 se halla cerca del vértice *C3*, que representa la categoría de los lectores *muy minuciosos* —ya hemos visto que este grupo contiene la mayor proporción (0,615) de este tipo de lectores—. Abajo, el nivel de educación más bajo, no lejos del borde del triángulo que sabemos que muestra perfiles con cero lectores *C3* (por ejemplo el punto [1/2 1/2 0] de la imagen 3.3 es un ejemplo de uno de estos puntos). La interpretación de esta distribución hay que buscarla en el hecho de que cuando nos movemos, de abajo hacia arriba, en general, los perfiles cambian respecto a la frecuencia relativa de la categoría *C3*, en contraste con la combinación de las categorías *C1* y *C2*. No observamos tendencia particular alguna hacia *C1* o hacia *C2*.

Unión de filas o de columnas

Supongamos que queremos reunir los dos niveles de educación primaria, E1 y E2, de la imagen 3.1, en una nueva fila que llamaremos E1&E2. Existen dos posibilidades de unión: la primera es sumar ambas filas, para obtener una nueva fila de frecuencias [23 53 22] con un total de 98 individuos y de perfil [0,235 0,541 0,224]; la segunda posibilidad es que consideremos al perfil E1&E2 como la media ponderada de los perfiles E1 y E2:

$$[0{,}235\ \ 0{,}541\ \ 0{,}224] = \frac{0{,}045}{0{,}314} \times [0{,}357\ \ 0{,}500\ \ 0{,}143] + \frac{0{,}269}{0{,}314} \times [0{,}214\ \ 0{,}548\ \ 0{,}238]$$

donde las masas de E1 y E2 son 0,045 y 0,269, respectivamente, que sumadas dan 0,314 (fijémonos en que los pesos de esta media ponderada son iguales a 14/98 y 84/98, siendo 14 y 84 los totales de las filas E1 y E2, respectivamente). Geométricamente, el perfil E1&E2 se halla en una línea que une E1 y E2, pero más cerca de E2, como podemos ver en la imagen 3.5. Las distancias de E1 a E1&E2 y de

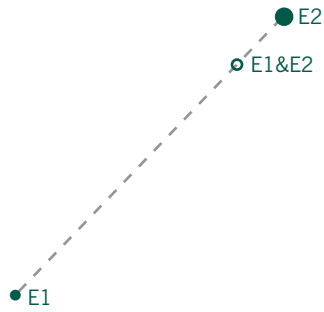


Imagen 3.5:
Ampliación de las posiciones de E1 *y* E2 *en la imagen 3.2, que muestra la posición del punto* E1&E2 *al unir ambas categorías.* E2 *tiene seis veces más masa que* E1*, en consecuencia* E1&E2 *se halla más cerca de* E2*, en un punto que divide el segmento que une* E1 *con* E2 *de acuerdo con la proporción 84:14 = 6:1*

E2 a E1&E2 se hallan en la misma proporción que los totales 84 y 14, es decir de 6 a 1. Por tanto, podemos considerar E1&E2 como el punto de equilibrio de las dos masas situadas en E1 y en E2, con la mayor masa en E2.

Distribuciones equivalentes de filas o de columnas

Supongamos que añadimos una fila a los datos de la imagen 3.1, una categoría de «sin educación reglada» que simbolizaremos por E0, de frecuencias [10 14 4] con relación a los tipos de lectores. El perfil de E0 es idéntico al perfil de E1, ya que las frecuencias de E0 son simplemente el doble de las de E1. Los dos conjuntos de frecuencias son *distribucionalmente equivalentes*. Por tanto, los perfiles de E0 y de E1 se hallan exactamente en el mismo punto del espacio de perfiles. Los podemos unir para dar a este punto una masa igual a la combinación de las masas de los dos perfiles, es decir un punto de frecuencias [15 21 6].

Cambio de masas

Las masas de las filas y las de las columnas son proporcionales a las sumas marginales de la tabla. Si por alguna razón importante tenemos que modificar las masas es fácil transformar la tabla. Por ejemplo, supongamos que queremos que los cinco niveles de educación de la imagen 3.1 tengan masas proporcionales a los tamaños de las poblaciones de procedencia y no proporcionales al tamaño de sus muestras. En tal caso, podemos cambiar los valores de la tabla multiplicando los perfiles de los niveles de educación por el tamaño de sus respectivas poblaciones de origen. Los perfiles de esta nueva tabla serán idénticos a los perfiles originales; sin embargo, las masas de las filas serán proporcionales a los tamaños de las poblaciones. Alternativamente, supongamos que, en vez de ponderar de forma distinta los niveles de educación, como hasta ahora, queremos ponderarlos de igual forma. Para ello podemos tomar la tabla de perfiles fila (o de forma equivalente, de porcentajes en cada fila), como si fuera la tabla original. En esta tabla, la suma total de los elementos de las filas es 1 (o el 100%), es decir, todos los niveles educativos tendrán la misma ponderación.

RESUMEN: Masas y centroides

1. Supongamos que queremos representar gráficamente los perfiles fila, es decir representamos los perfiles fila en el espacio símplex definido por los vértices de las columnas. En tal caso, cada vértice representa una categoría de las columnas: un perfil fila completamente concentrado en esa categoría.
2. Podemos interpretar cada perfil como un centroide (o media ponderada) de los vértices, en el que los pesos son los elementos de su perfil. Por tanto, los perfiles tenderán a hallarse cerca de los vértices para los que tengan valores mayores.
3. Cada perfil fila tiene asociado un peso, llamado *masa*, proporcional a la suma de los elementos de la fila de la tabla original. Podemos obtener el perfil fila medio como el centroide de los perfiles fila, ponderando cada perfil con su correspondiente masa.

4. Todo lo que hemos visto hasta ahora para los perfiles fila, lo podemos aplicar de la misma manera a las columnas de la tabla. En realidad, la mejor manera de pasar del análisis de filas al de columnas es transponer la tabla, es decir, que las columnas sean las filas y viceversa, y hacer lo que hemos visto para las filas.
5. Las filas (o las columnas) formadas sumando las frecuencias de filas (o de columnas) de la tabla tienen un perfil igual a las medias ponderadas de los perfiles de las filas (o de las columnas) que lo componen.
6. Las filas (o las columnas) con perfiles iguales son *distribucionalmente equivalentes*. Las podemos agregar en un solo punto.
7. Podemos modificar las masas de las filas (o de las columnas) para que sean proporcionales a determinados valores, simplemente multiplicando por un factor de escala.

CAPÍTULO 4

Distancia ji-cuadrado e inercia

En los capítulos 2 y 3 representamos gráficamente los perfiles, medimos de forma implícita las distancias entre ellos y luego interpretamos sus posiciones. En el AC las distancias entre los perfiles las medimos de forma algo más complicada, mediante la *distancia ji-cuadrado*. Esta distancia es la clave de muchas de las propiedades interesantes del AC. Existen varias maneras de justificar la utilización de la distancia ji-cuadrado. Algunas de ellas, más técnicas, quedan fuera del alcance de este libro, otras son más intuitivas. En este capítulo nos decantamos por estas últimas. Así, empezaremos por la interpretación geométrica del conocido *estadístico ji-cuadrado*, calculado a partir de los datos de una tabla de contingencia. Las ideas que hay detrás del estadístico ji-cuadrado nos llevan al concepto de distancia ji-cuadrado y al concepto de inercia. En el AC medimos la variabilidad de una tabla de datos mediante la inercia, un concepto muy relacionado con la distancia ji-cuadrado.

Contenido

La hipótesis de independencia, o de homogeneidad, en tablas de contingencia

Consideremos, otra vez, los datos de la imagen 3.1. Fijémonos en que, de los 312 individuos de la muestra, 57 (el 18,3%) están situados en el grupo de lectores *C1*, 129 (el 41,3%) en el de lectores *C2* y 126 (el 40,4%) en el de lectores *C3*; así pues, el perfil fila medio está constituido por las proporciones [0,183 0,413 0,404]. Si no existieran diferencias entre los niveles de educación, por lo que concierne al tipo de lectores, los perfiles de todas las filas deberían ser más o menos iguales al perfil fila medio. Las diferencias que observamos se deberían sólo a fluctuaciones del muestreo aleatorio. Si suponemos que no hay diferencias, o dicho de otra manera, si suponemos que los niveles de educación son *homogéneos* con relación al tipo de lectores, ¿cuáles serán, las frecuencias esperadas de la fila E5? El nivel de educación E5 consta de 26 individuos, por tanto esperaríamos que el 18,3% de éstos pertenecieran a la categoría *C1*; es decir, 26 × 0,183 = 4,76 (aunque no tenga sentido considerar 0,76 individuos, para estos cálculos es conveniente mantener los valores decimales). De la misma forma, esperaríamos que 26 × 0,413 = 10,74 individuos de E5, pertenecieran a la categoría *C2*, y 26 × 0,404 = 10,50 a la categoría *C3*. En estadística, el «supuesto de no diferencias» entre las filas de una tabla de contingencia (o de forma similar, entre las columnas) tiene distintas denominaciones, —«hipótesis de independencia» es una de ellas, o quizás, en este contexto, es más adecuada la denominación «supuesto (o asunción) de homogeneidad»—. Mediante el supuesto de homogeneidad, las frecuencias esperadas para la fila E5 serían [4,76 10,74 10,50], sin embargo, los valores observados son [3 7 16]. De forma similar, por el mismo supuesto de homogeneidad, podemos calcular las frecuencias esperadas de las restantes filas. En la imagen 4.1, mostramos, debajo de los valores observados de todas las filas, los correspondientes valores esperados. Llegaríamos exactamente a las mismas frecuencias esperadas si trabajáramos con los perfiles columna, es decir, suponiendo que los grupos de lectores son homogéneos.

Imagen 4.1: *Frecuencias observadas, tal como aparecen en la imagen 3.1 junto con las frecuencias esperadas (entre paréntesis) calculadas suponiendo que se cumple el supuesto de homogeneidad*

	Tipo de Lector				
Nivel de educación	*Rápidos C1*	*Minuciosos C2*	*Muy minuciosos C3*	*Total*	*Masas de las filas*
Educación primaria incompleta	5	7	2	*14*	*0,045*
E1	(2,56)	(5,78)	(5,66)		
Educación primaria	18	46	20	*84*	*0,269*
E2	(15,37)	(34,69)	(33,94)		
Educación secundaria incompleta	19	29	39	*87*	*0,279*
E3	(15,92)	(35,93)	(35,15)		
Educación secundaria	12	40	49	*101*	*0,324*
E4	(18,48)	(41,71)	(40,80)		
Educación universitaria incompleta	3	7	16	*26*	*0,083*
E5	(4,76)	(10,74)	(10,50)		
Total	*57*	*129*	*126*	*312*	
Perfil fila medio	*(0,183)*	*(0,413)*	*(0,404)*		

Contraste de la hipótesis de homogeneidad utilizando el estadístico ji-cuadrado

Las frecuencias observadas siempre serán distintas de las frecuencias esperadas. Sin embargo, en estadística queremos saber si estas diferencias son suficientemente grandes como para contradecir la hipótesis de que las filas son homogéneas. Es decir, queremos saber si es poco probable que las discrepancias entre las frecuencias observadas y las frecuencias esperadas se deban sólo al azar. Para responder a esta pregunta calcularemos una medida de discrepancia entre las frecuencias observadas y las frecuencias esperadas. Concretamente, calcularemos las diferencias entre cada par de frecuencias observadas y esperadas, las elevaremos al cuadrado, las dividiremos por las frecuencias esperadas e iremos acumulando los resultados hasta llegar a un valor final —el *estadístico ji-cuadrado*, que simbolizaremos por χ^2—:

$$\chi^2 = \sum \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}}$$

Cálculo de χ^2

Dado que en una tabla de 5 por 3 (5 × 3) hay 15 células, el cálculo constará de 15 términos. En el siguiente cálculo mostramos solamente los tres primeros, correspondientes a la fila E1, y los tres últimos, correspondientes a la fila E5:

$$\chi^2 = \frac{(5-2{,}56)^2}{2{,}56} + \frac{(7-5{,}78)^2}{5{,}78} + \frac{(2-5{,}66)^2}{5{,}66} + \cdots + \frac{(3-4{,}76)^2}{4{,}76} + \frac{(7-10{,}74)^2}{10{,}74} + \frac{(16-10{,}50)^2}{10{,}50} \qquad (4.1)$$

En este cálculo, la suma de los 15 términos es igual a 26,0. Cuanto mayor sea este valor, mayores serán las discrepancias entre las frecuencias observadas y las frecuencias esperadas y, en consecuencia, estaremos menos convencidos de la certeza del supuesto de homogeneidad. Para valorar si 26,0 es grande o pequeño utilizamos las tablas de la distribución ji-cuadrado, con sus correspondientes «grados de libertad». Así, para una tabla de 5 × 3, los grados de libertad son 4 × 2 = 8 (el número de filas menos uno, multiplicado por el número de columnas menos uno). A un valor del estadístico χ^2 de 26,0, con 8 grados de libertad, el valor *p* asociado es de 0,001. Este resultado nos indica que la probabilidad de que las frecuencias observadas en la imagen 4.1 se correspondan con el supuesto de homogeneidad es extremadamente baja —una entre mil—. Es decir, rechazamos la homogeneidad de la tabla y concluimos que es muy probable que existan diferencias reales entre los niveles de educación, en lo concerniente a los perfiles de los tipos de lectura.

Una expresión alternativa del estadístico χ^2, en términos de perfiles y de masas

De hecho, estamos más interesados en la capacidad del χ^2 para medir la falta de homogeneidad, es decir, para medir la heterogeneidad entre los perfiles, que en la prueba estadística de homogeneidad que acabamos de describir. Vamos a expresar de otra forma el estadístico χ^2. Para ello dividiremos el numerador y el denominador de cada uno de los tres términos de cada fila por

el cuadrado del total de la fila. Por ejemplo, fijémonos en los tres últimos términos del cálculo del estadístico χ^2 que mostramos en (4.1); dividimos el numerador y el denominador de cada uno de estos tres términos por el cuadrado del total de la fila E5, es decir, por 26^2, de esta forma, en vez de tener las frecuencias absolutas originales, tenemos los perfiles observados y los perfiles esperados;

$$\begin{aligned}\chi^2 &= 12 \text{ términos similares} \cdots + \frac{\left(\frac{3}{26}-\frac{4{,}76}{26}\right)^2}{\frac{4{,}76}{26^2}} + \frac{\left(\frac{7}{26}-\frac{10{,}74}{26}\right)^2}{\frac{10{,}74}{26^2}} + \frac{\left(\frac{16}{26}-\frac{10{,}50}{26}\right)^2}{\frac{10{,}50}{26^2}} \\ &= 12 \text{ términos similares} \cdots + \\ &+ 26 \times \frac{(0{,}115-0{,}183)^2}{0{,}183} + 26 \times \frac{(0{,}269-0{,}413)^2}{0{,}413} + 26 \times \frac{(0{,}615-0{,}404)^2}{0{,}404} \end{aligned} \quad (4.2)$$

Fijémonos en que hemos eliminado uno de los 26 que aparecía dividiendo en el denominador de cada uno de los tres términos y que ahora aparece como un factor que multiplica a cada término. De esta manera hemos conseguido expresar los términos como perfiles. Así, los 15 términos los calcularíamos de la manera siguiente:

$$\text{total de la fila} \times \frac{(\text{perfiles observados de la fila} - \text{perfiles esperados de la fila})^2}{\text{perfiles esperados de la fila}}$$

La inercia (total) es igual al estadístico χ^2 dividido por el tamaño de la muestra

Hagamos una modificación más en el cálculo del estadístico χ^2 que mostramos anteriormente, con el fin de ponerlo en sintonía con los conceptos que hasta ahora hemos visto en el AC. Dividamos los dos lados de la ecuación (4.2) por el tamaño total de la muestra, de manera que en cada término de la derecha de la ecuación aparezca en primer lugar un factor que, en vez de corresponder al total de la fila, corresponda a la masa de la misma.

$$\begin{aligned}\frac{\chi^2}{312} &= 12 \text{ términos similares} \cdots + \\ &+ 0{,}083 \times \frac{(0{,}115-0{,}183)^2}{0{,}183} + 0{,}083 \times \frac{(0{,}269-0{,}413)^2}{0{,}413} + 0{,}083 \times \frac{(0{,}615-0{,}404)^2}{0{,}404} \end{aligned} \quad (4.3)$$

donde 0,083 = 26/312 es la masa de la fila E5 (imagen 4.1). En el AC, llamamos *inercia total*, o simplemente *inercia*, al valor χ^2/n de la izquierda, donde n es el total de la tabla. Este valor es una medida de la varianza total de la tabla independiente de su tamaño. En estadística, este valor recibe diferentes nombres, uno de ellos es el de «coeficiente medio cuadrático de contingencia». A su raíz cuadrada la denominamos «coeficiente phi» (ϕ); por tanto, podemos expresar la inercia como ϕ^2. Si en la expresión (4.3) agrupamos los tres términos de cada fila, obtenemos la siguiente expresión para la inercia:

$$\frac{\chi^2}{312} = \phi^2 = 4 \text{ grupos similares de términos} \cdots + \\ + 0{,}083 \times \left[\frac{(0{,}115 - 0{,}183)^2}{0{,}183} + \frac{(0{,}269 - 0{,}413)^2}{0{,}413} + \frac{(0{,}615 - 0{,}404)^2}{0{,}404} \right] \qquad (4.4)$$

Ahora, los cinco grupos de términos de esta fórmula, uno de cada fila de la tabla, son iguales a la masa correspondiente de su fila (por ejemplo 0,083 para la fila E5), multiplicada por un valor al cuadrado, entre paréntesis, que tiene el aspecto de una distancia (para ser más precisos, el cuadrado de una distancia).

La distancia euclídea o pitagórica

En la expresión (4.4) que acabamos de ver, si no fuera por el hecho de que dividimos el cuadrado de las diferencias entre los elementos observados y los esperados del perfil por los elementos esperados, el valor entre los corchetes, sería exactamente el cuadrado de la distancia «directa» entre el perfil fila E5 y el perfil fila medio en un espacio físico tridimensional, es decir la *distancia euclídea* o *pitagórica*. Para comprenderlo mejor, vamos a verlo de otra manera, supongamos que representamos gráficamente los dos perfiles [0,115 0,269 0,615] y [0,183 0,413 0,404] con respecto a tres ejes perpendiculares. La distancia entre ellos sería la raíz cuadrada de la suma de los cuadrados de las diferencias entre las coordenadas de cada perfil, es decir:

$$\text{Distancia euclídea} = \sqrt{(0{,}115 - 0{,}183)^2 + (0{,}269 - 0{,}413)^2 + (0{,}615 - 0{,}404)^2} \qquad (4.5)$$

Esta distancia, cuyo valor es 0,264, corresponde exactamente a la distancia entre el punto E5 y la media de los perfiles que representamos gráficamente en la Imagen 3.2

La distancia ji-cuadrado: un ejemplo de distancia euclídea ponderada

Sin embargo, la expresión (4.4) no es la distancia euclídea —contiene un factor extra, en el denominador de cada término al cuadrado—. Dado que este factor redimensiona o *repondera* cada una de las diferencias al cuadrado, denominamos a esta variante de la distancia euclídea, *distancia euclídea ponderada*. En este caso en particular en el que los factores de ponderación que aparecen en el denominador son los elementos esperados del perfil, la denominamos *distancia ji-cuadrado*, o de forma sintética, distancia χ^2. Por ejemplo, la distancia χ^2 entre la fila E5 y el centroide es:

$$\text{Distancia } \chi^2 = \sqrt{\frac{(0{,}115 - 0{,}183)^2}{0{,}183} + \frac{(0{,}269 - 0{,}413)^2}{0{,}413} + \frac{(0{,}615 - 0{,}404)^2}{0{,}404}} \qquad (4.6)$$

su valor es de 0,431, mayor que la distancia euclídea que calculamos en (4.5), ya que los términos contenidos en la raíz cuadrada han aumentado de valor. En el próximo capítulo veremos cómo visualizar las distancias ji-cuadrado.

Interpretación geométrica de la inercia

A partir de (4.4) y (4.6) podemos expresar la inercia de la siguiente forma:

$$\text{Inercia} = \sum_i (i\text{-ésimo masa}) \times \left(\text{distancia } \chi^2 \text{ de } i\text{-ésimo perfil media}\right)^2 \qquad (4.7)$$

efectuando la suma con relación a las cinco filas de la tabla. Dado que la suma de las masas es 1, podemos decir que la inercia es la media ponderada de los cuadrados de las distancias χ^2 entre los perfiles fila y su perfil media. Por tanto, la inercia será alta cuando los perfiles fila presenten grandes desviaciones con relación a su media, y será baja cuando éstos se hallen cerca de la media. En la imagen 4.2 mostramos una secuencia de cuatro pequeñas matrices de datos, de baja a alta inercia total, cada una de ellas con cinco filas y tres columnas. También hemos representado gráficamente cada una de las matrices en coordenadas triangulares. Hemos escogido estos ejemplos, esencialmente, para visualizar incrementos de magnitud de las inercias. Esta secuencia de mapas también ilustra el concepto de asociación, o de correlación, entre las filas y las columnas de una matriz. Cuando la inercia es baja, los perfiles fila presentan poca variación y se hallan cerca de su perfil medio. En tal caso, decimos que existe poca asociación, o correlación, entre las filas y las columnas. Cuanto

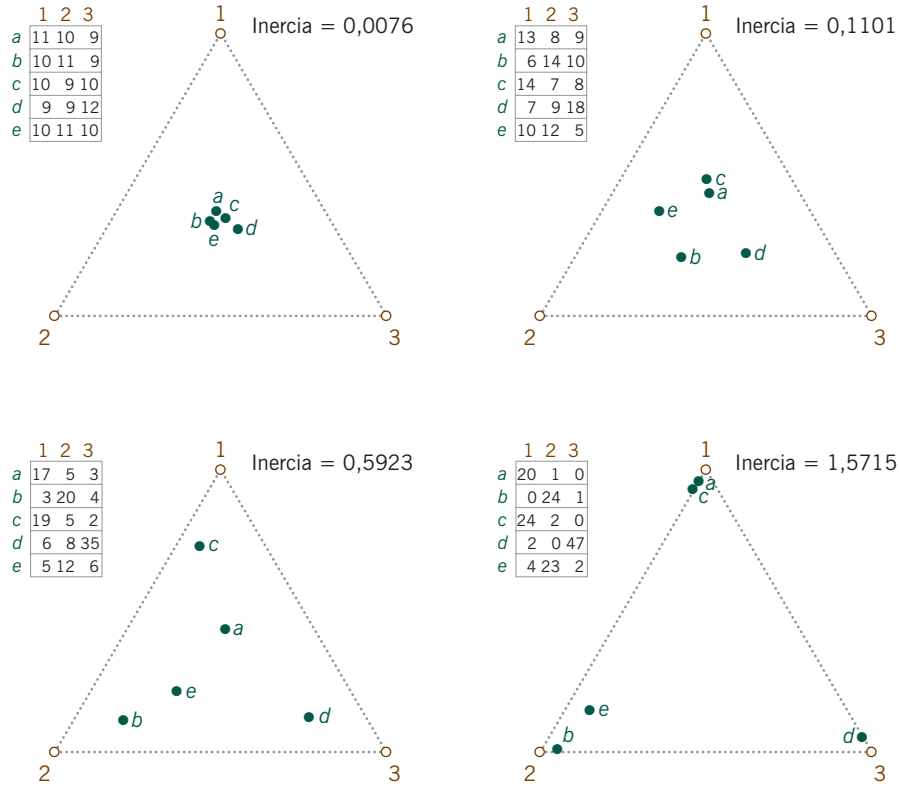


Imagen 4.2: *Serie de tablas de datos con inercia total en aumento. Cuanto mayor sea la inercia total, mayor será la asociación entre las filas y las columnas. Visualizamos este hecho con una mayor dispersión de los puntos en el espacio de perfiles. Hemos escogido los valores de estas tablas de manera que las sumas de las columnas sean todas iguales, y así también lo serán los pesos en la formulación de la distancia χ^2. Por tanto las distancias que observamos en estos mapas son distancias χ^2*

mayor sea la inercia, más cerca se hallarán los perfiles fila de los vértices columna. Es decir, mayor será la asociación entre las filas y las columnas. Más adelante, en el capítulo 8, describiremos de manera más formal la relación existente entre la inercia y el coeficiente de correlación entre las filas y las columnas.

Las inercias máxima y mínima

Si todos los perfiles fueran idénticos, y por tanto todos se hallaran en el mismo punto (su media), todas las distancias ji-cuadrado serían cero, también lo sería la inercia total. Por otro lado, se llegaría a la inercia máxima cuando todos los perfiles se hallaran exactamente en los vértices del espacio de perfiles. En tal caso, la inercia máxima sería igual a la dimensionalidad del espacio (en los ejemplos triangulares de la imagen 4.2, este valor máximo sería igual a 2).

La inercia de las filas es igual a la inercia de las columnas

Hasta ahora hemos visto los conceptos de perfil, de masa, de distancia χ^2 y de inercia, en términos de las filas de una tabla. Tal como comentamos en el capítulo 3, todo lo que hemos descrito hasta ahora para las filas lo podríamos aplicar, de forma equivalente, a las columnas de la tabla (en la imagen 3.4 podemos ver los valores de los perfiles columna, el perfil columna medio y las masas de las columnas). Podríamos comprobar que el resultado del cálculo de la inercia, según la ecuación (4.7) sería idéntico si lo calculáramos a partir de los perfiles columna. Es decir, la inercia total de la tabla, sería igual a la media ponderada de los cuadrados de las distancias χ^2 entre los perfiles columna y su perfil media, ponderadas ahora con las masas de las columnas.

Algo de notación

Esta sección no es imprescindible para la comprensión de los aspectos prácticos del análisis de correspondencias, por tanto la podemos obviar. Sin embargo, será útil para los lectores que quieran comprender la teoría y la literatura sobre el análisis de correspondencias (utilizaremos esta notación en el capítulo 14). Introduciremos un poco de notación estándar para los conceptos definidos hasta el momento aprovechando los datos de la imagen 3.1 (que repetimos en la imagen 4.1).

- n_{ij}: elemento de la tabla de contingencia situado en la *i*-ésima fila y en la *j*-ésima columna, por ejemplo $n_{21} = 18$.
- n_{i+}: el total de la *i*-ésima fila, por ejemplo $n_{3+} = 87$ (el subíndice + indica suma de los elementos del correspondiente índice).
- n_{+j}: el total de la *j*-ésima columna, por ejemplo $n_{+2} = 129$.
- n_{++}: o simplemente n, el total de la tabla, por ejemplo $n = 312$.
- p_{ij}: n_{ij} dividido por el total de la tabla, así, $p_{21} = n_{21}/n = 18/312 = 0{,}0577$.
- r_i: la masa de la *i*-ésima fila, así $r_i = n_{i+}/n$ (lo que equivale a p_{i+}, la suma de frecuencias relativas de la *i*-ésima fila p_{ij}); así $r_3 = 87/312 = 0{,}279$; simbolizamos al vector de masas como **r**.

- c_j: la masa de la j-ésima columna, es decir $c_j = n_{+j}/n$ (lo que equivale a p_{+j}, la suma de las frecuencias relativas de la j-ésima columna p_{ij}); por ejemplo c_2 = 129/312 = 0,414; simbolizamos al vector de las masas de las columnas como **c**.

- a_{ij}: el j-ésimo elemento del perfil de la fila i, así $a_{ij} = n_{ij}/n_{i+}$; así a_{21} = 18/84 = 0,214; simbolizamos al perfil de la fila i por el vector $\mathbf{a}_i$.

- b_{ij}: el i-ésimo elemento del perfil de la columna j, así $b_{ij} = n_{ij}/n_{+j}$; así b_{21} = 18/57 = 0,316; simbolizaremos el perfil de la columna j por el vector $\mathbf{b}_j$.

- $\sqrt{\sum_j (a_{ij} - a_{i'j})^2/c_j}$: la distancia χ^2 entre el i-ésimo y el i'-ésimo perfil fila, lo simbolizamos por $\|\mathbf{a}_i - \mathbf{a}_{i'}\|_c$; así de la imagen 3.1

  $$\|\mathbf{a}_1 - \mathbf{a}_2\|_c = \sqrt{\frac{(0,357-0,214)^2}{0,183} + \frac{(0,500-0,548)^2}{0,413} + \frac{(0,143-0,238)^2}{0,404}} = 0,374.$$

- $\sqrt{\sum_i (b_{ij} - b_{ij'})^2/r_i}$: distancia χ^2 entre el j-ésimo y la j'-ésimo perfil columna, lo simbolizamos por $\|\mathbf{b}_j - \mathbf{b}_{j'}\|_r$; así de la imagen 3.4

  $$\|\mathbf{b}_1 - \mathbf{b}_2\|_r = \sqrt{\frac{(0,088-0,054)^2}{0,045} + \frac{(0,316-0,357)^2}{0,269} + \ldots\text{etc.}} = 0,323$$

  donde 0,088 = 5/57; 0,054 = 7/129; 0,045 = 14/312; etc.

- $\sqrt{\sum_j (a_{ij} - c_j)^2/c_j}$: distancia χ^2 entre el i-ésimo perfil fila $\mathbf{a}_i$ y el perfil fila medio **c** (el vector de las masas de las columnas), lo simbolizamos por $\|\mathbf{a}_i - \mathbf{c}\|_c$; así de la imagen 3.1

  $$\|\mathbf{a}_1 - \mathbf{c}\|_c = \sqrt{\frac{(0,357-0,183)^2}{0,183} + \frac{(0,500-0,413)^2}{0,413} + \frac{(0,143-0,404)^2}{0,404}} = 0,594.$$

- $\sqrt{\sum_i (b_{ij} - r_i)^2/r_i}$: distancia χ^2 entre el j-ésimo perfil columna $\mathbf{b}_j$ y el perfil columna medio **r** (el vector de las masas de las filas), lo simbolizamos por $\|\mathbf{b}_j - \mathbf{r}\|_r$; así de la imagen 3.4

  $$\|\mathbf{b}_1 - \mathbf{r}\|_r = \sqrt{\frac{(0,088-0,045)^2}{0,045} + \frac{(0,316-0,269)^2}{0,269} + \ldots\text{etc.}} = 0,332.$$

Con esta notación, la fórmula de la inercia total (4.7) es:

$$\phi^2 = \frac{\chi^2}{n} = \sum_i r_i \|\mathbf{a}_i - \mathbf{c}\|_c^2 = \sum_i r_i \sum_j \left(\frac{p_{ij}}{r_i} - c_j\right)^2 / c_j \qquad \text{(por fila)} \tag{4.8}$$

$$= \sum_j c_j \|\mathbf{b}_i - \mathbf{r}\|_r^2 = \sum_j c_j \sum_i \left(\frac{p_{ij}}{c_j} - r_i\right)^2 / r_i \qquad \text{(por columna)} \tag{4.9}$$

y su valor 0,0833, por lo tanto, χ^2 = 0,0833 × 312 = 26,0.

RESUMEN:
Distancia ji-cuadrado e inercia

1. El estadístico ji-cuadrado (χ^2) es una medida global de las diferencias entre las frecuencias observadas y las frecuencias esperadas de una tabla de contingencia. Calculamos las frecuencias esperadas mediante la hipótesis de homogeneidad de los perfiles fila (o de los perfiles columna)
2. La *inercia* (*total*) de una tabla de contingencia es igual al estadístico χ^2 dividido por el total de la tabla.
3. Geométricamente, la inercia mide lo «lejos» que se hallan los perfiles fila (o los perfiles columna) de su perfil medio. Podemos considerar que el perfil medio simboliza la hipótesis de homogeneidad (es decir, de igualdad) de los perfiles.
4. Medimos las distancias entre los perfiles utilizando la *distancia ji-cuadrado* (distancia χ^2). La formulación de esta distancia es similar a la *distancia euclídea* (o *pitagórica*) entre puntos en un espacio físico, salvo por el hecho de que dividimos cada cuadrado de la diferencia entre coordenadas por su correspondiente elemento del perfil medio.
5. Podemos expresar la inercia de manera que la podamos interpretar como una media ponderada de las distancias χ^2 entre los perfiles fila y su perfil medio (de forma similar, entre los perfiles columna y su media).

CAPÍTULO 5

Representación gráfica de distancias ji-cuadrado

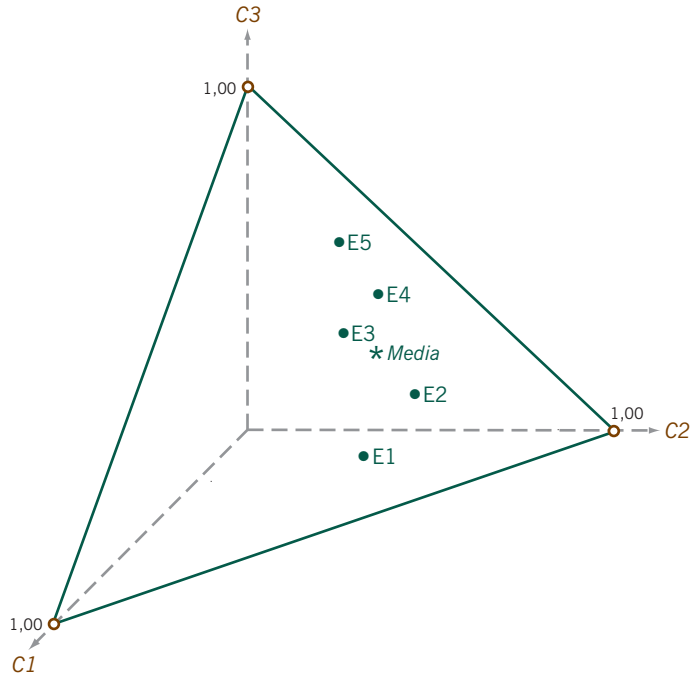
En el capítulo 3 interpretamos las posiciones de los perfiles bidimensionales en un sistema de coordenadas triangular mediante distancias euclídeas. En el capítulo 4 definimos la distancia ji-cuadrado (distancia χ^2) entre perfiles. Vimos la relación existente entre la distancia χ^2, el estadístico ji-cuadrado y la inercia de una matriz de datos. La distancia χ^2 es una distancia euclídea ponderada, en la que ponderamos los cuadrados de las diferencias entre coordenadas, con el inverso del correspondiente elemento del perfil medio. Hasta ahora, no hemos visualizado realmente las distancias χ^2 entre perfiles. Sólo lo hemos hecho en la imagen 4.2, en la que los elementos del perfil medio eran iguales y, por tanto, en este caso particular, las distancias χ^2 también eran distancias euclídeas. En este capítulo veremos cómo con una simple transformación del espacio de perfiles, las distancias que observamos en nuestras representaciones gráficas son distancias χ^2.

Contenido

Diferencia entre la distancia χ^2 y la distancia euclídea usual

En la imagen 5.1 hemos representado gráficamente los perfiles fila de la imagen 3.1 en unos ejes de coordenadas perpendiculares, en el espacio físico tridimensional habitual. Aquí las distancias entre perfiles no son distancias χ^2, son distancias euclídeas (sin ponderar) [fórmula (4.5)]. En este tipo de espacio, calculamos las distancias entre dos perfiles con elementos x_j e y_j, respectivamente

Imagen 5.1:
Espacio de perfiles, que muestra los perfiles de los niveles de educación en un triángulo equilátero en un espacio tridimensional; las distancias son euclídeas



(donde $j = 1, ..., J$), sumando los cuadrados de las diferencias de las coordenadas $(x_j - y_j)^2$, en todas las dimensiones j y calculando finalmente la raíz cuadrada de la suma resultante. Esta es la manera como usualmente calculamos las distancias «directamente» en el espacio físico con el que estamos familiarizados. Como hemos visto, el cálculo de la distancia χ^2 es distinto, ya que dividimos cada diferencia al cuadrado por el correspondiente elemento del perfil medio. Es decir, cada término es igual a $(x_j - y_j)^2/c_j$, donde c_j es el correspondiente elemento del perfil medio. Dado que solamente podemos interpretar y comparar distancias en nuestro espacio físico habitual, sería deseable algún tipo de modificación del mapa que hiciera que las distancias «directas» habituales se conviertieran en distancias χ^2. Afortunadamente, como veremos a continuación, esto es posible mediante transformaciones simples de los perfiles.

Transformación de las coordenadas antes de representarlas gráficamente

En el cálculo de la distancia χ^2, podemos reescribir cada término de la forma $(x_j - y_j)^2/c_j$, como $(x_j/\sqrt{c_j} - y_j/\sqrt{c_j})^2$. Esta forma equivalente de expresar el término general en el cálculo de la distancia es formalmente idéntica a la de la distancia euclídea usual; es decir, como una diferencia al cuadrado. El único cambio es que ahora las coordenadas no son los valores originales x_j e y_j, sino que las hemos transformado en $x_j/\sqrt{c_j}$ e $y_j/\sqrt{c_j}$. Ello sugiere que, en vez de utilizar como coordenadas los elementos originales de los perfiles, podríamos utilizar estos elementos divididos por las raíces cuadradas de los correspondientes elementos del per-

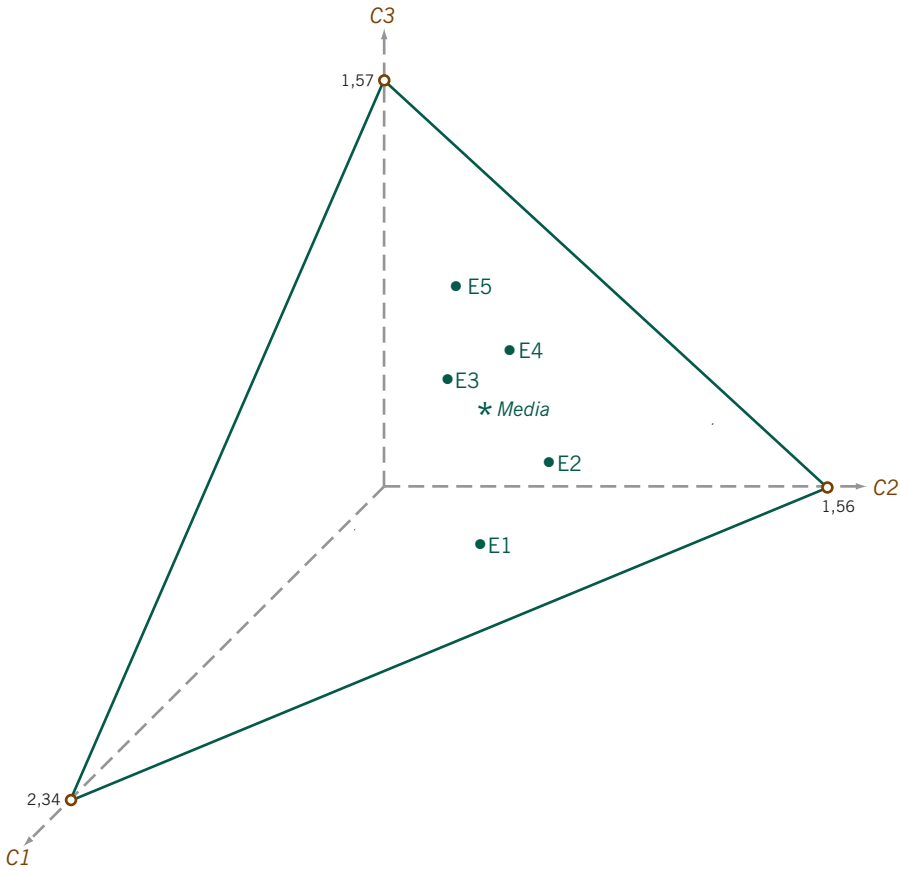


Imagen 5.2:
El espacio de perfiles muestra los ejes extendidos en distinta proporción, de manera que las distancias entre perfiles se convierten en distancias χ^2

fil medio. En tal caso, la distancia euclídea habitual entre estas coordenadas transformadas sería la distancia χ^2 que buscamos.

Efecto práctico de la transformación

Los valores c_j son los elementos del perfil medio, por tanto, todos son menores que 1. En consecuencia, la transformación consistente en dividir los elementos del perfil por $\sqrt{c_j}$ comportará un incremento del valor de todas las coordenadas. De todas maneras, unas aumentarán más que las otras. Si un determinado c_j es relativamente pequeño en comparación con los otros (es decir, la frecuencia de la *j*-ésima categoría de la columna es relativamente pequeña), entonces las correspondientes coordenadas $x_j/\sqrt{c_j}$ e $y_j/\sqrt{c_j}$ aumentarán de forma relativamente grande. A la inversa, una c_j grande, correspondiente a una categoría más frecuente, comportará un incremento relativamente menor de las coordenadas transformadas. Por tanto, la transformación aumenta los valores de las categorías con frecuencias bajas, relativamente más que los de las categorías con frecuencias altas. En el espacio sin transformar de la imagen 5.1, los vértices se hallan a una unidad

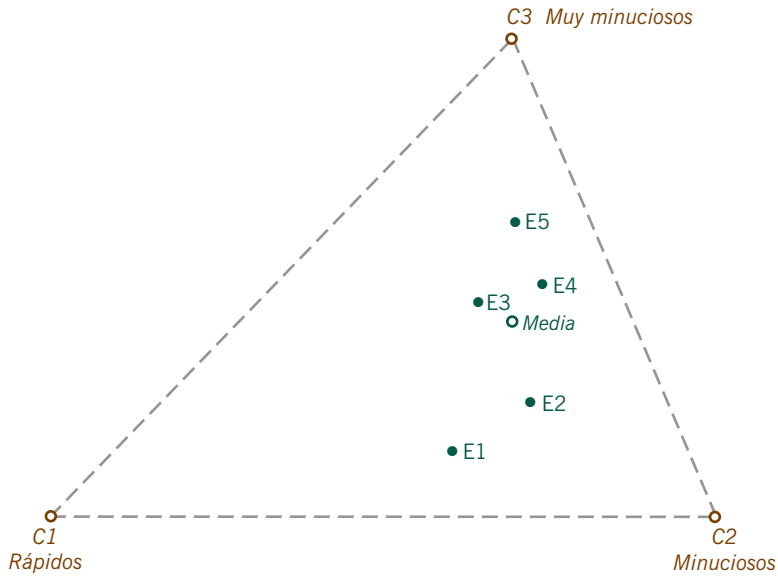


Imagen 5.3:
Espacio triangular de perfiles del espacio extendido de la imagen 5.2 situado en un «plano» (comparemos con la imagen 3.2). El triángulo se ha estirado más en la dirección de C1, la categoría menos frecuente

física del origen (es decir, del punto cero) de los tres ejes de coordenadas. El primer vértice, de coordenadas [1 0 0], se convertirá en la posición $[1/\sqrt{c_1}\ 0\ 0]$; es decir, su posición en el primer eje se estira hasta llegar al valor $1/\sqrt{0{,}183} = 2{,}34$. De forma similar, el segundo y el tercer vértice se estiran hasta los valores $1/\sqrt{c_2} = 1/\sqrt{0{,}413} = 1{,}56$ y $1/\sqrt{c_3} = 1/\sqrt{0{,}404} = 1{,}57$, respectivamente. En la imagen 5.2 mostramos estos valores junto a los vértices de los correspondientes ejes. Los perfiles transformados ocupan nuevas posiciones en el espacio, pero siguen dentro del triángulo definido por los vértices transformados. Fijémonos en que el estiramiento es mayor en la dirección de *C1*, la categoría con menor frecuencia marginal.

Interpretación alternativa en términos de ejes de coordenadas recalibrados

Geométricamente, podríamos ver la situación anterior de otra manera. En los tres ejes de los sistemas de coordenadas sin transformar de las imágenes 2.4 y 5.1, las marcas que indican las escalas (por ejemplo, los valores 0,1; 0,2; 0,3; etc.) se hallan separadas a intervalos iguales; sin embargo, como indica la imagen 5.2, la transformación provoca una extensión de los tres vértices. Aunque después de la transformación las escalas en los tres ejes son distintas, podríamos seguir considerando los tres vértices como si fueran perfiles unitarios, es decir, a una unidad del origen. En el eje *C1*, un intervalo de 0,1 entre dos marcas sería una longitud física de 0,234, mientras que, en los ejes *C2* y *C3*, estos intervalos serían de 0,156 y de 0,157, respectivamente. Por tanto, el intervalo unidad en el eje *C1* sería aproximadamente un 50% más largo que el mismo intervalo en los otros

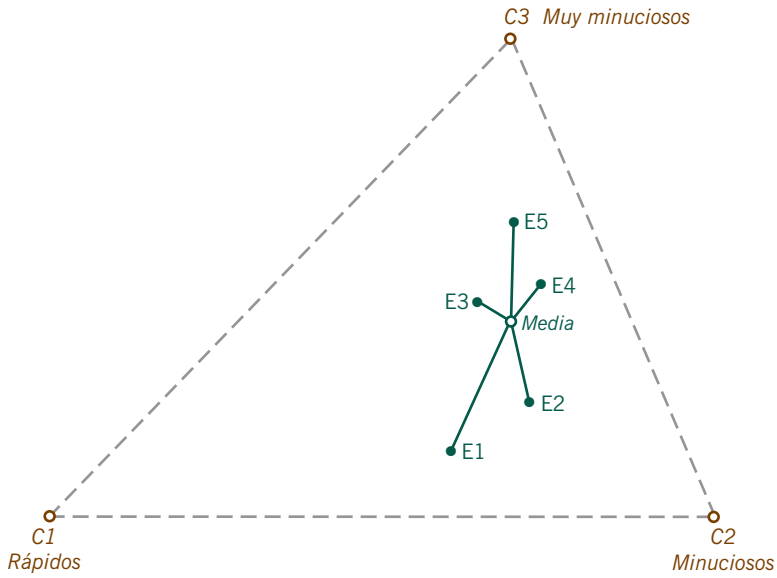


Imagen 5.4:
Espacio de perfiles «extendido» que muestra las distancias χ^2 de los perfiles a su centroide; la inercia es la media ponderada de la suma de los cuadrados de estas distancias y el estadístico χ^2 es la inercia multiplicada por el tamaño de la muestra (en este ejemplo, $n = 312$)

dos ejes. A pesar de ello, seguiríamos utilizando los valores originales de los perfiles para situarlos en el espacio tridimensional. En definitiva, sea cual sea la manera cómo veamos la transformación —ya sea como una transformación de los valores de los perfiles, o como un estiramiento y una recalibración de los ejes—, el resultado es el mismo: ahora situamos los perfiles en el espacio triangular estirado que mostramos en la imagen 5.2. En la imagen 5.3 hemos representado gráficamente el triángulo extendido en un plano. Queda claro que el vértice *C1*, correspondiente a la categoría más rara de lectores *rápidos*, la categoría que más se ha estirado.

Interpretación geométrica de la inercia y del estadístico χ^2

Ahora que en el espacio transformado las distancias son distancias χ^2, podemos trazar líneas de unión entre los perfiles y su media para mostrar, así, las distancias χ^2 entre los perfiles y su media (imagen 5.4). Por la fórmula (4.7), sabemos que la suma de las distancias de las filas a su media, ponderadas con sus respectivas masas, es igual a la inercia total de la tabla. Si en vez de ponderar con las masas, ponderamos con las frecuencias totales de las filas (la frecuencia total de una fila es n veces la masa de la fila, siendo n la suma total de la tabla), entonces la suma ponderada de los cuadrados de estas distancias es igual al estadístico χ^2. Obtenemos resultados equivalentes con los perfiles columna y el perfil columna medio. Por tanto, geométricamente, podemos interpretar la inercia y el estadístico χ^2 como medidas del grado de dispersión de los perfiles (de las filas o de las columnas) con relación a su media.

El principio de equivalencia distribucional

Para explicar este principio consideremos otra vez la imagen 3.1. Supongamos que podemos distinguir dos tipos de lectores *minuciosos*, los que se fijan más en la información política y los que se fijan más en la información cultural y deportiva. Simbolizaremos estas dos categorías por *C2a* y *C2b*, respectivamente. Supongamos, además, que en estas dos nuevas columnas, las frecuencias relativas de los niveles de educación son las mismas. Es decir, que no hay diferencias entre ambas subdivisiones del grupo de lectores *minuciosos* por lo que respecta a su educación. En el capítulo 3 dijimos que este tipo de columnas eran distribucionalmente equivalentes, en tanto que tienen los mismos perfiles. La subdivisión de la columna *C2* en *C2a* y *C2b* no aporta nueva información sobre las diferencias entre los niveles de educación. Por tanto, cualquier análisis de estos datos debería dar los mismos resultados, tanto si subdividimos *C2*, como si lo dejamos como una sola categoría. Decimos que un análisis que satisface esta propiedad cumple el *principio de equivalencia distribucional*. Si hubiéramos utilizado las distancias euclídeas habituales para medir las distancias entre los perfiles de los niveles de educación, no se cumpliría este principio ya que si hubiéramos hecho la mencionada subdivisión, hubiésemos obtenido resultados distintos. En cambio, la distancia χ^2 cumple siempre este principio, no se ve afectada por este tipo de subdivisiones de las categorías de la matriz de datos. Es decir, si unimos dos columnas distribucionalmente equivalentes, no cambian las distancias χ^2 entre las filas. En la práctica, esto significa que podemos unir columnas con perfiles similares sin que la geometría de las filas se vea afectada, y viceversa. El hecho de que en este tipo de análisis la introducción de arbitrariedades técnicas que modifiquen el número de categorías no afecte al resultado, y que éste sólo se vea modificado si introducimos modificaciones sustanciales, da ciertas garantías a los investigadores.

Las distancias χ^2 hacen que las contribuciones de las categorías sean más parecidas

Ya conocemos cómo organizar una representación gráfica para visualizar las distancias χ^2, pero ¿por qué tenemos que visualizar las distancias χ^2? ¿Por qué no utilizamos directamente distancias euclídeas? Podemos justificar la utilización de las distancias χ^2 de muchas maneras, unas más técnicas que otras. Existen razones más profundas que van más allá de las derivadas de la visualización del estadístico χ^2 que acabamos de presentar. Una de ellas se basa en la constatación de que existen diferencias importantes en las varianzas de los valores de las frecuencias de las distintas categorías. Así, por ejemplo, en la imagen 3.1 podemos ver el recorrido de los valores de los perfiles de la columna *C1* (de 0,115 a 0,357), una columna con frecuencias pequeñas, que es menor que el de la columna *C3* (de 0,143 a 0,615), una columna con frecuencias mayores. Esta observación ilustra una regla general sobre los datos de frecuencias: los conjuntos de frecuencias pequeñas presentan menor dispersión que los conjuntos de frecuencias grandes. Lo podemos ver calculando las contribuciones de las categorías de la imagen 3.1 a los cuadrados de las distancias euclídeas y χ^2, distancias entre los perfiles de los

	Euclídea			χ^2		
FILA	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*
E1	28,7	7,1	64,2	47,1	5,1	47,7
E2	2,1	38,7	59,1	4,7	37,2	58,1
E3	13,2	66,4	20,4	25,5	56,7	17,8
E4	37,1	2,8	60,1	56,6	1,9	41,5
E5	6,5	29,7	63,9	13,3	27,1	59,6
Global	*17,0*	*21,8*	*61,2*	*31,3*	*17,7*	*51,0*

Imagen 5.5:
Porcentajes de contribución de las categorías de las columnas a los cuadrados de las distancias euclídea y χ^2 de los perfiles fila a su centroide (datos de la imagen 3.1)

niveles de educación y su centroide (perfil medio). Por ejemplo, el cuadrado de la distancia euclídea entre el perfil del quinto nivel de educación E5 y el centroide es:

$$\begin{aligned}(\text{Distancia euclídea})^2 &= (0{,}115-0{,}183)^2+(0{,}269-0{,}413)^2+(0{,}615-0{,}404)^2\\ &= 0{,}00453+0{,}02080+0{,}04475\\ &= 0{,}07008\end{aligned}$$

mientras que el cuadrado de la distancia χ^2 es:

$$\begin{aligned}(\text{Distancia }\chi^2)^2 &= \frac{(0{,}115-0{,}183)^2}{0{,}183}+\frac{(0{,}269-0{,}413)^2}{0{,}413}+\frac{(0{,}615-0{,}404)^2}{0{,}404}\\ &= 0{,}02480+0{,}05031+0{,}11081\\ &= 0{,}18592\end{aligned}$$

[véanse ecuaciones (4.5) y (4.6)]. Cada una de estas distancias al cuadrado es la suma de tres valores correspondientes a las tres categorías de las columnas. Para valorar la contribución de cada tipo de lector, podemos expresar estos tres valores como porcentajes respecto de la distancia total. Por ejemplo, la contribución de la categoría *C1*, al cuadrado de la distancia euclídea es de 0,00453 sobre un total de 0,07008, es decir el 6,5%; mientras que la contribución de *C1* al cuadrado de la distancia χ^2 es de 0,02480 sobre un total de 0,18592, es decir el 13,3% (fila E5 de la imagen 5.5). En la imagen 5.5 mostramos las contribuciones de todos los términos, así como la contribución global de cada categoría, calculada considerando conjuntamente todos los términos de la misma (última fila de la imagen 5.5). Así, vemos que la contribución global de la categoría *C1* a la distancia euclídea es del 17,0%, mientras que la contribución global de esta categoría a la distancia χ^2 es del 31,3%. Este ejercicio ilustra el fenómeno general de que *C1*, la categoría con frecuencias más pequeñas, contribuye menos a la distancia euclídea que, por ejemplo, *C3*. Sin embargo, la contribución de *C1* a la distancia χ^2 se ve incrementada gracias a la división por las frecuencias medias.

Distancia euclídea ponderada

Como vimos en el capítulo 4, la distancia χ^2 es un ejemplo de distancia euclídea ponderada, su definición general es la siguiente:

$$\text{Distancia euclídea ponderada} = \sqrt{\sum_{j=1}^{p} w_j \left(x_j - y_j\right)^2} \qquad (5.1)$$

donde w_j son los valores positivos de los pesos y x_j, con $j = 1, ..., p$ e y_j, con $j = 1, ..., p$ son dos puntos en un espacio p-dimensional. En el análisis de componentes principales (ACP), un método muy relacionado con el AC, las p dimensiones vienen definidas por variables continuas, a menudo en diferentes escalas de medida. En el ACP eliminamos el efecto de la escala sobre la varianza dividiendo los datos por las desviaciones estándar s_j de las respectivas variables. De esta manera, reemplazamos las observaciones x_j e y_j de la variable j por x_j/s_j e y_j/s_j. Podemos ver esta operación como la utilización de una distancia euclídea ponderada con pesos $w_j = 1/s_j^2$, los inversos de las varianzas. En la definición de la distancia χ^2 entre perfiles, los pesos son iguales a $w_j = 1/c_j$, es decir, son iguales a los inversos de los elementos del perfil medio.

Justificación teórica de la distancia χ^2

A pesar de que en el AC los perfiles se hallan en la misma escala de frecuencias relativas, seguimos teniendo la necesidad de compensar las diferencias entre varianzas, situación similar a la del ACP. En la *distribución de Poisson*, una de la distibuciones estadísticas estándar para recuentos, es inherente el hecho de que los conjuntos de frecuencias con medias más elevadas tienen varianzas mayores que los conjuntos de frecuencias con medias menores. Precisamente, una característica de la distribución de Poisson es que la varianza es igual a su media. En nuestro contexto, podemos interpretar la transformación de las frecuencias —consistente en dividir por la raíz cuadrada de la frecuencia esperada (media)— como una estandarización de los datos, ya que la raíz cuadrada de la frecuencia media es un equivalente a la desviación típica. De todas formas, existen otros procedimientos de estandarización. Pero, ¿por qué la distancia χ^2 es tan especial? Aparte de cumplir el principio de equivalencia distribucional y de hacer que el análisis de filas y el de columnas sean simétricos, otra ventaja de la utilización de la distancia χ^2 hay que buscarla en las propiedades de la *distribución multinomial*, una distribución estadística multivariante para recuentos. En el apéndice teórico (A) vemos este tema con más profundidad.

RESUMEN: Representación gráfica de distancias ji-cuadrado

1. Podemos visualizar las distancias χ^2 entre perfiles, en el espacio físico habitual (euclídeo) transformando los perfiles antes de representarlos gráficamente. Esta transformación consiste en dividir cada elemento del perfil por la raíz cuadrada del correspondiente elemento del perfil medio.
2. Otra posibilidad para visualizar las distancias χ^2 entre perfiles consiste, en vez de transformar los elementos del perfil antes de representarlos, en estirar los

ejes de manera que, en cada eje, la unidad tenga una longitud inversamente proporcional a la raíz cuadrada del correspondiente elemento del perfil medio.

3. La distancia χ^2 es un caso especial de distancia euclídea ponderada en la que los pesos son los inversos de los correspondientes elementos del perfil medio.
4. Cuando representamos gráficamente los perfiles fila, podemos ver el redimensionamiento de las coordenadas (o la extensión de los ejes) como una estandarización de las columnas de la tabla, que hace que las comparaciones entre los perfiles fila sean más equitativas.
5. Las distancias χ^2 cumplen el *principio de equivalencia distribucional*, que garantiza la estabilidad de las distancias entre las filas, cuando dividimos las columnas en componentes similares, o cuando unimos columnas similares.

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

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.

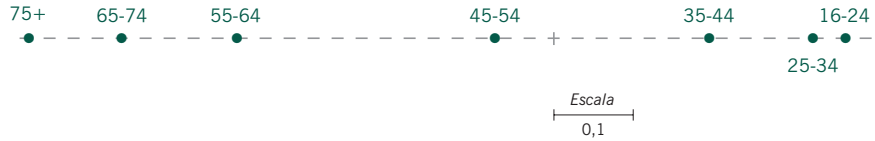
Proyección de los perfiles en subespacios

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

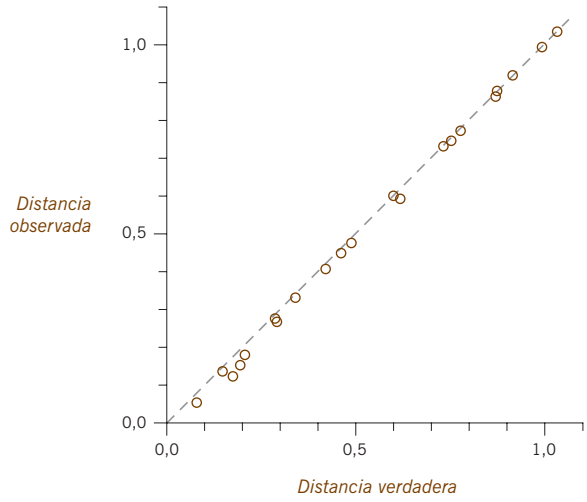


Imagen 6.4:
Distancias observadas entre todos los pares de puntos del mapa de la imagen 6.3, representadas con relación a las correspondientes distancias χ^2 verdaderas entre los perfiles fila del mapa de la imagen 6.3

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.

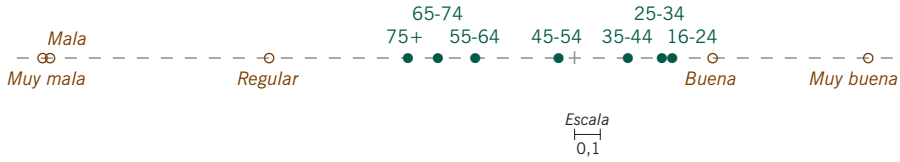


Imagen 6.5:
Mapa óptimo del mapa de la imagen 6.3, que muestra las proyecciones de los vértices de las categorías de salud

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

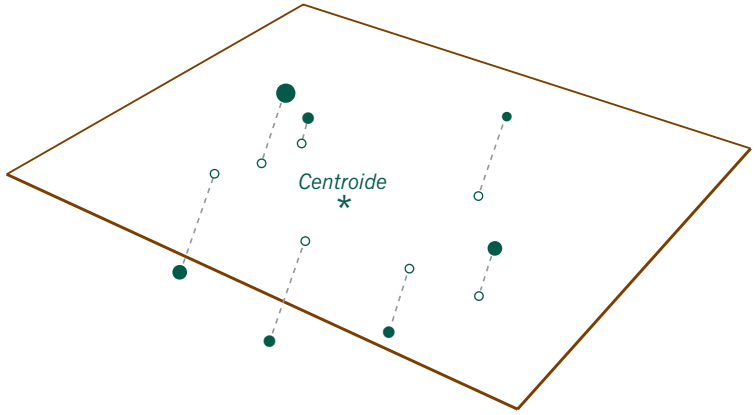


Imagen 6.6:
Perfiles en un espacio multidimensional y un plano que corta dicho espacio; el plano que mejor se ajuste en el sentido mínimo-cuadrático debe pasar por el centroide de los puntos (Los perfiles tienen masas diferentes tal como indican los tamaños de los puntos.)

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.

CAPÍTULO 7

Escalado óptimo

Hasta ahora hemos presentado el AC como un método geométrico de análisis de datos. Hemos destacado tres conceptos fundamentales: perfil, masa y distancia χ^2, y cuatro conceptos derivados: centroide (media ponderada), inercia, subespacio y proyección. Los perfiles son puntos multidimensionales, ponderados por masas. Medimos las distancias entre perfiles mediante distancias χ^2. Visualizamos los perfiles proyectándolos sobre el subespacio de pocas dimensiones que mejor se ajusta a los perfiles. A continuación, para su interpretación, proyectamos los vértices como puntos de referencia en dicho subespacio. De todas formas, existen muchas maneras distintas de definir y de interpretar el AC. Por ello, la misma metodología de base se ha redescubierto muchas veces en diferentes contextos. Una de estas metodologías alternativas es el *escalado óptimo*. Una discusión sobre esta aproximación nos permitirá profundizar en el conocimiento del AC.

Contenido

Cuantificación de un conjunto de categorías

Consideremos una vez más el ejemplo de la tabla de la imagen 6.1, la tabla de contingencia que cruza los grupos de edad con las categorías sobre la autopercepción de la salud. Tanto las variables fila como las variables columna son variables categóricas, que hemos guardado en un archivo de datos utilizando códigos de 1 a 7 para la edad, y de 1 a 5 para la salud. Si queremos calcular estadísticos como la media y la varianza o hacer un análisis de regresión en el que intervenga, por ejemplo, la variable autopercepción de la salud, necesitamos valores numéricos para cada categoría de salud. Si utilizamos los valores de 1 a 5, estamos asumiendo de forma implícita que la separación entre estas categorías es exactamente la unidad, lo que no tiene porqué ser cierto. El hecho de que hayamos ordenado las categorías de salud (la autopercepción de la salud es una variable categórica ordinal), justifica, en parte, que hayamos utilizado los valores de 1 a 5, pero, ¿que ocurriría si la variable fuera nominal, como, por ejemplo, la variable país que vimos en el capítulo 1 (imagen 1.3)? ¿Y si fuera el estado civil?* La variable grupo de edad es también una variable ordinal establecida mediante intervalos en la escala original de la edad, de manera que podemos razonablemente utilizar los puntos medios de cada intervalo de la edad como valores de escala. Sin embargo, no es claro el valor que hemos asignado al grupo de edad 7, que hemos dejado abierto (de 75 o más años). Cuando no haya una alternativa mejor, y las categorías presenten una ordenación natural, como en nuestro caso, en los cálculos utilizaremos por defecto valores enteros (en nuestro caso de 1 a 7 y de 1 a 5), que denominan una *escala entera*. Vamos a ver cómo el escalado óptimo ofrece un camino, resultante de un determinado criterio de optimización, que nos permite asignar valores numéricos a una variable categórica.

Cálculo de la media global utilizando una escala entera

Vamos a utilizar la escala entera para efectuar algunos cálculos simples. Sin embargo, primero invertiremos la codificación de las categorías de salud, de manera que el mayor valor corresponda a la mejor salud; así, 5 indicará *muy buena* salud, descendiendo hasta 1, que indicará *muy mala* salud. En la encuesta constituida por 6371 individuos, hay 817 individuos con *muy buena* salud (código 5), 3542 con *buena* salud (código 4), y así sucesivamente. Utilizando estos códigos enteros como escala para las categorías de salud, podemos calcular la *salud media* de la siguiente manera:

$$[(817 \times 5) + (3542 \times 4) + ... + (103 \times 1)] \,/\, 6371 = 3{,}72$$

Es decir,

$$(0{,}128 \times 5) + (0{,}556 \times 4) + ... + (0{,}016 \times 1) = 3{,}72 \qquad (7.1)$$

donde 817/6371 = 0,128, 3542/6371 = 0,556, etc. son los elementos del perfil fila medio (última fila de la tabla de la imagen 6.2). Por tanto, esta media de todos

* En mi experiencia como consultor en estadística, una vez me mostraron una encuesta con la variable «afiliación religiosa» tomando los valores: 0 = ninguna, 1 = católica, 2 = protestante, etc. ¡El investigador calculó la religión media de la muestra!

los encuestados no es más que el centroide de los valores de la escala obtenido ponderando con los elementos del perfil fila medio.

Cálculo de la media de los grupos de edad mediante una escala entera

Consideremos ahora un determinado grupo de edad, por ejemplo, el de 16 a 24 años. En la primera fila de datos de la tabla de la imagen 6.1, vemos que de los 1223 encuestados de este grupo, hay 243 individuos con *muy buena* salud, 789 con *buena*, y así sucesivamente. Utilizando otra vez los valores enteros de la escala de 5 a 1 para las categorías de la salud, la *salud media* de este grupo de 16-24 es:

$$[(243 \times 5) + (789 \times 4) + ... + (6 \times 1)]/1223 = 4{,}02$$

es decir

$$(0{,}199 \times 5) + (0{,}645 \times 4) + ... + (0{,}005 \times 1) = 4{,}02 \qquad (7.2)$$

donde, en la segunda línea, aparecen nuevamente los valores del perfil (del grupo de edad de 16 a 24 años), 243/1223 = 0,199, 789/1223 = 0,645, etc., que hemos utilizado como pesos. Vemos que el grupo de edad más joven tiene una autopercepción media de la salud mayor que la media general, 4,02 con relación a 3,72. Podríamos repetir el cálculo anterior para los restantes seis grupos de edad, obteniendo las medias siguientes:

16-24	25-34	35-44	45-54	55-64	65-74	75+	*Media global*
4,02	3,97	3,86	3,66	3,39	3,30	3,19	*3,72*

Cálculo de la varianza utilizando la escala entera

Ahora que ya hemos calculado las medias de las categorías de salud de cada grupo de edad, podemos calcular su varianza. Este cálculo es similar al cálculo de inercia del capítulo 4 ya que ponderaremos cada grupo de edad proporcionalmente al tamaño de su muestra. Otra posibilidad sería asignar a cada uno de los 6371 encuestados el valor correspondiente a su respectivo grupo de edad, y hacer el cálculo habitual de la varianza. Hemos calculado la varianza como (véase la fila de totales de la tabla de la imagen 6.1):

$$\frac{1223}{6371}(4{,}02-3{,}72)^2+\frac{1234}{6371}(3{,}97-3{,}72)^2+\ldots+\frac{396}{6371}(3{,}19-3{,}72)^2=0{,}0857$$

con desviación típica $\sqrt{0{,}0857}=0{,}293$.

Cálculo de las puntuaciones en una escala desconocida

Todos los cálculos anteriores dependen de la escala entera asignada a las categorías de salud, una elección arbitraria verdaderamente difícil de justificar, especialmente después de ver los resultados del capítulo 6. La pregunta es: ¿existe una escala más justificable o, al menos, más interesante? La respuesta depende de lo que entendamos por «más interesante». Vamos a considerar un posible criterio que

nos lleve a una escala con valores directamente relacionados con el AC. Supongamos que indicamos la escala asignada a las categorías de la salud por los valores desconocidos v_1, v_2, v_3, v_4 y v_5. La media de todos los encuestados sería, en función de estos valores desconocidos, igual que en (7.1):

$$\text{media global de la salud} = (0{,}128 \times v_1) + (0{,}556 \times v_2) + \ldots + (0{,}016 \times v_5) \qquad (7.3)$$

y la media del grupo de edad de 16 a 24 años sería, igual que en (7.2):

$$\text{media de salud 16-24 años} = (0{,}199 \times v_1) + (0{,}645 \times v_2) + \ldots + (0{,}005 \times v_5) \qquad (7.4)$$

Llamamos *puntuaciones* a las medias calculadas de esta manera, así (7.3) es la puntuación media y (7.4) es la puntuación del primer grupo de edad, que indicamos como s_1. Calcularíamos las puntuaciones, s_1, s_2, ..., s_7, de todos los grupos de edad de la misma manera, siempre en función de los valores desconocidos de la escala. Dado que cada uno de los 6371 encuestados pertenece a un grupo de edad, les podemos asignar la puntuación correspondiente de la escala de salud. Por ejemplo, a los 1223 encuestados del grupo de edad de 16 a 24 años les asignaremos la puntuación calculada en (7.4). Podríamos imaginar a los 6371 encuestados distribuidos en las siete puntuaciones de la escala de salud, independientemente de los valores que éstas tomen.

La maximización de la varianza proporciona la escala óptima

Para determinar los valores de la escala, propondremos que las 6371 puntuaciones cumplan determinadas propiedades. Una propiedad deseable sería que las puntuaciones estuvieran bien separadas entre sí, de manera que pudiéramos distinguir al máximo los grupos de edad. Dicho de otra forma, sería muy indeseable que las puntuaciones se encontraran muy cerca entre sí, de manera que nos fuera difícil distinguir entre grupos de edad en términos de sus categorías de salud. Una manera de expresar este requerimiento de forma más precisa es exigir que la varianza de las puntuaciones de los 6371 encuestados sea máxima. En términos numéricos, tenemos 1223 encuestados del primer grupo de edad (primera fila de la imagen 6.1) a los que hemos asignado la puntuación s_1, 1234 del segundo grupo de edad a los que hemos asignado la puntuación s_2, y así sucesivamente. Calcularemos la varianza de las 6371 puntuaciones, como hicimos en la página anterior. La *escala óptima* vendrá definida por los valores v_1, v_2, ..., v_5 que hagan que la varianza de las s_1, s_2, ..., s_7 puntuaciones sea máxima.

Los valores de la escala óptima de la dimensión del AC que mejor se ajusta

Afortunadamente, las posiciones de las categorías de salud, en la dimensión del AC que mejor se ajusta, resuelve de forma exacta este problema del escalado óptimo. La varianza máxima es igual a la inercia de esta dimensión óptima del AC. Es decir, los valores de las coordenadas de los vértices de la imagen 6.5 son los valores de la escala óptima de v_1 a v_5. A partir de los elementos de los perfiles de las filas podemos calcular sus correspondientes puntuaciones, de s_1 a s_7.

Categoría de la salud	*Coordenadas de vértices*
Muy buena	1,144
Buena	0,537
Regular	–1,188
Mala	–2,043
Muy mala	–2,076

Grupo de edad	*Coordenadas de perfiles*
16–24	0,371
25–34	0,330
35–44	0,199
45–54	–0,071
55–64	–0,396
65–74	–0,541
75+	–0,658

Imagen 7.1:
Valores de las coordenadas de los puntos de la imagen 6.5, es decir, las coordenadas de los vértices de las columnas y de los perfiles de las filas en la dimensión que mejor se ajusta a los perfiles de las filas

En la tabla de la imagen 7.1 se muestran los valores de las coordenadas de los vértices y los de las coordenadas de los perfiles. En el capítulo 3, vimos que la posición de un determinado grupo de edad corresponde al centroide de los vértices de los cinco grupos de la salud. Esta propiedad también se cumple para las proyecciones de los perfiles en cualquier subespacio. Así, por ejemplo, obtenemos la puntuación del grupo de edad de 16 a 24 años (imagen 6.2), ponderando las posiciones de los vértices de las cinco categorías de salud con los perfiles correspondientes de este grupo de edad (imagen 7.1), de la siguiente manera:

$$(0{,}199 \times 1{,}144) + (0{,}645 \times 0{,}537) + \ldots + (0{,}005 \times -2{,}076) = 0{,}371$$

lo que concuerda con la coordenada del perfil 16-24 de la imagen 7.1

También podríamos plantear el escalado óptimo al revés; es decir buscaríamos los valores de la escala de los grupos de edad que maximizaran la varianza de las categorías de salud. La solución viene dada por las coordenadas de los vértices de los cinco grupos de edad, siendo las coordenadas de los perfiles las puntuaciones de las categorías de la salud. En el siguiente capítulo veremos más a fondo la simetría existente entre el análisis de filas y el análisis de columnas. Esta simetría, o *dualidad*, del escalado óptimo ha llevado a algunos autores a denominar este método como *optimización dual de la escala.*

Interpretación de la escala óptima

A diferencia de la escala entera original, la escala óptima no sitúa las cinco categorías de salud a distancias iguales. En el mapa de la imagen 6.5 vimos que existía una gran diferencia entre *buena* y *regular*, y una diferencia muy pequeña entre *mala* y *muy mala*. Estos valores de la escala óptima de las categorías de la salud hacen que las puntuciones de los grupos de salud estén lo más separadas posible según el criterio de la varianza. Es decir, utilizando la escala óptima de las categorías de salud, obtenemos la máxima discriminación entre los grupos de edad. En el mapa de la imagen 6.3, en la que sólo representamos las puntuaciones de los

grupos de edad, vemos que hasta el grupo de edad de 34 a 45 años, existen pequeños cambios en la autopercepción de la salud, luego vemos grandes cambios en los grupos de edad media, especialmente entre los grupos de 45 a 54 y de 55 a 64 años, y finalmente cambios más pequeños en los grupos de mayor edad. Revisando otra vez los perfiles de la tabla de la imagen 6.2, podemos comprobar que entre los grupos de edad de 45 a 54 años y de 55 a 64 años se produce una caída de aproximadamente el 50% en la categoría *muy buena* y un incremento de más del doble en la categoría *mala*. Estos cambios en los valores de los perfiles explican los cambios observados en las puntuaciones.

Condiciones de identificación de una escala óptima

Los valores de la escala óptima obtenidos para las categorías de la salud son 1,144, 0,537, –1,188, –2,043 y –2,076, respectivamente (imagen 7.1). Hemos calculado estos valores en determinadas restricciones, necesarias para poder hallar una sola solución. Estas restricciones son que, para los 6371 encuestados, la media de los valores de la escala de salud sea 0 y que su varianza sea 1:

$$(0{,}128 \times 1{,}144) + (0{,}556 \times 0{,}537) + \ldots + (0{,}016 \times (-2{,}076)) = 0 \qquad \text{(media 0)}$$

$$(0{,}128 \times 1{,}144^2) + (0{,}556 \times 0{,}537^2) + \ldots + (0{,}016 \times (-2{,}076)^2) = 1 \qquad \text{(varianza 1)}$$

Estos prerrequisitos para los valores de la escala son las *condiciones de identificación* o *restricciones* en el lenguaje utilizado en la teoría matemática de optimización. La primera condición es necesaria, ya que podrían existir dos escalas distintas que tuvieran la misma varianza pero medias diferentes. Es decir, sería imposible identificar una solución sin especificar la media. La segunda condición es necesaria ya que, si multiplicamos de forma arbitraria los valores de la escala por un valor grande, la varianza de las eventuales puntuaciones se incrementaría mucho, lo que no tendría sentido alguno, pues estamos intentando maximizar la varianza. En consecuencia, es necesario que busquemos una escala que tengan una determinada media y un determinado rango de variación. Aunque las condiciones de «media 0 y varianza 1» son una elección arbitraria, conducen a unas coordenadas adecuadas para los vértices del AC, que también cumplen estas condiciones.

Cualquier transformación lineal de la escala sigue dando una escala óptima

Para determinar la escala óptima, las dos condiciones de identificación que hemos descrito anteriormente son simples instrumentos técnicos que aseguran una sola solución matemática de nuestro problema. Sin embargo, una vez obtenidos los valores de la escala, tenemos la posibilidad de transformarlos en una escala más conveniente, siempre y cuando recordemos que la media y la varianza de la escala transformada no tienen ninguna trascendencia sustantiva o relevancia estadística. En general, llevamos a cabo la redefinición de esta escala, fijando los puntos extremos, de manera que tengan valores con algún significado. Por ejemplo, en este caso, podríamos dar el valor 0 a la categoría *muy mala* salud, y el va-

Categoría de la salud	*Valor en la escala óptima*	*Valor en la escala transformada*
Muy buena	1,144	100,0
Buena	0,537	81,1
Regular	–1,188	27,6
Mala	–2,043	1,0
Muy mala	–2,076	0,0

Imagen 7.2: *Valores de la escala óptima del AC y valores transformados para que la escala esté entre 0 y 100*

lor 100 a la de *muy buena*. En tal caso, necesitamos hacer que una transformación asigne el valor 0 a –2,076 y el valor 100 a 1,144. Para ello, en primer lugar, podemos sumar 2,076 a todos los valores de la escala, de manera que el valor más pequeño sea 0. Ahora la escala va de 0 a 1,144 + 2,076 = 3,220. Para asignar 100 al mayor valor de la escala, podemos multiplicar todos los valores por 100/3,220. En este caso en concreto la fórmula de cálculo para pasar de la antigua a la nueva escala es simplemente:

$$nueva = (antigua + 2{,}076) \times \frac{100}{3{,}220}$$

o, para el caso general:

$$nueva = \left[(antigua - \text{antiguo límite inferior}) \times \frac{\text{rango nuevo}}{\text{rango antiguo}} \right] + \text{nuevo límite inferior} \quad (7.5)$$

(en nuestro ejemplo el nuevo límite inferior es 0). Aplicando esta fórmula a los cinco valores de la escala óptima, obtenemos los valores transformados de la imagen 7.2.

La escala anterior de 5 a 1, con cuatro intervalos iguales entre los puntos de la escala, tendría los valores 100, 75, 50, 25, 0 en la escala transformada de rango 100 (recordemos que hemos invertido la escala de manera que *muy buena* sea 100). Sin embargo, en la escala óptima transformada, *regular* no se halla en el punto medio (50) de la escala, se halla mucho más cerca del extremo «mala» salud de la escala.

La escala óptima no es única

Debemos insistir en que la escala óptima depende del criterio establecido para su determinación, así como de las condiciones de identificación escogidas. Aparte de los criterios puramente técnicos, también depende, de forma clara, de la tabla de contingencia original. Si tuviéramos una tabla de contingencia que cruzara autopercepción de la salud con otra variable demográfica, por ejemplo, nivel de educación, obtendríamos, una escala óptima distinta para las categorías de la salud, ya que ahora el procedimiento discriminaría de manera óptima las diferencias entre niveles de educación.

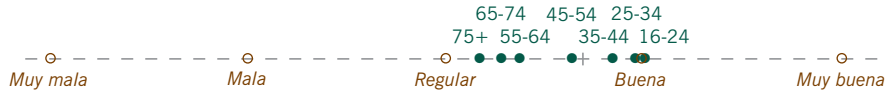
Un criterio basado en las distancias entre las filas y las columnas

En contraste con el criterio de maximización que hemos descrito anteriormente, a continuación presentamos un criterio de minimización para hallar los valores de la escala óptima que también conduce a la solución del AC. Este criterio se basa en las distancias entre las filas y las columnas —en el ejemplo que nos ocupa estas distancias serán las distancias entre las categorías de salud y los grupos de edad—. Imaginemos, en primer lugar, las categorías de salud en una determinada escala, por ejemplo, la escala entera de 1 a 5, de *muy mala* a *muy buena* salud, que representamos gráficamente en la imagen 7.3. Ahora el objetivo es hallar, en la misma escala, las posiciones de los grupos de edad que se hallen tan «próximos» como sea posible a las categorías de salud, en el sentido de que un grupo de edad, que tenga una frecuencia elevada para una determinada categoría de salud, tienda a aproximarse a esta categoría. Supongamos ahora que los valores de las categorías de salud (en este ejemplo inicial eran los valores de 1 a 5) son los valores h_1, h_2, ..., h_5 y que los valores de los grupos de edad son a_1, a_2, ..., a_7. La distancia entre un grupo de edad y una categoría de salud es igual al valor absoluto de la diferencia $|a_i - h_j|$; sin embargo, preferimos, como medida de proximidad, la distancia al cuadrado $(a_i - h_j)^2$.* Para que las distancias dependan de las frecuencias de ocurrencia de las tablas de contingencia, ponderaremos cada distancia al cuadrado con p_{ij}, la frecuencia relativa como la definimos en la página 51 del capítulo 4. Es decir, los valores de la imagen 6.1 divididos por la suma total 6371 (por tanto, la suma de todos los p_{ij} es 1). Nuestro objetivo sería entonces minimizar la función siguiente:

$$\sum_i \sum_j p_{ij} d_{ij}^2 = \sum_i p_{ij} (a_i - h_j)^2 \qquad (7.6)$$

que tenderá a acortar las distancias cuando p_{ij} sea mayor. Dados unos determinados valores h_j de las categorías de salud, es fácil demostrar que obtenemos un mínimo de (7.6) con las medias ponderadas de los grupos de edad. Para los valores de las categorías de salud de 1 a 5, estas medias ponderadas son las puntuaciones que calculamos antes —en la fórmula (7.2) y las puntuaciones que le siguen—, que también hemos representado en el mapa de la imagen 7.3. Las dos escalas que mostramos en el mapa de la imagen 7.3 minimizan (7.6), pero si los valores de la escala de la salud fueran otros, ¿cúal sería el valor del mínimo? Para poder

Imagen 7.3:
La escala 1-5 de las categorías de salud y las medias ponderadas de los grupos de edad



* De nuevo, como anteriormente, es siempre más fácil trabajar con las distancias al cuadrado —la raíz cuadrada de las expresiones que conducen a distancias euclídeas causan muchos problemas de optimización—; estas dificultades desaparecen cuando consideramos la optimización mínimo-cuadrática.

responder con sentido a esta pregunta, necesitamos, otra vez, definir unas condiciones de identificación; en caso contrario podríamos llegar a una solución que situara a todas las categorías de salud en el mismo punto. Si consideramos las mismas condiciones de identificación que vimos anteriormente para los valores de las categorías de la salud, es decir, media 0 y varianza 1, obtendremos, de nuevo, el mínimo con la dimensión óptima del AC. Comparando las posiciones de los grupos de edad en el mapa de la imagen 7.3 con las posiciones óptimas en el mapa de la imagen 6.5, vemos que la dispersión de los grupos de edad es mayor en el mapa 6.5, lo que significa que en el mapa 7.3, los grupos de edad quedan más cerca de las categorías de salud en términos del criterio (7.6). El valor del mínimo alcanzado en el mapa de la imagen 6.5 es igual a 1 menos la varianza (maximizada) de la dimensión óptima del AC, lo que llamamos *pérdida de homogeneidad*. (Volveremos a este concepto en el capítulo 20, cuando tratemos sobre el análisis de homogeneidad.) El criterio (7.6) lo podemos generalizar fácilmente a dos o más dimensiones, digamos K dimensiones, simplemente sustituyendo a_i y h_j por vectores con K elementos y sustituyendo los cuadrados de las diferencias $(a_i - h_j)^2$ por el cuadrado de las distancias euclídeas en un espacio de dimensión K.

RESUMEN:
Escalado óptimo

1. El *escalado óptimo* asigna valores a las categorías (o atributos) de una variable categórica mediante algún criterio de optimización que separe, o discrimine, los grupos de casos que hemos formado al cruzar los casos con dicha variable.
2. Las posiciones de las categorías son los vértices en la dimensión óptima del AC que proporcionan unos valores de escala óptimos, en el sentido de que maximizan la varianza entre los grupos. Las puntuaciones de los grupos son las proyecciones de sus perfiles sobre esta dimensión. La varianza máxima de las puntuaciones es igual a la inercia de las proyecciones de los perfiles.
3. Es característico de la geometría del AC que las posiciones de las coordenadas de las proyecciones de las categorías sobre la dimensión óptima estén estandarizadas. De todas formas, en la práctica podemos recentrar y redimensionar los valores de la escala, por ejemplo, para que sus valores vayan de 0 a 1 o de 0 a 100. En tal caso variarán los valores de la media y la varianza.
4. También podemos hallar la escala óptima a partir de un criterio basado en las distancias entre filas y columnas. Concretamente, se trata de situar en el mapa las coordenadas de filas y columnas de manera que se minimicen las distancias ponderadas entre filas y columnas —ponderadas con las frecuencias relativas obtenidas de la tabla de contingencia—. Este valor mínimo es igual a 1 menos la varianza (máxima) de las puntuaciones en la escala óptima.

CAPÍTULO 8

Simetría entre el análisis de filas y el de columnas

En todos los ejemplos y análisis que hemos mostrado hasta ahora, nos hemos centrado en el análisis de las filas de una tabla. Hemos visualizado e interpretado las posiciones de los perfiles fila, utilizando las columnas como puntos de referencia, es el «análisis de filas». Sin embargo, podemos aplicar el análisis anterior de forma completamente simétrica a las columnas de la tabla. Lo podemos ver como una transposición de la tabla, en la que intercambiamos las filas por las columnas y viceversa, para a continuación repetir los procedimientos que hemos descrito del capítulo 2 al 7. En este capítulo, veremos que los análisis de filas y columnas están muy relacionados. En realidad, si llevamos a cabo el análisis de filas, también estamos efectuando el análisis de columnas y viceversa. Por tanto, podemos ver el AC como un análisis simultáneo de las filas y de las columnas de una tabla.

Contenido

Resumen del análisis de las filas

Consideremos de nuevo los datos de la tabla de la imagen 6.1 sobre la autopercepción de la salud. En el capítulo 6 hicimos el análisis de filas de estos datos porque queríamos representar los perfiles de los grupos de edad con relación a las categorías de salud. Estos siete perfiles se hallaban en un espacio tetradimensional, delimitado por los cinco vértices que representan los perfiles unidad extremos de cada una de las categorías de salud. Llegamos a la conclusión que la mayor parte de la variación espacial de los perfiles se producía en una recta (imagen 6.3). Finalmente, proyectamos e interpretamos las proyecciones de los perfiles y de los cinco vértices sobre la mencionada recta (imagen 6.5).

Análisis de columnas: los valores de los perfiles tienen una interpretación simétrica

Consideremos ahora la posibilidad de analizar los perfiles columna de la tabla de la imagen 6.1. Es decir, los perfiles de las categorías de salud con relación a los grupos de edad que mostramos en la tabla de la imagen 8.1. Para cada categoría de salud, los perfiles columna proporcionan porcentajes de individuos con relación a los grupos de edad. Por ejemplo, en la categoría *mala* salud, el 4,3% de los individuos tiene de 16 a 24 años, el 8,5% de 25 a 34 años, y así sucesivamente. A pesar de que la tabla de perfiles columna tiene un aspecto completamente distinto que el de la tabla de perfiles fila de la imagen 6.2, cuando nos fijamos en valores concretos y los comparamos con sus medias, vemos que contienen la misma información (en el capítulo 2, con los datos sobre mis viajes, ya nos dimos cuenta de ello). Consideremos, por ejemplo, el 23,7% de la columna *mala* del grupo de edad de 65 a 74 años. Comparemos este valor con el porcentaje de individuos de ese grupo de edad para el total de la muestra, que podemos encontrar en la última columna: el 11,2%. Llegamos a la conclusión que, en el grupo de edad de 65 a 74 años, algo más del doble de los encuestados manifiestan que su salud es *mala* en comparación con la media global de ese grupo de edad (el cociente es 23,7/11,2 = 2,1). Si nos fijamos ahora en la misma celda de la tabla de la imagen 6.2, vemos que el 13,7% del grupo de 65 a 74 años manifiesta que su salud es *mala*, mientras que esta proproción en el total de la muestra es del 6,5% (última fila de la tabla de la imagen 6.2). De nuevo, llegamos a la conclusión de que, en este grupo de edad, algo más del doble de los

Imagen 8.1:
Perfiles de la columna de las categorías de salud con relación a los grupos de edad, expresados como porcentajes

Grupo de edad	*Muy buena*	*Buena*	*Regular*	*Mala*	*Muy mala*	*Media*
16–24	29,7	22,3	11,2	4,3	5,8	*19,2*
25–34	26,9	22,8	11,0	8,5	5,8	*19,4*
35–44	18,0	18,6	12,1	9,9	7,8	*16,2*
45–54	11,0	13,2	15,8	12,1	15,5	*13,5*
55–64	6,5	11,7	20,5	25,6	29,1	*14,3*
65–74	5,4	7,5	19,0	23,7	19,4	*11,2*
75+	2,4	3,8	10,5	15,9	16,5	*6,2*
Suma	*100*	*100*	*100*	*100*	*100*	*100*

encuestados manifiesta que su salud es *mala*, en comparación con la media global de esta categoría de salud, el cociente es idéntico: 13,7/6,5 = 2,1.

Análisis de columnas: la misma inercia total

En el capítulo 4 vimos que la inercia total de los perfiles columna es igual a la inercia total de los perfiles fila; ambos cálculos son tan sólo maneras alternativas de expresar la misma fórmula: el estadístico χ^2 dividido por el tamaño de la muestra. Para los datos sobre la autopercepción de la salud, la inercia total es de 0,1404.

Análisis de columnas: igual dimensionalidad

Los perfiles columna definen una nube de cinco puntos, cada uno de ellos con siete componentes, que deben hallarse en un espacio de seis dimensiones, ya que la suma de sus componentes es 1. Sin embargo, los cinco puntos no llegan a ocupar las seis dimensiones de este espacio, pues sólo ocupan cuatro. Podemos percibir este hecho de forma intuitiva, si tenemos en cuenta que dos puntos se hallan exactamente en una recta unidimensional, tres puntos en un plano bidimensional, cuatro puntos en un espacio tridimensional y, por tanto, cinco puntos se hallarán en un espacio tetradimensional. En consecuencia, aunque los perfiles fila y los perfiles columna se encuentran en espacios distintos, la dimensionalidad de estas dos nubes de puntos es la misma, en este caso, de cuatro. Se trata de la primera coincidencia geométrica entre el análisis de los perfiles fila y el de los perfiles columna. Pronto veremos muchas más similitudes.

Análisis de columnas: la misma aproximación para reducir la dimensionalidad

Consideremos todavía los perfiles de las cinco categorías de salud en un espacio tetradimensional. Nos planteamos ahora las mismas preguntas que antes: ¿se pueden representar, de forma aproximada, estos puntos en un subespacio de pocas dimensiones?, ¿cuál es la calidad de esa aproximación? Haciendo el mismo tipo de cálculos matemáticos que vimos en el capítulo 6, llegamos a que podemos representar los perfiles columna en un espacio unidimensional y que la calidad de la representación es del 97,3%, exactamente el mismo porcentaje que hallamos con los perfiles fila. Estamos ante una segunda coincidencia geométrica entre los dos análisis.

Análisis de columnas: los mismos valores de coordenadas pero redimensionados

En el mapa de la imagen 8.2 representamos las proyecciones de los perfiles columna sobre la recta que mejor se ajusta a los perfiles. Vemos que las categorías de salud se sitúan exactamente en el mismo orden que los vértices en el mapa de la imagen 6.5. Aunque los valores de sus coordenadas no son iguales, sus posiciones relativas son idénticas. Comparando las posiciones de las categorías de salud

Imagen 8.2:
Mapa unidimensional óptimo de los perfiles de las categorías de salud

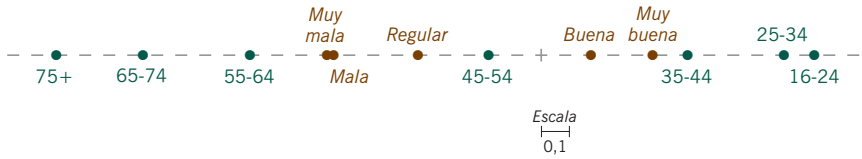


Imagen 8.3:
El mismo mapa de la imagen 8.2, que muestra las posiciones de las proyecciones de los vértices de los grupos de edad

del mapa de la imagen 8.2 con la escala del mapa de la imagen 6.5, vemos que las coordenadas de los perfiles son una versión encogida, o contraída, de las posiciones de los vértices. Pronto interpretaremos de forma específica este «factor de contracción». Sin embargo, profundicemos un poco más. En el mapa de la imagen 8.3 representamos las proyecciones de los siete vértices que corresponden a los grupos de edad sobre la misma recta. La comparación de las posiciones de los vértices aquí con la de los perfiles de los grupos de edad en el mapa de la imagen 6.5 (o con las posiciones del mapa de la imagen 6.3 en una escala mayor) pone de manifiesto el mismo fenómeno, pero para las filas; en el mapa de la imagen 6.5, las posiciones de los perfiles fila con relación a la recta que mejor se ajusta son una versión encogida de las posiciones de los vértices de los grupos de edad proyectados sobre la recta, que mejor se ajusta a los perfiles de las categorías de salud del mapa de la imagen 8.2. Es decir, en el análisis de columnas, las posiciones de los vértices fila son una expansión de las posiciones de los perfiles fila del análisis de filas. Estamos ante la tercera, y más importante, coincidencia geométrica entre los dos análisis.

Ejes e inercias principales

En estos análisis, la recta que mejor se ajusta se denomina *eje principal*. En los próximos capítulos veremos que existen otros ejes principales. Por ello, de forma más precisa, llamaremos a esta recta «primer eje principal». Hemos visto que tanto en el análisis de filas, como en el análisis de columnas la inercia total es de 0,1404 y que, en ambos casos, el porcentaje de inercia explicada por el primer eje es del 97,3%. También en ambos casos, el valor concreto de la inercia explicada por el primer eje es de 0,1366, por tanto, el porcentaje de inercia explicada es igual a $100 \times 0{,}1366/0{,}1404 = 97{,}3\%$. Llamamos *inercia principal* a la inercia explicada por un eje principal (de 0,1366 en este caso). En este ejemplo se trata de la primera inercia principal ya que nos referimos a la inercia del primer eje principal. La inercia principal también recibe el nombre de *valor propio*, ya que se puede calcular como un valor propio de una matriz cuadrada simétrica.

El factor de escala es la raíz cuadrada de la inercia principal

Parece pues, que tenemos que hacer un solo análisis: de filas o de columnas. Los resultados de uno de ellos se pueden obtener de los resultados del otro. Sin embargo, ¿cuál es exactamente la relación entre ambos? Dicho de otra manera, ¿cuál es el factor de escala que nos permite pasar de las posiciones de los vértices de un análisis a las posiciones de los perfiles del otro? Pues bien, este factor de escala es igual a la raíz cuadrada de la inercia principal. Así, en este ejemplo,

Categoría de la salud	*Coordenadas de perfiles*
Muy buena	0,423
Buena	0,198
Regular	–0,439
Mala	–0,755
Muy mala	–0,767

Grupos de edad	*Coordenadas de vértices*
16–24	1,004
25–34	0,893
35–44	0,538
45–54	–0,192
55–64	–1,070
65–74	–1,463
75+	–1,782

Imagen 8.4:
Valores de las coordenadas de los puntos del mapa de la imagen 8.2, es decir las coordenadas de los perfiles columna y de los vértices de las filas en el primer eje principal de los perfiles columna (compárese con las tablas de la imagen 7.1)

es $\sqrt{0{,}1366} = 0{,}3696$. En consecuencia, para pasar de los vértices fila de la imagen 8.3, a los perfiles fila de los mapas de las imágenes 6.3 o 6.5, simplemente multiplicamos los valores de las coordenadas por 0,3696, es decir, algo más de un tercio. A la inversa, para pasar de los perfiles columna del mapa de la imagen 8.3 a los vértices columna del mapa de la imagen 6.5, multiplicamos los valores de las coordenadas por el inverso de este valor, concretamente por $1/0{,}3696 = 2{,}706$. En las tablas de las imágenes 7.1 y 8.4 se muestran todos los valores numéricos de las coordenadas de los perfiles y de las coordenadas de los vértices. Comparando los valores de ambas imágenes llegamos a la expresión:

$$\text{Coordenada del perfil} = \text{coordenadas del vértice} \times \sqrt{\text{inercia principal}}$$

Fijémonos que en los mapas de las imágenes 6.5 y 8.3, los perfiles están más juntos que los vértices. El factor de escala es una medida directa de lo apretados que están los perfiles «interiores» en comparación con los vértices «exteriores». En este caso, un factor de escala de 0,3696 indica que la dispersión de los perfiles es aproximadamente un tercio de la de los vértices. Al final del capítulo 4 interpretamos la inercia total como una medida de la dispersión de los perfiles en relación a los vértices exteriores (imagen 4.2). Las inercias principales (o sus raíces cuadradas) son también medidas de dispersión, pero se refieren a los ejes principales de forma individual, no al espacio de perfiles en su conjunto. Cuanto mayor sea la inercia principal, y en consecuencia cuanto mayor sea el factor de escala, mayor será la dispersión de los perfiles con relación a los vértices en el eje principal. En consecuencia, es obvio que la inercia principal no puede ser mayor que 1, pues los perfiles deben hallarse en el «interior» de sus correspondientes vértices.

La correlación como una interpretación de la inercia principal

La raíz cuadrada de la inercia principal, que ya hemos señalado que siempre toma un valor menor de 1, tiene otra interpretación como coeficiente de correlación. En general, los coeficientes de correlación se calculan entre pares de medidas, como por ejemplo la correlación entre los ingresos y la edad. En el caso que nos ocupa, para cada encuestado tenemos dos observaciones —el grupo de

edad y la categoría de la salud—, pero se trata de observaciones categóricas, no de medidas. Podemos calcular el coeficiente de correlación entre estas dos variables recurriendo a los códigos enteros que utilizamos anteriormente por defecto, es decir de 1 a 7 para los grupos de edad y de 1 a 5 para las categorías de salud. En tal caso obtenemos una correlación de 0,3456. Utilizando otros valores, obtendríamos otras correlaciones. Por tanto, nos podemos plantear las siguientes preguntas: para obtener la máxima correlación, ¿qué valores debemos utilizar para los grupos de edad?, ¿y para las categorías de salud? Llamamos *correlación canónica* a la correlación máxima que obtenemos de esta manera. En este ejemplo, la correlación canónica es de 0,3696, exactamente la raíz cuadrada de la inercia, es decir, el factor de escala que vincula el análisis de filas con el de columnas. Los valores numéricos de los grupos de edad y de las categorías de salud que dan la máxima correlación son precisamente los valores de las coordenadas de los grupos de edad y de las categorías de salud en el eje principal del AC que aparecen en las imágenes 7.1 y 8.4 y que representamos gráficamente en las imágenes 6.3, 6.5, 8.2 y 8.3. Podemos utilizar las coordenadas de los perfiles o las coordenadas de los vértices, ya que la correlación no se ve afectada ni por un cambio de origen ni por redimensionamiento de las escalas. Sin embargo, en general, utilizamos escalas estandarizadas de media 0 y varianza 1.

Representación gráfica de la correlación

Es habitual mostrar gráficamente la correlación entre dos variables en un diagrama de dispersión de los casos, como por ejemplo los grupos de edad (eje *y*) y categorías de la salud (eje *x*). En este diagrama de dispersión tenemos 6371 casos, sin embargo, sólo aparecen 7 valores en el eje *y*, y 5 en el *x*. Por tanto, en el diagrama de dispersión tenemos sólo $7 \times 5 = 35$ posibles puntos (imagen 8.5). En cada punto hallamos todos los casos de la categoría de la salud y del grupo de edad de la celda correspondiente de la tabla de contingencia original (imagen 6.1). Aquí hemos representado los puntos como cuadrados de área proporcional a la frecuencia de la celda. En este diagrama de dispersión, la correlación canónica de los 6371 individuos, es igual a la correlación de Pearson. Cuando decimos que la correlación canónica es óptima queremos decir que no existen otros valores de las categorías de las filas y de las categorías de las columnas que proporcionen un coeficiente de correlación mayor. Obtendríamos una correlación canónica igual a 1 cuando todos los puntos se hallaran en una recta, en este caso significaría que cada grupo de edad está asociado sólo con una categoría de salud (los perfiles serían todos perfiles unidad, es decir, vértices).

Coordenadas principales y coordenadas estándares

En esta etapa es conveniente introducir algo de terminología para evitar tener que repetir continuamente las expresiones «coordenadas de las posiciones de los vértices» y «coordenadas de las posiciones de los perfiles». Hemos estandarizado las primeras para que tengan media 0 y varianza 1, y las llamaremos *coordenadas*

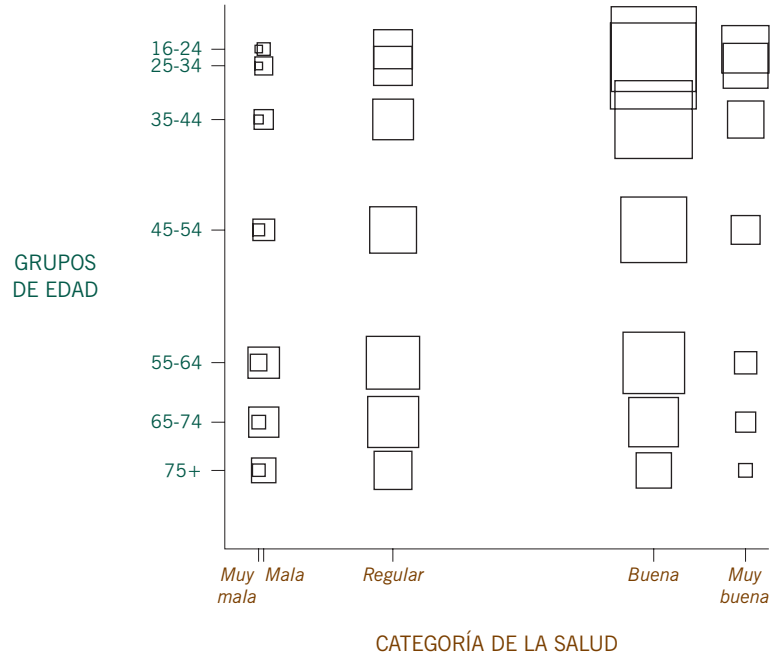


Imagen 8.5:
Diagrama de dispersión de los valores que maximizan la correlación entre las categorías de salud y los grupos de edad; los cuadrados correspondientes a cada combinación de valores tienen un área proporcional al número de individuos. La correlación es igual a 0,3456

estándares; las segundas son las coordenadas de los perfiles en los ejes principales, y las denominaremos *coordenadas principales*. Por ejemplo, en las tablas de la imagen 8.4, la primera columna de resultados numéricos contiene las coordenadas principales de las categorías de la salud, mientras que la segunda columna contiene las coordenadas estándar de los grupos de edad. En ambos casos son coordenadas en el primer eje principal del AC. Veremos que en capítulos posteriores tendremos, en general, más de un eje principal.

Maximización del cuadrado de las correlaciones con la media

Vamos a ver otra manera de calcular la correlación en el AC. Así, a cada uno de los 6371 individuos del ejemplo sobre la encuesta de salud le asignemos un par de valores: uno (a_i) para el grupo de edad y otro (h_j) para la categoría de salud. Como antes, estos valores son desconocidos, sin embargo, vamos a definir un criterio de optimización que nos permita determinarlos. Supongamos que cada individuo tiene una puntuación igual a la suma de los dos valores $a_i + h_j$. Por ejemplo, alguien del grupo de edad de 25 a 34 años con *muy buena* salud (segundo grupo de edad y primera categoría de salud) tendría una puntuación igual a $a_2 + h_1$. Supongamos ahora que indicamos la correlación entre todos los pares de valores $\{a_i, a_i + h_j\}$ como cor$(a, a + h)$, donde a y h indican los 6371 valores de la muestra. De forma similar, indicamos la correlación entre todos los pares $\{h_j, a_i + h_j\}$ como cor$(h, a + h)$. Habría que buscar unas escalas que optimizaran estas dos correlaciones. Se puede demostrar que la primera dimensión del AC

proporciona unos valores que son óptimos en el sentido de que maximizan la media de los cuadrados de estas correlaciones:

$$\text{media de los cuadrados de las correlaciones} = \frac{1}{2}[\text{cor}^2(a, a+h) + \text{cor}^2(h, a+h)] \quad (8.1)$$

Dado que para cualquier par de variables estandarizadas X e Y, la $\text{cor}(X, X+Y) = \sqrt{[1+\text{cor}(X,Y)]/2}$, en (8.1) la media de los cuadrados de las correlaciones será igual a:

$$\text{media de los cuadrados de las correlaciones} = \frac{1+\text{cor}(a,h)}{2} \quad (8.2)$$

En consecuencia, cuando con el AC maximizamos la $\text{cor}(a,h)$, es decir, obtenemos la correlación canónica (8.2), también maximizamos (8.1). Este resultado nos será útil más tarde ya que lo podemos generalizar fácilmente a más de dos variables, como veremos en el capítulo 20.

Minimización de la pérdida de homogeneidad entre variables

Utilizando la notación anterior, podemos ver otro criterio de optimización que también nos conduce a los resultados del AC. En primer lugar, en vez de calcular las sumas de los valores de cada individuo, calculemos la media de estos valores, $\frac{1}{2}(a_i + h_j)$. A continuación calculemos las diferencias entre los valores de la edad y de salud de cada individuo con su media: $a_i - \frac{1}{2}(a_i + h_j)$ y $h_j - \frac{1}{2}(a_i + h_j)$. Una medida de la similitud entre los valores de edad y los de la salud de cada individuo es la media de la suma de cuadrados de estas dos diferencias, lo que nos lleva a una medida de la varianza de los valores a_i y h_j:

$$\text{varianza (de un caso)} = \frac{1}{2}\left([a_i - \frac{1}{2}(a_i + h_j)]^2 + [h_j - \frac{1}{2}(a_i + h_j)]^2\right) \quad (8.3)$$

Sin embargo, en este contexto, preferimos el término *homogeneidad* porque si los valores a_i y h_j fueran iguales, su varianza sería cero; a un individuo con esta combinación de categorías le llamamos individuo *homogéneo*. Un término alternativo a homogeneidad es el de *consistencia interna*. Calculando la media de los valores (8.3) para todos los individuos de la muestra, obtenemos un valor llamado *pérdida de homogeneidad* (en la página 69, se usa este término en el mismo sentido). Si todos los valores de edad coincidieran con los de salud, la pérdida de homogeneidad sería cero, es decir la muestra sería completamente homogénea (o internamente consistente). El objetivo del AC es hallar una escala de valores que minimice esta pérdida. Una vez más, los valores que minimizan la pérdida de homogeneidad coinciden con las coordenadas de la edad y de salud de la primera dimensión del AC. Como veremos en el capítulo 20, podemos fácilmente extender esta definición a más de dos variables.

RESUMEN:
Simetría entre el análisis de filas y el de columnas

1. Todo lo que hemos hecho en el análisis de filas lo podemos aplicar de forma completamente simétrica a las columnas, como si repitiéramos todas las operaciones en la tabla transpuesta.
2. Con el análisis de columnas visualizamos los perfiles de las columnas y los vértices de las filas en el subespacio de representación óptimo de los perfiles de las columnas.
3. El (primer) *eje principal* de perfiles es la recta, o dimensión, que mejor se ajusta y la (primera) *inercia principal* es la inercia explicada por esta dimensión.
4. Las *coordenadas principales* son las posiciones de las coordenadas de los perfiles en un eje principal, y las *coordenadas estándares* son las posiciones de las coordenadas de los vértices en un eje principal.
5. Los dos análisis son equivalentes en el sentido de que tienen la misma inercia total, la misma dimensionalidad y la misma descomposición de la inercia total en inercias de los ejes principales.
6. Además, en ambos análisis, los perfiles y los vértices están íntimamente relacionados de la siguiente manera: en un eje principal, las posiciones de los perfiles (en coordenadas principales) tienen exactamente las mismas posiciones relativas que los correspondientes vértices (en coordenadas estándares) en el otro análisis, pero con valores contraídos. El factor de escala implicado es exactamente la raíz cuadrada de la inercia principal de ese eje.
7. Este factor de escala también se puede interpretar como una *correlación canónica*, especialmente cuando nos referimos al primer eje principal. Se trata de la máxima correlación que podemos obtener con las variables fila y las variables columna como resultado de la asignación de valores numéricos a las categorías de estas variables.

CAPÍTULO 9

Representaciones bidimensionales

Hasta ahora hemos estudiado con bastante extensión, las proyecciones de una nube de perfiles sobre un solo eje principal, la recta que mejor se ajusta. Sin embargo, en la práctica encontraremos que la mayor parte de representaciones del AC son bidimensionales. Es habitual representar el primer eje principal horizontalmente (eje x) y el segundo eje principal verticalmente (eje y). Aunque podemos proyectar la nube de puntos sobre cualquier subespacio de pocas dimensiones, las proyecciones bidimensionales son especiales debido a que son nuestra forma habitual de representar gráficos sobre papel o en las pantallas de ordenador. De todas formas, en el apéndice de cálculo mostramos algunos ejemplos de cómo utilizar el lenguaje de programación R para hacer representaciones en tres dimensiones (imagen B.5, en pág. 306).

Contenido

Conjunto de datos 4: hábitos fumadores de grupos de empleados

El ejemplo que veremos a continuación, que apareció originalmente en mi libro de 1984, *Theory and Applications of Correspondence Analysis*, ha sido utilizado como ejemplo para ilustrar el AC en los principales programas estadísticos comerciales. Este ejemplo, a pesar de que corresponde a un conjunto de datos ficticios, se ha citado en bastantes artículos científicos y todavía lo podemos utilizar como intro-

Imagen 9.1: *Clasificación de los empleados de una empresa según su nivel profesional y sus hábitos fumadores, que muestra los perfiles de las filas, el perfil fila medio, entre paréntesis, así como las masas de las filas*

	Tipos de fumadores					
Grupo de empleados	*No fumadores*	*Fumadores ocasionales*	*Fumadores medios*	*Fumadores compulsivos*	*Total de las filas*	*Masas*
Directivos séniors	4	2	3	2	*11*	*0,057*
DS	(0,364)	(0,182)	(0,273)	(0,182)		
Directivos jóvenes	4	3	7	4	*18*	*0,093*
DJ	(0,222)	(0,167)	(0,389)	(0,222)		
Empleados séniors	25	10	12	4	*51*	*0,279*
ES	(0,490)	(0,196)	(0,235)	(0,078)		
Empleados jóvenes	18	24	33	13	*88*	*0,456*
EJ	(0,205)	(0,273)	(0,375)	(0,148)		
Secretarias	10	6	7	2	*25*	*0,130*
SC	(0,400)	(0,240)	(0,280)	(0,080)		
Total	*61*	*45*	*62*	*25*	*193*	
Perfil medio	*(0,316)*	*(0,233)*	*(0,321)*	*(0,130)*		

ducción a las representaciones bidimensionales. Los datos tratan sobre una encuesta a 193 empleados de una empresa que tiene como objetivo conocer los hábitos de los fumadores de la empresa. Clasificamos a los empleados de la empresa de acuerdo con su nivel profesional (cinco grupos) y sus hábitos fumadores (cuatro grupos) (en la imagen 9.1 reproducimos la correspondiente tabla de contingencia). Dado que se trata de una tabla de 5 × 4, sus perfiles fila y sus perfiles columna se hallan exactamente en un espacio tridimensional.

Análisis de filas

Como hemos visto anteriormente, podemos ver esta tabla como un conjunto de filas o como un conjunto de columnas. Supongamos que el análisis de filas es más relevante, es decir, estamos interesados en representar, para cada grupo de empleados, los porcentajes de no fumadores, de fumadores ocasionales, etc. El espacio de perfiles constituye un símplex de cuatro puntos, es decir, un tetraedro de tres dimensiones que es el equivalente tridimensional al espacio triangular que vimos anteriormente (lo podemos visualizar utilizando las representaciones tridimensionales que describimos en el apéndice de cálculo, B). Para reducir la dimensionalidad de los perfiles, los podemos proyectar sobre el plano que mejor se ajuste (imagen 6.6). En el mapa de la imagen 9.2 también representamos gráficamente las proyecciones de los cuatro vértices que representan los hábitos fumadores de los empleados. Fijémonos en que, como es habitual, hemos situado el primer eje principal horizontalmente y el segundo eje principal verticalmente. Junto a los ejes indicamos las inercias principales (de 0,07476 y de 0,01002, respectivamente), así como los correspondientes porcentajes de inercia. Podemos sumar estos valores para conocer el porcentaje de inercia explicado por esta representación. Así vemos que la inercia explicada por el plano es de

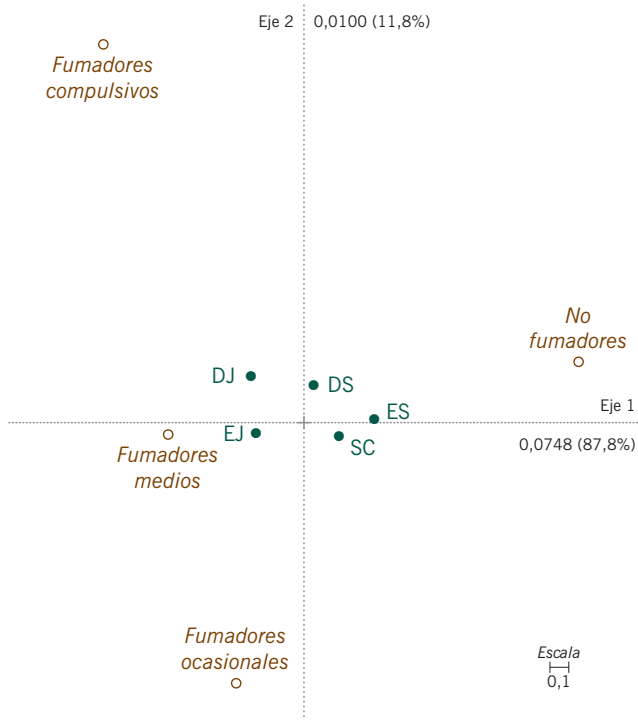


Imagen 9.2:
Mapa óptimo del AC bidimensional de los datos sobre los hábitos de los fumadores de la imagen 9.1, con las filas en coordenadas principales (proyecciones de los perfiles) y las columnas en coordenadas estándares (proyecciones de los vértices)

0,08478, lo que representa el 99,5% de la inercia total de 0,08519. Es decir, sacrificando una dimensión, hemos perdido sólo el 0,5% de la inercia de los perfiles. Está claro, pues, que los cinco perfiles fila se hallan muy cerca del plano representado, tan cerca, que cuando exploremos sus posiciones relativas, podremos ignorar las distancias de éstos al plano.

Interpretación de los perfiles fila y de los vértices columna

Si nos fijamos sólo en las posiciones de los perfiles, podemos comprobar que los grupos que se hallan más separados son, por un lado, los empleados jóvenes (EJ) y los directivos jóvenes (DJ) situados a la izquierda, y, por otro, los empleados séniors (ES) situados a la derecha; por tanto, las mayores diferencias en los hábitos de los fumadores se hallan entre estos dos extremos. Los directivos séniors (DS) se hallan entre los directivos jóvenes y los empleados séniors, mientras que las secretarias (SC) se hallan muy cerca de los empleados séniors. No obstante, para poder explicar las similitudes y las diferencias entre los grupos de empleados, es necesario que nos fijemos en las posiciones de los perfiles con relación a las de los vértices. Dado que las tres categorías de fumadores se hallan a la izquierda y la de no fumadores se halla a la derecha, la distinción entre derecha e izquierda es equivalente a la distinción entre fumadores y no fumadores. Los grupos EJ y DJ son diferentes del grupo de ES, ya que los primeros son relativamente fumadores,

mientras que el grupo ES es relativamente más «no fumador». El centro de este tipo de representaciones es siempre el perfil medio, de manera que podemos considerar las desviaciones de los grupos de empleados en distintas direcciones a partir del perfil medio, las mayores desviaciones se producen de izquierda a derecha.

Anidado de los ejes principales

La representación bidimensional también contiene la mejor representación unidimensional. Si proyectáramos todos los puntos de la imagen 9.2 verticalmente sobre el eje horizontal, esta representación unidimensional sería la misma que habríamos obtenido si, de entrada, hubiésemos estado interesados sólo en la mejor representación unidimensional. Decimos que los ejes están *anidados*. Es decir, la representación óptima de una determinada dimensionalidad contiene todas las representaciones óptimas de menor dimensionalidad. Fijémonos en que las proyecciones, sobre el primer eje, de los tres grupos de fumadores situados a la izquierda, quedan muy cerca entre sí y bastante separadas del punto situado a la derecha correspondiente a los no fumadores. Ésta es la característica más importante de los datos. Utilizando la terminología que vimos en el capítulo 7, podemos decir que la «escala de fumadores» que mejor diferencia los cinco grupos de empleados no es la que asigna intervalos iguales a las cuatro categorías de fumadores, sino la que sitúa a los tres grupos de fumadores muy cerca y muy separados del grupo de no fumadores. Efectivamente, existe una dicotomía entre fumadores y no fumadores.

Interpretación de la segunda dimensión

Continuando con la interpretación bidimensional, vemos que el segundo eje principal (eje vertical) separa los tres grupos de fumadores. Como indica el porcentaje de inercia explicada por el eje vertical, muy inferior, los perfiles no difieren tanto vertical como horizontalmente. No obstante, a pesar de que los EJ y los DJ tienen porcentajes similares de fumadores, como se puede ver por su posición similar en el eje horizontal, llegamos a la conclusión de que el perfil de los EJ tiene relativamente más fumadores ocasionales que fumadores compulsivos en comparación con el perfil de los DJ. Podemos verificar fácilmente estas conclusiones a partir de los datos originales de la imagen 9.1.

Verificación de la interpretación perfiles-vértices

Midiendo las distancias entre los perfiles y los vértices de la imagen 9.2, y comparando posteriormente dichas mediciones con los valores de los perfiles, podemos verificar la interpretación sobre las posiciones de los perfiles con relación a la de los vértices. Tenemos que hacer esta verificación vértice a vértice, por ejemplo, midiendo las cinco distancias de los grupos de empleados al vértice *fumadores ocasionales*. Como regla general, suponiendo que la representación sea de buena calidad, lo que es cierto en esta ocasión, cuanto más cerca se halle un perfil de un vértice, más se identifica este perfil con el grupo representado por el vértice. Así, por ejemplo, en el párrafo anterior dijimos que debido a que EJ se

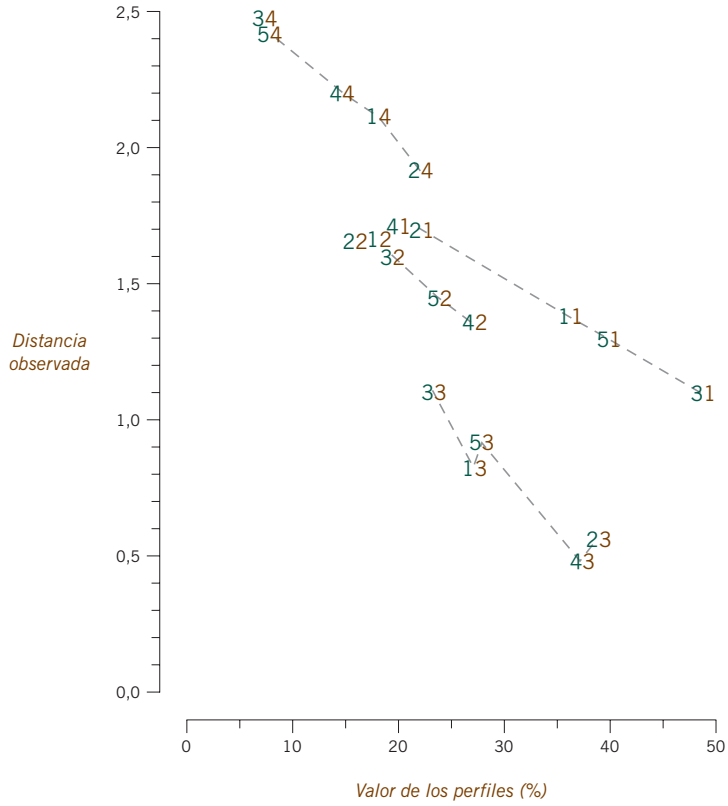


Imagen 9.3:
Distancias observadas de los perfiles a los vértices de la imagen 9.2, representadas con relación a los correspondientes valores de los perfiles fila de la imagen 9.1. Hemos etiquetado cada par fila-columna con sus números de categoría correspondiente; por ejemplo, el perfil fila 3 *(empleados séniors) y el vértice columna* 4 *(fumadores compulsivos) se denota como* 34*. Fijémonos en que, en cada vértice, salvo alguna excepción, a medida que aumentan los valores de los perfiles disminuyen las distancias*

halla más cerca del vértice *fumadores ocasionales* que DJ, EJ debe contener relativamente más fumadores ocasionales que DJ. Los datos muestran que 24/88, el 27%, de los individuos de EJ son fumadores ocasionales, mientras que sólo 3/18, el 17%, de los DJ lo son, lo que concuerda con nuestra interpretación. En la imagen 9.3 comparamos, de forma gráfica, las distancias de los perfiles a los vértices con los correspondientes valores de los elementos de los perfiles expresados como porcentaje. Utilizamos la abreviación 42, para indicar la distancia observada de EJ-a-*fumadores ocasionales* (fila 4, columna 2) y 22 para indicar la de DJ-a-*fumadores ocasionales* (fila 2, columna 2). Así vemos que los EJ quedan más cerca del vértice *fumadores ocasionales* que los DJ, para esta categoría el valor del elemento correspondiente del perfil es de 0,27 para los EJ y sólo de 0,17 para los DJ. En cada vértice, los elementos de los perfiles se relacionan de forma *monotónicamente inversa* con las distancias de los perfiles a los vértices. Gráficamente (imagen 9.3) ello significa que, en cada vértice, los cinco perfiles se disponen de forma descendiente, de izquierda a derecha. Por ejemplo, en el cuarto vértice *(fumadores habituales),* los perfiles con etiquetas 34, 54, 44, 14 y 24, se disponen en este tipo de secuencia descendente.

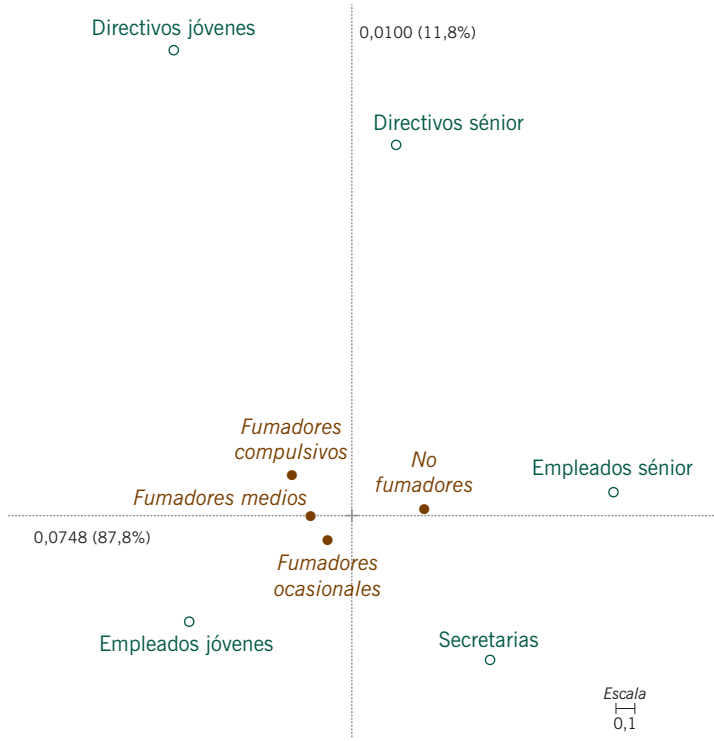


Imagen 9.4:
Mapa asimétrico del AC de los datos sobre los hábitos de los fumadores de la tabla 9.1, con las columnas en coordenadas principales y las filas en coordenadas estándares

Mapas asimétricos

Decimos que el mapa de la imagen 9.2 es un *mapa asimétrico*, o un mapa con *escalas asimétricas*, ya que es una representación conjunta de perfiles y vértices. En un mapa asimétrico, representamos las filas en coordenadas principales, y las columnas en coordenadas estándares o viceversa. Es decir, si estuviéramos más interesados en el análisis de las columnas que en el de filas, representaríamos las columnas en coordenadas principales, y las filas en coordenadas estándares. Lo que dijimos en el capítulo 8 sobre el factor de escala entre las filas y las columnas se cumple para todos los ejes principales. En consecuencia, la representación bidimensional de los perfiles columna sería una versión encogida de las posiciones de los vértices mostrados en el mapa de la imagen 9.2. Sin embargo los «factores de contracción» (es decir, las correlaciones canónicas, que son iguales a las raíces cuadradas de las inercias principales) de los dos ejes no son los mismos: $\sqrt{0,07476} = 0,273$ y $\sqrt{0,01002} = 0,1000$, respectivamente. Por tanto, el factor de contracción del primer eje es de 0,273 (una contracción de poco menos de cuatro veces), y el factor de contracción del segundo eje es de 0,1 (diez veces). Siguiendo el mismo razonamiento, para pasar de los perfiles fila de la imagen 9.2 a las posiciones de sus vértices, simplemente tenemos que expandirlos aproximadamente cuatro veces en el primer eje y diez veces en el segundo eje. Aparte de estos factores de escala, las posiciones relativas de los perfiles y los vértices son las mismas. En la ima-

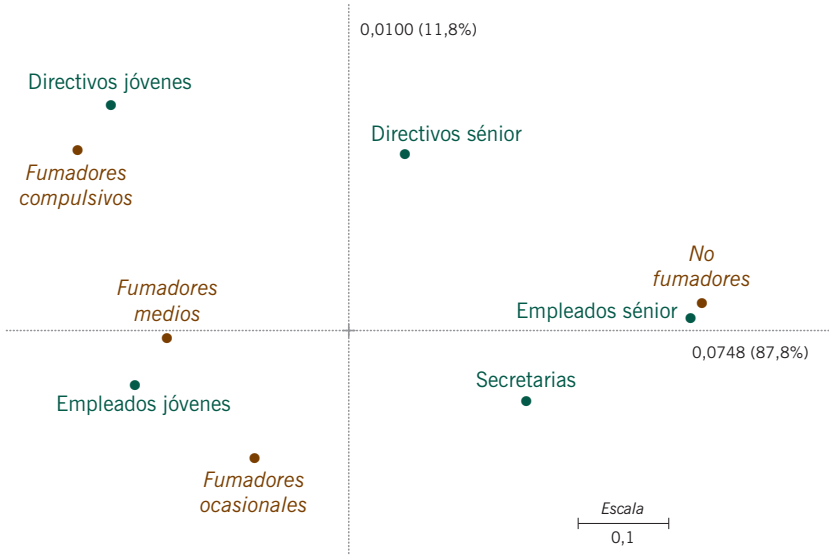


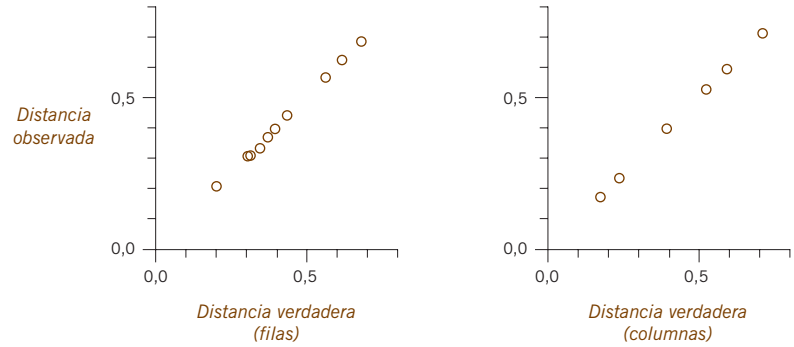
Imagen 9.5:
Mapa simétrico de los datos sobre los hábitos de los fumadores. Hemos representado tanto las filas como las columnas en coordenadas principales

gen 9.4 podemos ver otro posible mapa asimétrico, en el que hemos representado las columnas como perfiles en coordenadas principales, y las filas como vértices en coordenadas estándares. En este último mapa, las posiciones de los perfiles columna se hallan a medias ponderadas de los vértices de las filas, los pesos son los elementos de los perfiles de las columnas. Al mapa asimétrico de la imagen 9.2 lo llamamos mapa en *filas principales* (ya que expresamos las filas en coordenadas principales), en cambio al mapa de la imagen 9.4 lo denominamos mapa en *columnas principales*.

Mapa simétrico

Una vez examinada con bastante profundidad la explicación geométrica de las representaciones asimétricas, vamos ahora a introducir otra posibilidad de representación de los resultados, el *mapa simétrico*. Esta opción es, de lejos, la más popular en la literatura sobre el AC, especialmente entre los investigadores franceses. En los mapas simétricos solapamos en la misma representación, los perfiles fila y los perfiles columna, a pesar de que, en sentido estricto, las configuraciones de filas y columnas emanan de espacios distintos. Por tanto, en los mapas simétricos representamos tanto las filas como las columnas en coordenadas principales. Por ejemplo, el mapa de la imagen 9.5 es un mapa simétrico concerniente a los datos sobre los hábitos de los fumadores, en el que hemos solapado los dos conjuntos de puntos «interiores», que representamos mediante los círculos llenos en los mapas de las imágenes 9.2 y 9.4. La justificación de la representación conjunta de filas y de columnas hay que buscarla en la estrecha relación que existe entre el análisis de filas y el de columnas. Es decir, en la existencia de un solo factor de escala entre las filas y sus correspondientes vértices. La conveniencia de

Imagen 9.6:
Distancias observadas entre las filas y las columnas en la imagen 9.5, representadas con relación a las correspondientes verdaderas distancias χ^2 entre los perfiles fila y los perfiles columna de la tabla 9.1



este tipo de representación radica en que cualquiera que sea el nivel de asociación, en los mapas simétricos la dispersión de los perfiles fila y perfiles columna es similar. Por tanto, es más difícil que en la representación gráfica se produzca un solapamiento de etiquetas. En cambio, en los mapas asimétricos, los perfiles (que en general son los puntos de principal interés) se hallan apretados en el centro de la representación, lejos de los vértices exteriores, lo que hace que la visualización sea menos estética.

Verificación de la distancia ji-cuadrado entre los puntos en un mapa simétrico

En el mapa de la imagen 9.5, en la que hemos representado conjuntamente los perfiles de las filas y los de las columnas, las distancias entre las filas que aparecen en el mapa son aproximadamente distancias χ^2, de la misma manera que las distancias entre las columnas, son también aproximadamente distancias χ^2. Al ser la representación de las filas idéntica a la de la imagen 9.2, podemos aplicar la misma interpretación sobre las distancias entre las filas (fijémonos, sin embargo, en la diferencia de escala de estos dos mapas), o sea, lo mismo es aplicable a las columnas de la imagen 9.4. Esta similitud de distancias entre puntos se puede verificar representando gráficamente las distancias observadas *versus* las verdaderas (imagen 9.6). Vemos que existe una excelente concordancia, esperable en tanto que, en ambos casos, la calidad de la representación de los perfiles es del 99,5%.

El peligro de interpretar las distancias entre las filas y las columnas en un mapa simétrico

La comodidad de los mapas simétricos, como el de la imagen 9.5, tiene un precio que deriva del riesgo de querer interpretar, de forma directa, las distancias entre filas y columnas. En estos mapas, no hemos definido ni tenemos previsto definir este tipo de distancias. Se trata de una peculiaridad del AC que, a menudo, es mal comprendida y que, frecuentemente, causa confusión entre los usuarios de los mapas simétricos a los que les gustaría realizar grupos formados por filas y columnas (en este sentido, véase el epílogo de la obra). De forma rigurosa, no es posible deducir a partir de la proximidad entre un punto fila y un punto columna, que la fila y la columna correspondientes presenten una asociación elevada. Este tipo de interpretación es, hasta cierto punto, posible sólo en el caso de mapas asi-

métricos como el de la imagen 9.3. Una regla de oro para la interpretación de este tipo de mapas es que podemos interpretar la distancia entre puntos siempre que éstos se hallen en el mismo espacio, como es el caso de los perfiles fila y de los vértices columna en el espacio de perfiles fila. Cuando interpretemos mapas simétricos, tenemos que tener siempre bien presente que un mapa simétrico no es más que el «solapamiento de dos mapas distintos». En el capítulo 13, describiremos «el biplot»; un mapa que nos permite interpretar de forma más precisa la visualización conjunta de las filas y las columnas.

RESUMEN: Representaciones bidimensionales

1. Cuando en una representación gráfica aumenta la dimensionalidad de un subespacio, se incrementa la precisión de la representación de los perfiles. Sin embargo, al aumentar la dimensionalidad, la visualización de los puntos es más y más compleja. En general, preferimos las representaciones bidimensionales.
2. Los ejes principales están *anidados*; es decir, el eje principal de una representación unidimensional es idéntico al primer eje principal de una solución bidimensional, y así sucesivamente. Aumentar la dimensionalidad de una representación implica, simplemente, que añadimos nuevos ejes principales a los que ya hemos hallado.
3. Un *mapa asimétrico* es aquel en que representamos las filas y las columnas en escalas distintas, por ejemplo, las filas en coordenadas principales y las columnas en coordenadas estándares (son los vértices columna). Existen dos posibilidades, según sean de interés principal las filas o las columnas.
4. En un mapa asimétrico en el que, por ejemplo, representamos las filas en coordenadas principales (es decir, un análisis de filas), las distancias entre las filas son, aproximadamente, distancias χ^2. Las distancias entre las filas y un vértice columna son, en general, inversamente proporcionales a los valores de los elementos del perfil de esa columna.
5. Sin embargo, en los *mapas simétricos*, la representación gráfica más frecuente, expresamos tanto las filas como las columnas en coordenadas principales.
6. En los mapas simétricos, las distancias entre las filas y las distancias entre las columnas son aproximadamente distancias χ^2 de sus respectivos perfiles. No obstante, en un mapa simétrico no existe una interpretación específica para las distancias entre las filas y las columnas.

CAPÍTULO 10

Tres ejemplos más

Para terminar los primeros 10 capítulos de introducción al AC aplicado a tablas de contingencia de dos entradas, vamos a ver tres ejemplos más: a) una tabla que resume la clasificación de científicos de diez disciplinas de investigación en distintas categorías de financiación; b) una tabla de recuentos de 92 especies marinas en diferentes puntos de muestreo en el fondo marino; y c) un ejemplo lingüístico, en el que se ha llevado a cabo un recuento de las letras del alfabeto en muestras de textos en inglés de seis autores. Con el análisis de estos ejemplos, avanzaremos en la discusión sobre temas relacionados con las representaciones bidimensionales, como la interpretación de las dimensiones, la diferencia entre los mapas asimétricos y los mapas simétricos, y la importancia de la razón de escalas del mapa.

Contenido

Conjunto de datos 5: evaluación de investigadores científicos

Los datos proceden de una organización de investigación y desarrollo que clasificó a 796 investigadores científicos en cinco categorías de acuerdo con los recursos financieros de que disponían para su investigación (imagen 10.1). Hemos clasificado los investigadores según su disciplina científica (las 10 filas de la tabla) y según el tipo de financiación (las cinco columnas de la tabla). Asimismo, hemos

Imagen 10.1: *Frecuencias de las categorías de financiación de 796 investigadores que solicitaron fondos para la investigación: la categoría A corresponde a los que recibieron más recursos, la D a los que recibieron menos y la E a los que no recibieron*

	Categoría de financiación					
Disciplina científica	*A*	*B*	*C*	*D*	*E*	*Suma*
Geología	3	19	39	14	10	*85*
Bioquímica	1	2	13	1	12	*29*
Química	6	25	49	21	29	*130*
Zoología	3	15	41	35	26	*120*
Física	10	22	47	9	26	*114*
Ingeniería	3	11	25	15	34	*88*
Microbiología	1	6	14	5	11	*37*
Botánica	0	12	34	17	23	*86*
Estadística	2	5	11	4	7	*29*
Matemáticas	2	11	37	8	20	*78*
Suma	*31*	*128*	*310*	*129*	*198*	*796*
Perfil fila medio	*3,9%*	*16,1%*	*38,9%*	*16,2%*	*24,9%*	

etiquetado las categorías de financiación como *A*, *B*, *C*, *D* y *E*, de más a menos recursos financieros. En realidad, las categorías de la *A* a la *D* corresponden a investigadores que han disfrutado de recursos de investigación, de la *A* (los que recibieron más) hasta la *D* (los que recibieron menos), mientras que categoría *E* corresponde a los científicos que no consiguieron financiación (es decir, sus proyectos de investigación fueron rechazados).

Descomposición de la inercia

Esta tabla de 10 × 5 se halla exactamente en un espacio tetradimensional. La descomposición de la inercia en los dos primeros ejes principales es la siguiente:

Dimensión	*Inercia principal*	*Porcentaje de inercia*
1	0,03912	47,2%
2	0,03038	36,7%

Expresamos la inercia explicada por cada eje como porcentaje. Así, las dos primeras dimensiones explican casi el 84% de la inercia. La suma de las inercias principales es de 0,082879; por tanto, el estadístico χ^2 es igual a 0,082879 × 796 = 65,97. Si hiciéramos una prueba estadística, utilizando la distribución χ^2 con 9 × 4 = 36 grados de libertad, veríamos que se trata de un valor altamente significativo (p = 0,002).

Mapa asimétrico de perfiles fila

En la imagen 10.2 hemos representado el mapa asimétrico de los perfiles fila y los vértices columna. En esta representación gráfica vemos que el grado de la asociación entre las disciplinas científicas y las categorías de financiación es bastante baja; es decir, los perfiles no se alejan demasiado de la media (compárese con las figuras de la imagen 4.2). Esta situación es bastante típica de los datos derivados

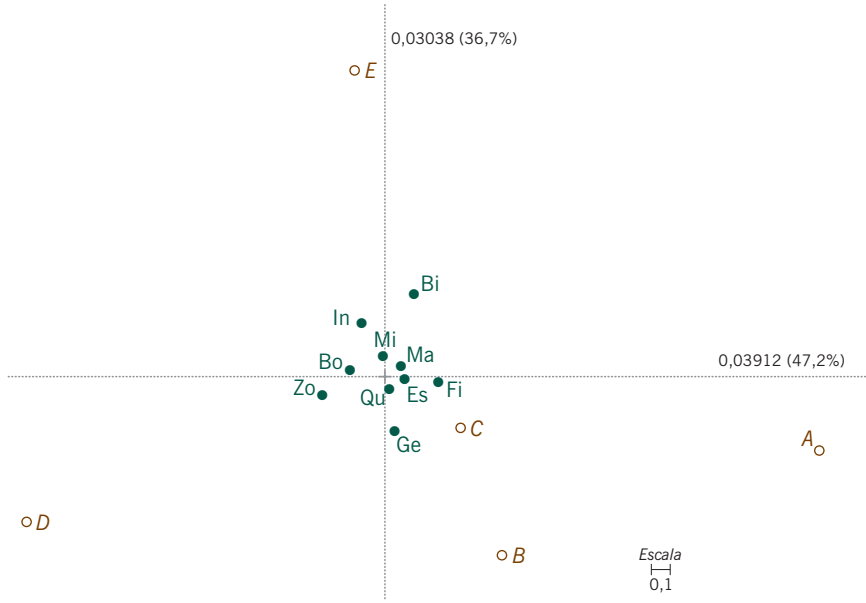
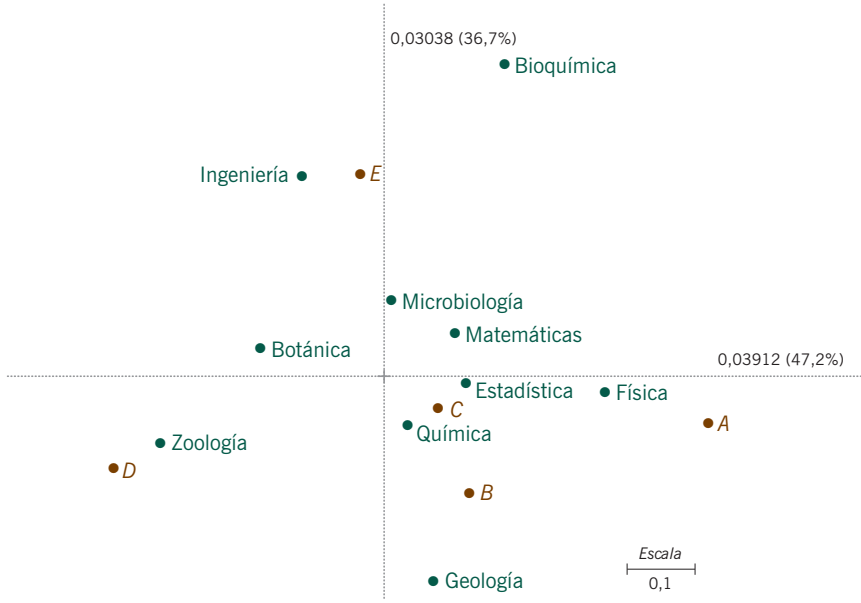


Imagen 10.2:
Mapa asimétrico de los perfiles fila de la tabla 10.1 (datos sobre la financiación de la investigación científica)

de las ciencias sociales. Por tanto, para esta tabla, el mapa asimétrico no es muy útil ya que todos los perfiles se hallan apelotonados en el centro de la figura —están tan cerca unos de otros que no podemos escribir las etiquetas completas, nos tenemos que conformar con las dos primeras letras de cada disciplina—. Sin embargo, si nos fijamos en las posiciones de los vértices, podemos interpretar fácilmente el espacio. La dimensión horizontal alinea en su orden natural las cuatro categorías de financiación, de la *D* (menor financiación) a la *A* (mayor financiación), las categorías *B* y *C* se hallan juntas, en medio. La dimensión vertical opone la categoría *E* (sin financiación) a las otras categorías, de manera que la interpretación es bastante directa. Cuanto más arriba se halle una disciplina, menos proyectos de los investigadores habrán conseguido financiación. Cuanto más a la derecha se halle una disciplina, más financiación habrán recibido los proyectos de los investigadores. Utilizando la terminología de la investigación de mercados, podríamos decir que «el punto ideal» se hallaría abajo a la derecha: más proyectos de investigación aceptados (abajo) y proyectos con buena financiación (a la derecha). Por tanto, si hiciéramos un estudio de tendencias en función del tiempo, las disciplinas tendrían que desplazarse hacia abajo a la derecha para mostrar una mejora en su estatus de financiación. De momento no existen disciplinas en esta zona, aunque la Física es la que se halla más a la derecha (mayor porcentaje [10 de 114, el 8%] de investigadores de la categoría *A*), pero verticalmente se halla en medio, ya que el porcentaje de investigadores que no han conseguido recursos se halla cerca de la media (26 de 114 no han conseguido financiación, el 22,8%, en comparación con la media global que es de 198 de 796, el 24,9%).

Imagen 10.3:
Mapa simétrico de la tabla de la imagen 10.1 (datos sobre la financiación de la investigación científica)



Mapa simétrico

En la imagen 10.3 mostramos el mapa simétrico de los mismos datos. La única diferencia entre este último mapa y el mapa de la imagen 10.2 es que, en vez de representar las columnas como vértices, las hemos representado como perfiles. De manera que se produce un cambio de escala que amplía la representación de los perfiles fila. Esta ampliación en la configuración de las disciplinas nos facilita la interpretación de sus posiciones relativas y nos deja más espacio para escribir las etiquetas completas. Ahora podemos ver más fácilmente las posiciones relativas de las disciplinas. Por ejemplo, Geología, Estadística, Matemáticas y Bioquímica se hallan todas en la misma posición, con relación al primer eje, lo que no ocurre con relación al segundo eje. Esto significa que los investigadores de estas disciplinas, cuyos proyectos de investigación han sido financiados, tienen posiciones similares con relación a las categorías de financiación de la *A* a la *D*. Sin embargo, Geología tiene muchos menos proyectos «no aceptados» (el 11,8% en la categoría *E*) que Bioquímica (41,4%). En esta representación simétrica no podemos valorar gráficamente el grado de asociación total (inercia) entre filas y columnas. Solamente podemos valorar la asociación a partir del valor de las inercias principales de cada eje, o a partir de sus raíces cuadradas, es decir, de las correlaciones canónicas en cada eje, concretamente $\sqrt{0{,}039117} = 0{,}198$ y $\sqrt{0{,}030381} = 0{,}174$, respectivamente. Sólo podemos evaluar gráficamente el grado de asociación entre filas y columnas en mapas asimétricos como el mapa de la imagen 10.2 (comparemos de nuevo los distintos grados de asociación que vimos en las figuras de la imagen 4.2).

	Estación (muestra)												
Especies	E4	E8	E9	E12	E13	E14	E15	E18	E19	E23	E24	R40	R42
Myri.ocul.	193	79	150	72	141	302	114	136	267	271	992	5	12
Chae.seto.	34	4	247	19	52	250	331	12	125	37	12	8	3
Amph.falc.	49	58	66	47	78	92	113	38	96	76	37	0	5
Myse.bide.	30	11	36	65	35	37	21	3	20	156	12	58	43
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Eucl.sp.	0	0	0	0	1	0	0	1	1	0	0	0	0
Scal.infl.	0	1	0	0	0	1	0	0	0	0	0	0	1
Eumi.ocke.	0	0	1	0	0	1	1	0	0	0	0	0	0
Modi.modi.	0	0	0	1	1	0	0	1	0	0	0	0	0

Imagen 10.4: *Frecuencias de 92 especies marinas en 13 muestras (las dos últimas son muestras de referencia); hemos ordenado las especies (filas) en orden descendiente de abundancia total; mostramos las cuatro especies más abundantes y las cuatro menos abundantes*

Interpretación de las dimensiones de los mapas

Tanto si recurrimos a mapas simétricos como asimétricos, la interpretación *dimensional* de los mapas siempre es la misma. Es decir, tenemos que interpretar los ejes uno a uno, como hicimos anteriormente. Así es, por ejemplo, lo habitual en el análisis factorial, cuando utilizamos las posiciones relativas de los puntos —«las variables» de la tabla— para asignar nombres descriptivos a los ejes. Es lo que hicimos anteriormente al describir en primer lugar los ejes a partir de las posiciones relativas de las categorías de financiación para, a continuación, interpretar las posiciones relativas de las disciplinas con relación a los ejes. Sin embargo, en este tipo de interpretación, todas las afirmaciones son relativas. Es decir, no podemos evaluar de forma absoluta las diferencias entre los perfiles de financiación de las distintas disciplinas a no ser que nos refiramos a los datos originales. Dicho de otra manera, con otros datos, aunque obtuviéramos mapas simétricos similares a los de la imagen 10.3, el grado de asociación entre los perfiles de financiación y las disciplinas podría ser mucho mayor (o menor).

Conjunto de datos 6: abundancia de especies en muestras del fondo marino

El AC se utiliza ampliamente para analizar datos en ecología. El segundo ejemplo que presentamos hace referencia a un conjunto de datos usual en biología marina. Los datos, que mostramos parcialmente en la tabla de la imagen 10.4, corresponden a los recuentos de 92 especies marinas identificadas en 13 muestras del fondo marino del mar del Norte. La mayor parte de las muestras se obtuvieron cerca de una plataforma petrolífera que producía una cierta contaminación del fondo marino. Existen dos muestras, utilizadas como referencia, supuestamente no contaminadas, que se obtuvieron lejos de la zona de influencia de la plataforma petrolífera. Estos datos, y en general este tipo de datos biológicos, se caracterizan por presentar una gran variabilidad, que podemos percibir fácilmente inspeccionando la pequeña parte de datos que aquí proporcionamos. La inercia total de esta tabla es de 0,7826, mucho mayor que la de los ejemplos anteriores. En consecuencia, cabe esperar que los perfiles mostrarán mucha más dispersión con relación a los vértices. Fijémonos que, en este ejemplo, no pode-

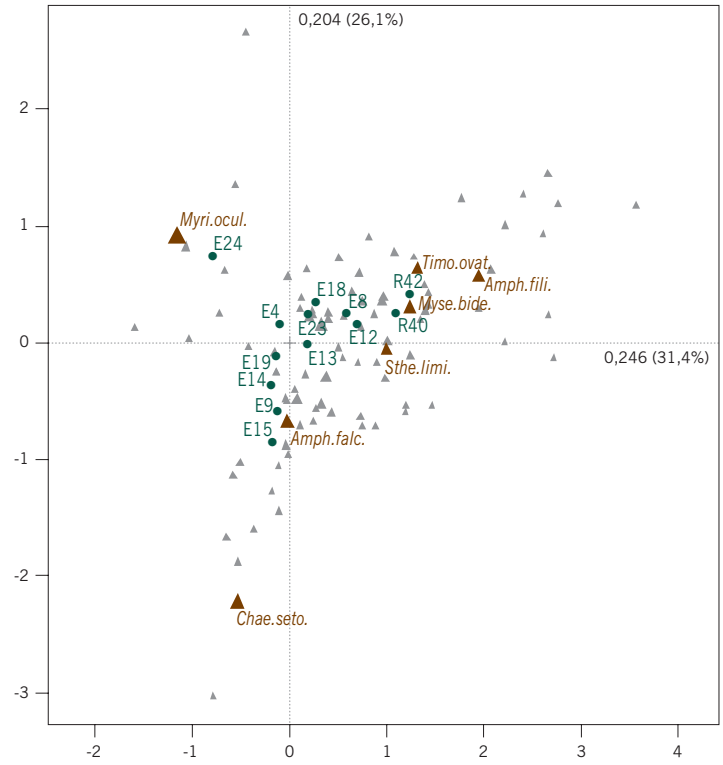


Imagen 10.5:
Mapa asimétrico del AC, con las estaciones de muestreo en coordenadas principales y las especies en coordenadas estándares. Los símbolos de las especies son proporcionales a la abundancia de las especies (masa); hemos etiquetado con las primeras letras de su nombre científico a algunas especies importantes para el análisis, situando la etiqueta al lado de su símbolo triangular. La inercia explicada por el mapa es del 57,5%

mos utilizar la prueba χ^2, ya que los datos no constituyen una verdadera tabla de contingencia (los recuentos no son independientes, ya que los organismos marinos a menudo se presentan agrupados en los puntos de muestreo).

Mapa asimétrico del AC de los datos sobre abundancia de especies

En el mapa de la imagen 10.5 podemos ver el mapa asimétrico de los perfiles de las muestras (columnas) y de los vértices de las especies (filas). Dado que tenemos 92 especies, no hemos podido etiquetar todos los puntos. Sólo hemos etiquetado las especies que contribuyen de forma más notable a la configuración del mapa, en general son las especies más abundantes. En el capítulo 11 veremos cómo podemos medir la contribución de cada punto; por el momento digamos simplemente que 10 de las 92 especies contribuyen en la construcción de este mapa en más de un 85% (podríamos eliminar las restantes 82 especies sin que el mapa cambiara demasiado). Podemos observar que las estaciones de muestreo describen una curva desde la parte baja a la izquierda (las estaciones de muestreo más contaminadas) hasta la parte superior derecha (las menos contaminadas). Las estaciones de referencia quedan arriba a la derecha. Una excepción es la estación de muestreo 24, que claramente se separa de las restantes, principalmente debido a su gran abundancia en *Myri.ocul.* (*Myriochele oculata*), lo podemos comprobar en la primera fila de la imagen 10.4.

Libro	Letras										
	a	*b*	*c*	*d*	*e*	⋯	*w*	*x*	*y*	*z*	*Suma*
TD-Buck	550	116	147	374	1.015	⋯	155	5	150	3	*7144*
EW-Buck	557	129	128	343	996	⋯	187	10	184	4	*7479*
Dr-Mich	515	109	172	311	827	⋯	156	14	137	5	*6669*
As-Mich	554	108	206	243	797	⋯	149	2	80	6	*6510*
LW-Clar	590	112	181	265	940	⋯	146	13	162	10	*7100*
PF-Clar	592	151	251	238	985	⋯	106	15	142	20	*7505*
FA-Hemi	589	72	129	339	866	⋯	225	1	155	2	*6877*
Is-Hemi	576	120	136	404	873	⋯	250	3	104	5	*6924*
SF7-Faul	541	109	136	228	763	⋯	160	11	280	1	*6885*
SF6-Faul	517	96	127	356	771	⋯	216	12	171	5	*6971*
Pe3-Holt	557	97	145	354	909	⋯	194	9	140	4	*6650*
Pe2-Holt	541	93	149	390	887	⋯	218	2	127	2	*6933*

Abreviaciones:
TD (Three Daughters), EW (East Wind) -Buck (Pearl S. Buck)
Dr (Drifters), As (Asia) -Mich (James Michener)
LW (Lost World), PF (Profiles of Future) -Clar (Arthur C. Clarke)
FA (Farewell to Arms), Is (Islands) -Hemi (Ernest Hemingway)
SF7 y SF6 (Sound and Fury, capítulos 7 y 6) -Faul (William Faulkner)
Pe3 y Pe2 (Bride of Pendorric, capítulos 3 y 2) -Holt (Victoria Holt)

Imagen 10.6:
Recuento de las letras en 12 muestras de textos de libros de seis autores distintos, que muestran datos de 9 de las 26 letras del alfabeto inglés

Como comentábamos anteriormente, hemos etiquetado las especies más abundantes, que son precisamente las que más determinan el mapa. Fijémonos en que los datos de este ejemplo se adaptan bien a un mapa asimétrico, sin duda debido a que existe una gran variabilidad entre las muestras, típico de los datos en ecología y, por tanto, la inercia es muy grande. ¡En el próximo ejemplo ocurre todo lo contrario!

Conjunto de datos 7: frecuencia de las letras en libros de seis autores

Hemos incluido este sorprendente ejemplo en el paquete **ca** del programa R (véase el apéndice de cálculo, B). Los datos forman una matriz de 12 × 26, las filas representan 12 textos que configuran seis pares, cada par contiene textos de un mismo autor (en la la imagen 10.6 mostramos parte de esta matriz). Las columnas son las 26 letras del alfabeto inglés, de la *a* a la *z*. Los datos son recuentos de letras en muestras constituidas por un texto de cada uno de los libros. Tenemos aproximadamente 6500-7500 recuentos de letras de cada libro o capítulo.

Una de las inercias más bajas, pero con una estructura significativa

Este conjunto de datos tiene una de las menores inercias totales que nunca he visto en mi experiencia con el AC. La inercia total es de 0,01873, lo que significa que los datos se hallan muy cerca de los valores esperados calculados a partir de las frecuencias marginales. Es decir, los perfiles son prácticamente idénticos. En la imagen 10.7, mostramos el mapa asimétrico de estos datos, con las letras en los vértices y los 12 textos formando una especie de pequeña mancha alrededor del origen, lo que indica

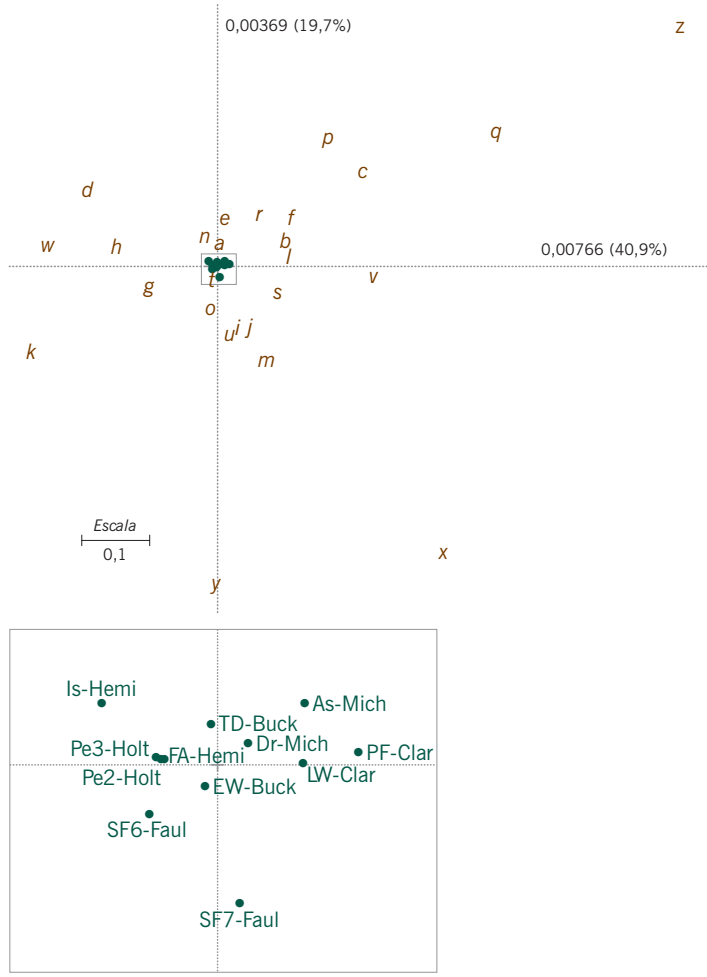


Imagen 10.7: *Mapa asimétrico de los datos sobre autores de la imagen 10.6, con las filas (textos) en coordenadas principales. La muy baja inercia de la tabla queda patente por la proximidad de los perfiles fila al centroide. Una «ampliación» del rectángulo situado en el centro del mapa muestra las posiciones relativas de los perfiles fila*

la poca variación que existe entre los textos en términos de distribución de las letras, por otra parte un resultado esperable. Es sorprendente ver cómo al ampliar esta pequeña mancha de puntos, de una pequeña variación, surge una estructura muy marcada. Efectivamente, los pares de textos del mismo autor aparecen juntos, resultado muy significativo desde un punto de vista estadístico (en el capítulo 25 veremos la prueba de permutación que nos permitirá contrastar esta afirmación).

La necesidad de mantener una razón de escalas de los mapas igual a 1

Vamos a realizar un importante comentario final sobre las representaciones gráficas de los mapas bidimensionales en el análisis de correspondencias. Dado que, en los mapas, las distancias son especialmente importantes, es evidente que una unidad de longitud en el eje horizontal debe ser igual a una unidad en el eje vertical. A pesar de que este requisito es obvio, en muchos programas informáticos y

hojas de cálculo se pasa por alto esta consideración y se dibujan diagramas de dispersión con los ejes en escalas distintas. Sabemos que, en general, los puntos presentan poca variación en el segundo eje (vertical); sin embargo, se suelen representar los mapas en rectángulos predefinidos que exageran este segundo eje. La *razón de escalas* del mapa, es decir el cociente entre una unidad de longitud horizontal y una unidad en el eje vertical debería ser igual a 1. Al final del apéndice de cálculo, B, presentamos algunas opciones para generar mapas de buena calidad.

RESUMEN:
Tres ejemplos más

1. Cuando sea posible, es útil contrastar —utilizando la prueba χ^2— si la asociación entre las filas y las columnas de una tabla de contingencia es significativa. Sin embargo, la significación estadística no es un requerimiento crucial para el análisis de mapas. Podemos ver el AC como una manera de expresar datos en forma gráfica para facilitar su interpretación; así tiene sentido representar cualquier tabla.
2. La interpretación *dimensional* de los mapas es siempre igual, tanto si recurrimos a mapas simétricos como a mapas asimétricos. Es decir, tenemos que interpretar los ejes uno a uno. Basamos la interpretación en asignar nombres descriptivos a los ejes principales a partir de las posiciones relativas de los puntos de uno de los dos conjuntos de coordenadas. A continuación, interpretamos las posiciones relativas del otro conjunto de coordenadas con relación a las dimensiones que previamente hemos asignado nombres descriptivos.
3. Los mapas asimétricos van bien cuando la inercia es alta, pero resultan problemáticos cuando la inercia total es pequeña. Ello es debido a que las coordenadas principales se hallan demasiado cerca del origen, lo que complica el etiquetado.
4. Es importante que las utilidades de representación gráfica mantengan la *razón de escalas* de los mapas. Una unidad en el eje horizontal debe aproximarse tanto como sea posible a una unidad en el eje vertical. Cuando las escalas son distintas, las distancias se distorsionan.

CAPÍTULO 11

Contribuciones a la inercia

La inercia total de una tabla de contingencia mide la variación existente en la tabla. Hasta ahora, hemos visto cómo descomponer la inercia en ejes principales, también hemos visto cómo descomponerla en filas o en columnas. Un paso más consiste en descomponer la inercia de filas o columnas en ejes principales. La investigación de estos componentes de la inercia (similar a un análisis de la varianza) desempeña un importante papel en la interpretación del AC. Además, proporciona herramientas de diagnóstico para identificar los puntos que más contribuyen a la definición de los ejes principales. Igualmente, nos permite valorar la calidad de la representación de los puntos.

Contenido

La inercia total mide la variación total de los perfiles

En el capítulo 4, la ecuación (4.7) nos mostró que, geométricamente, la inercia total es una media ponderada de las distancia χ^2 entre los perfiles y el perfil media. Con los perfiles columna obtenemos los mismos resultados. Si sólo existen pequeñas diferencias entre los perfiles y su media, la inercia toma un valor próxi-

Imagen 11.1: *Contribuciones de las filas y las columnas a la inercia, en valores absolutos que sumados dan la inercia total, y en valores relativos en tantos por mil (‰) que sumados dan 1000*

Filas	*Inercia*	*‰ inercia*	Columnas	*Inercia*	*‰ inercia*
Geología	0,01135	137	*A*	0,01551	187
Bioquímica	0,00990	119	*B*	0,00911	110
Química	0,00172	21	*C*	0,00778	94
Zoología	0,01909	230	*D*	0,02877	347
Física	0,01621	196	*E*	0,02171	262
Ingeniería	0,01256	152			
Microbiología	0,00083	10			
Botánica	0,00552	67			
Estadística	0,00102	12			
Matemáticas	0,00466	56			
Total	*0,08288*	*1000*	*Total*	*0,08288*	*1000*

mo a cero; es decir, existe poca variación (lo podemos ver en la imagen 4.2, en el diagrama triangular de arriba a la izquierda). El caso extremo ocurre cuando los perfiles se concentran en unas pocas categorías, y además en categorías distintas en los diferentes perfiles, en tal caso la inercia es grande (imagen 4.2, diagrama triangular de abajo a la derecha). La inercia es una medida de la dispersión de los perfiles en el espacio de perfiles.

Inercia de filas e inercia de columnas

La descomposición de la inercia en sumas de componentes positivos nos permite llevar a cabo un «análisis de inercia» útil para la interpretación de los resultados del AC. De acuerdo con la ecuación (4.7), la contribución de cada fila a la inercia es igual a su masa multiplicada por el cuadrado de su distancia al centroide de las filas, que llamaremos *inercia de filas*. Lo mismo se cumple para las columnas, y así obtendremos la *inercia de columnas*. Para facilitar la interpretación expresamos los componentes de la inercia con relación a la inercia total; se pueden expresar como porcentajes o, mejor, en *tantos por mil* (que indicaremos por ‰). En la tabla de la imagen 11.1 mostramos las inercias de las filas y de las columnas de los datos de la imagen 10.1 sobre la financiación de la investigación científica. Primero como «valores absolutos» y luego en forma relativa en tantos por mil. En nuestra implementación del AC en R utilizamos ampiamente los tantos por mil (lo podemos ver en el apéndice de cálculo, B), ya que nos permite incluir tres dígitos significativos sin utilizar decimales, lo que facilita la lectura de resultados.

Contribuciones grandes y contribuciones pequeñas a la inercia

Podemos ver fácilmente a partir de *«‰ de inercia»* de la imagen 11.1 que Zoología, Física, Ingeniería, Geología y Bioquímica son las filas, por este orden, que más contribuyen a la inercia; mientras que las categorías *D* y *E* son las columnas que más lo hacen. Como pauta general, para decidir qué contribuciones son grandes y cuáles son pequeñas, utilizaremos como valor umbral la media de las contribu-

Disciplina científica	Categoría de financiación					Suma
	A	*B*	*C*	*D*	*E*	
Geología	0	32	16	0	89	*137*
Bioquímica	0	23	4	44	48	*119*
Química	3	12	1	0	5	*21*
Zoología	9	14	11	189	8	*230*
Física	106	11	2	74	3	*196*
Ingeniería	1	11	38	1	102	*152*
Microbiología	2	0	0	3	5	*10*
Botánica	51	4	0	10	2	*67*
Estadística	10	0	0	2	0	*12*
Matemáticas	5	3	22	26	0	*56*
Suma	*187*	*110*	*94*	*347*	*262*	*1000*

Imagen 11.2: *Contribuciones de las celdas a la inercia, expresadas en tantos por mil. La suma de las filas y la de las columnas de esta tabla son idénticas a las inercias de las filas y columnas, expresadas en tantos por mil, de la imagen 11.1*

ciones a la inercia. Así, las contribuciones de las 10 filas suman 1000, su media será de 100, por tanto consideraremos contribuyentes principales las filas con contribuciones mayores del 100‰. Por otro lado, tenemos cinco columnas, lo que da una media de 200‰, por tanto, las columnas *D* y *E* son las que más contribuyen.

Contribuciones de las celdas a la inercia

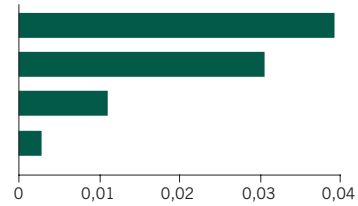
Podemos afinar más en el análisis de las contribuciones a la inercia examinando la contribución de cada celda. Como describimos en el capítulo 4, cada celda de la tabla contribuye con un valor positivo a la inercia total que, de nuevo, podemos expresar en tantos por mil (imagen 11.2). Vemos que las celdas [Zoología, *D*] y [Física, *A*] contribuyen mucho a la inercia; estas dos celdas juntas contribuyen casi al 30% de la inercia total de la tabla (189 + 106 = 295‰, es decir el 0,295 de la inercia total, es decir el 29,5%). Podemos denominar contribuciones ji-cuadrado a las contribuciones de las celdas, ya que sus valores son idénticos a las contribuciones relativas de cada celda al estadístico χ^2. Sumando las filas, o las columnas de la tabla de la imagen 11.2, llegamos a las mismas contribuciones, expresadas en tantos por mil, que vimos en la imagen 11.1.

Descomposición de la inercia en ejes principales

Otra descomposición importante de la inercia es con respecto a los ejes principales. En la página 106, dimos los valores de las inercias de los dos ejes principales de esta tabla de 10 × 5, que tiene cuatro dimensiones. En la imagen 11.3 mostramos los valores de todas las inercias principales, en porcentajes y en forma de diagrama de barras (lo llamaremos *diagrama de descomposición*). Hemos visto que las inercias principales se pueden interpretar por ellas mismas, por ejemplo, como el cuadrado de correlaciones canónicas (cap. 8, pág. 89). Sin embargo, en general, interpretaremos sus valores con relación a la inercia total, en general expresadas en porcentajes y no en tantos por mil.

Imagen 11.3:
Inercias principales de todas las dimensiones de los datos sobre la financiación científica expresadas en valores absolutos, en porcentajes y en porcentajes acumulados, y diagrama de descomposición

Dimensión	*Inercia principal*	*% inercia*	*% acumulado*
1	0,03912	47,2%	47,2%
2	0,03038	36,7%	83,9%
3	0,01087	13,1%	97,0%
4	0,00251	3,0%	100,0%



Componentes de cada inercia principal

Cada inercia principal es por ella misma una inercia, que hemos calculado a partir de las proyecciones de los perfiles fila (o de los perfiles columna) sobre los ejes principales. Por ejemplo, los 10 perfiles fila de los datos sobre la financiación de la investigación científica ocupan *un espacio completo* de dimensionalidad 4, uno menos que el número de columnas. La suma ponderada de los cuadrados de las distancias de los perfiles fila a su centroide es igual a la inercia total: su valor es de 0,08288. El primer eje principal es la recta mínimo-cuadrática más próxima a los perfiles. Este eje pasa por el centroide de las filas que se halla en el *origen*, o punto cero de la representación. Supongamos que proyectamos todos los perfiles fila sobre este eje. En tal caso, la primera inercia principal es la suma ponderada de los cuadrados de las distancias de estas proyecciones al centroide. Es decir, la primera inercia principal, igual a 0,03912, es la inercia de los puntos proyectados sobre el eje principal unidimensional. En la tabla de la imagen 11.4, mostramos las contribuciones de las filas y de las columnas a la primera inercia principal, calculadas a partir de las coordenadas principales de las filas y columnas. Así, vemos que la categoría *D* es la que más contribuye al primer eje, seguida de la *A*, mientras que las restantes categorías contribuyen muy poco. Con relación a las filas, vemos que Zoología (muy asociada con *D*) y Física (muy asociada con *A*) contribuyen en casi el 78% a la inercia del primer eje.

Imagen 11.4:
Contribución de las filas y de las columnas a la primera inercia principal; en valores absolutos, cuya suma es igual a la primera inercia principal, y expresadas de forma relativa en tantos por mil (‰)

Filas	*Inercia*	*‰ inercia*	Columnas	*Inercia*	*‰ inercia*
Geología	0,00062	16	*A*	0,00890	228
Bioquímica	0,00118	30	*B*	0,00260	67
Química	0,00023	6	*C*	0,00265	68
Zoología	0,01616	413	*D*	0,02471	632
Física	0,01426	365	*E*	0,00025	6
Ingeniería	0,00153	39			
Microbiología	0,00001	0			
Botánica	0,00345	88			
Estadística	0,00057	14			
Matemáticas	0,00112	29			
Total	*0,03912*	*1000*	*Total*	*0,03912*	*1000*

Descomposición completa de la inercia en los perfiles en los ejes principales

Podemos repetir lo anterior en todos los ejes principales. Así, en la imagen 11.5 mostramos los valores de la inercia de las filas descompuesta en los cuatro ejes principales (podemos construir una tabla similar para las columnas). Igual que en la imagen 11.4, en la que expresamos la descomposición de la inercia del primer eje principal en valores absolutos y en valores relativos, en tantos por mil, podríamos hacer lo mismo para todos los ejes principales. Así, veríamos que las filas que más contribuyen al eje 2 son Geología, Ingeniería y Bioquímica. El examen de las contribuciones de las filas (o de las columnas) a la inercia de los ejes principales nos proporciona un respaldo numérico a la interpretación de los mapas.

Componentes de la inercia de cada perfil

En la tabla de la imagen 11.5, los totales de las columnas nos dan las inercias principales de los ejes, mientras que los totales de las filas nos dan las inercias de los perfiles (por tanto, los valores de los totales de las filas deben ser iguales a los valores de la primera columna de la imagen 11.1). También podemos expresar las contribuciones a la inercia en términos relativos con relación a la inercia de las filas, como proporciones, en porcentajes o en tantos por mil. Estos resultados nos informarán sobre la inercia de las filas explicada por cada eje. Se trata de una miniversión de lo que hacíamos al determinar el porcentaje de inercia total que explicaba cada eje —aquí hacemos lo mismo pero fila a fila—. En la imagen 11.6 mostramos estos valores relativos en tantos por mil, es decir, el total de cada fila es 1000. Así, por ejemplo, vemos que el eje 2 es el que mejor explica la fila Geología, mientras que el eje 1 es el que mejor explica Física. Por otro lado, ni el eje 1 ni el 2 explican Matemáticas, puesto que su inercia se halla principalmente en la tercera dimensión.

Álgebra de la descomposición de la inercia

En la imagen 11.7 hemos representado gráficamente la descomposición de la inercia al mismo tiempo que introducimos un poco de notación. El punto $\mathbf{a}_i$, representa un perfil en un espacio multidimensional, por ejemplo el *i*-ésimo

	Eje principal				
Disciplina científica	*Eje 1*	*Eje 2*	*Eje 3*	*Eje 4*	*Total*
Geología	0,00062	0,00978	0,00082	0,00013	*0,01135*
Bioquímica	0,00118	0,00754	0,00084	0,00034	*0,00990*
Química	0,00023	0,00088	0,00029	0,00032	*0,00172*
Zoología	0,01616	0,00158	0,00063	0,00073	*0,01909*
Física	0,01426	0,00010	0,00169	0,00016	*0,01621*
Ingeniería	0,00153	0,00941	0,00127	0,00036	*0,01256*
Microbiología	0,00001	0,00056	0,00008	0,00019	*0,00083*
Botánica	0,00345	0,00016	0,00180	0,00011	*0,00552*
Estadística	0,00057	0,00001	0,00042	0,00003	*0,00102*
Matemáticas	0,00112	0,00037	0,00302	0,00015	*0,00466*
Total	*0,03912*	*0,03038*	*0,01087*	*0,00251*	*0,08288*

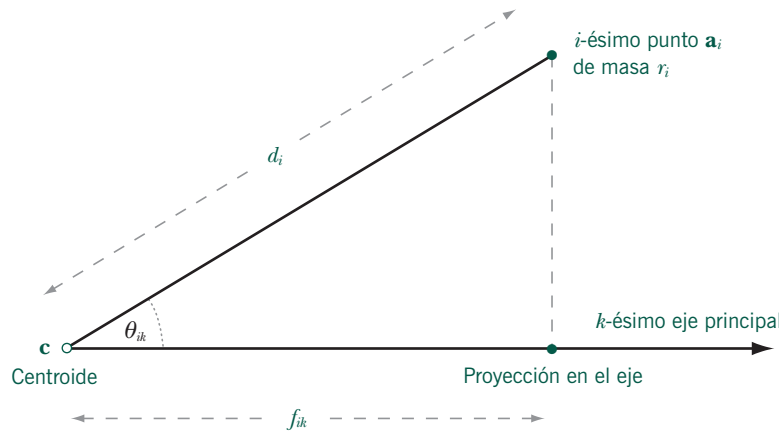
Imagen 11.5:
Descomposición en valores absolutos de la inercia de las filas (disciplinas científicas) en los cuatro ejes principales. La suma de las contribuciones de los ejes a las filas (totales de las filas) son las inercias de las filas de la imagen 11.1. Las sumas de las contribuciones de filas a los ejes (totales de las columnas) con las inercias principales de la imagen 11.3. La suma total de la tabla es la inercia total

Imagen 11.6: *Contribuciones relativas (en ‰) de los ejes principales a la inercia de las filas. En la última fila expresamos las inercias principales también en valores relativos que podemos interpretar como contribuciones relativas medias (comparar estos valores con los de la imagen 11.3)*

	Eje principal				
Disciplina científica	*Eje 1*	*Eje 2*	*Eje 3*	*Eje 4*	*Total*
Geología	55	861	72	11	*1000*
Bioquímica	119	762	85	35	*1000*
Química	134	510	170	186	*1000*
Zoología	846	83	33	38	*1000*
Física	880	6	104	10	*1000*
Ingeniería	121	749	101	28	*1000*
Microbiología	9	671	96	224	*1000*
Botánica	625	29	326	20	*1000*
Estadística	554	7	410	30	*1000*
Matemáticas	240	79	649	33	*1000*
Media	*472*	*367*	*131*	*30*	*1000*

perfil, de masa r_i, a una distancia d_i del perfil fila medio **c**. Por la ecuación (4.7), sabemos que la inercia total es igual a $\sum_i r_i d_i^2$. Hemos simbolizado la coordenada principal de $\mathbf{a}_i$ en el eje principal k por f_{ik}. Por tanto, la inercia en este eje (es decir la inercia principal k-ésima) será $\sum_i r_i f_{ik}^2$, en general simbolizada como λ_k. La contribución relativa de un punto i a la inercia principal del eje k sera $r_i f_{ik}^2$ dividido por λ_k (en la imagen 11.4 damos, en tantos por mil, las contribuciones relativas de los puntos al eje 1). En la imagen 11.5, mostramos los valores absolutos $r_i f_{ik}^2$ de las 10 filas y los 4 ejes principales de los datos sobre la financiación de la investigación científica. Los totales de las columnas de la imagen 11.5 son iguales a λ_k, mientras que los totales de las filas son la suma de las inercias de los ejes en esta fila $r_i d_i^2$. Gracias al teorema de Pitágoras, sabemos que $d_i^2 = \sum_k f_{ik}^2$, por tanto la contribución de los ejes a la inercia de las filas será:

Imagen 11.7: *Representación gráfica de un perfil $\mathbf{a}_i$ en un espacio multidimensional, a una distancia χ^2 d_i del centroide **c**, proyectado en la coordenada f_{ik} sobre el k-ésimo eje principal*



$$\sum_k r_i f_{ik}^2 = r_i d_i^2$$

Por tanto, la contribución relativa del eje k a la inercia del punto i es $r_i f_{ik}^2$ dividido por $r_i d_i^2$ (en la tabla de la imagen 11.6 damos estos valores relativos en tantos por mil).

Las contribuciones relativas como cuadrados de los cosenos de los ángulos

También podemos interpretar geométricamente las contribuciones relativas que hemos mostrado en la imagen 11.6. Dado que la proporción de inercia del punto i explicada por el eje k es $r_i f_{ik}^2 / r_i d_i^2 = (f_{ik}/d_i)^2$, viendo la imagen 11.7, queda claro que este valor es el cuadrado del coseno del ángulo formado por el punto i y el eje k. Supongamos que θ_{ik} sea dicho ángulo, entonces la contribución relativa del eje a la inercia del punto es $\cos^2(\theta_{ik})$. Por ejemplo, la contribución relativa del eje 1 a Física es de 0,880, por tanto, $\cos^2(\theta_{51}) = 0{,}880$, así pues, $\cos(\theta_{51}) = 0{,}938$ y, en consecuencia, el ángulo es $\theta_{51} = 20^\circ$. Este resultado muestra que Física, explicada principalmente por el eje 1, se halla cerca del eje 1, y forma un pequeño ángulo de 20º con dicho eje. En cambio, la contribución relativa del eje 1 a Geología es sólo de 0,055, lo que corresponde a un ángulo de $\theta_{11} = 76^\circ$ entre el eje 1 y Geología, luego esta disciplina científica no se halla cerca de este eje, más bien se encuentra en otras dimensiones del espacio (de hecho, se halla principalmente en el eje 2, lo que podemos deducir por la contribución relativa, de 0,861, de este eje a Geología).

Las contribuciones relativas como correlaciones al cuadrado

Existe todavía otra interpretación de las contribuciones relativas. Podemos interpretar los cosenos de los ángulos entre vectores como coeficientes de correlación. Por tanto, las contribuciones relativas son correlaciones al cuadrado. Así pues, podemos decir que la correlación de Física con el eje 1 es $\sqrt{0{,}880} = 0{,}938$, mientras que la correlación de Geología con dicho eje es de sólo $\sqrt{0{,}055} = 0{,}234$. Si la correlación es 1, el perfil se halla sobre el eje principal, mientras que si la correlación es 0, el perfil es perpendicular al eje principal (forma un ángulo de 90º).

Calidad de la representación en un subespacio

Gracias al teorema de Pitágoras, podemos sumar los cuadrados de los cosenos de los ángulos formados por un perfil y cada uno de los ejes, para obtener una suma de cosenos al cuadrado que relaciona el perfil con el subespacio definido por estos ejes. Así por ejemplo, podemos calcular el ángulo entre un perfil fila y el plano principal a partir de la suma de las contribuciones relativas de los dos ejes principales. Así, en la imagen 11.8 hemos sumado las dos primeras columnas de la tabla de la imagen 11.6. Interpretamos estas sumas como una medida de la *calidad* de la representación de los perfiles en los mapas bidimensionales que vimos en el capítulo 10, de la misma manera que la suma de los dos primeros porcentajes de inercia nos da una medida de la calidad global (o media) de la representación. Además, podemos ver qué perfiles están bien representados y cuáles no. Así, por ejemplo, en la última fila de la imagen 11.8 podemos ver que la calidad global del mapa bidimensional es del 83,9% y que, por tanto, no explicaría el 16,1%

Imagen 11.8:
Calidad de la representación (en tantos por mil) de los perfiles fila en dos dimensiones; solamente para Matemáticas *la inercia explicada es menor del 50%*

Disciplina científica	*Calidad*	Categoría de financiación	*Calidad*
Geología	916	*A*	587
Bioquímica	881	*B*	816
Química	644	*C*	465
Zoología	929	*D*	968
Física	886	*E*	990
Ingeniería	870		
Microbiología	680		
Botánica	654		
Estadística	561		
Matemáticas	319		
Global	*839*	*Global*	*839*

de la inercia de los perfiles. Algunos perfiles no estarán bien representados por hallarse más en el tercer y cuarto ejes que en los dos primeros. Tenemos, por ejemplo, que las Matemáticas están mal representadas, pues dos tercios de su inercia se encuentran fuera del plano. En las imágenes 10.2 y 10.3 el perfil de Matemáticas se parece al de Estadística, pero en este caso la proyección de su posición no es un reflejo preciso de su verdadera posición.

Analogías con el análisis factorial

Este apartado va dirigido, principalmente, a los lectores que conozcan el análisis factorial. Varios elementos del AC tienen elementos análogos a los del análisis factorial.

- El análogo al *coeficiente de carga del factor* es el coseno del ángulo formado por un perfil y un eje, es decir, la raíz cuadrada de la correlación al cuadrado con el signo de la coordenada del perfil. Por ejemplo, a partir de las imágenes 11.1 y 11.4, podemos calcular las correlaciones al cuadrado entre las categorías *A*, *B*, *C*, *D* y *E* con el primer eje principal.

$$A: \frac{0{,}00890}{0{,}01551} = 0{,}574 \qquad B: \frac{0{,}00260}{0{,}00911} = 0{,}286 \qquad C: \frac{0{,}00265}{0{,}00778} = 0{,}341$$

$$D: \frac{0{,}02471}{0{,}02877} = 0{,}859 \qquad E: \frac{0{,}00025}{0{,}02171} = 0{,}012$$

  Con los signos de las coordenadas de las columnas del mapa de la imagen 10.3, los «coeficientes de carga de los factores» serían las raíces cuadradas de los valores anteriores con sus correspondientes signos:

$$A: 0{,}758 \qquad B: 0{,}535 \qquad C: 0{,}584 \qquad D: -0{,}927 \qquad E: -0{,}108$$

- El análogo a la *comunalidad* es la calidad, expresada de 0 a 1. Por ejemplo, en la última columna de la imagen 11.8, mostramos las «comunalidades» de las

cinco categorías de las columnas de la solución bidimensional: 0,587; 0,816; 0,465; 0,968 y 0,990, para A, B, C, D y E, respectivamente.

- El análogo a la *unicidad* es 1 menos la calidad, expresada de 0 a 1. Por ejemplo, en la solución bidimensional, las «unicidades» de las cinco categorías de las columnas serán: 0,413; 0,184; 0,535; 0,032 y 0,010, para A, B, C, D y E, respectivamente.

RESUMEN: Contribuciones a la inercia

1. La inercia (total) de una tabla cuantifica la variación existente en los perfiles fila o en los perfiles columna.
2. Cada una de las filas y cada una de las columnas contribuye a la inercia total. Denominamos a estas contribuciones *inercias de las filas* e *inercias de las columnas*, respectivamente.
3. El AC se lleva a cabo con el objetivo de explicar la máxima inercia posible en el primer eje. El segundo eje explica el máximo de la inercia restante, y así sucesivamente. Por tanto, los ejes principales también descomponen la inercia total; a las inercias de los ejes principales las llamamos *inercias principales*.
4. A su vez, podemos descomponer las inercias principales con relación a las filas (o a las columnas). Tenemos dos posibilidades para expresar la *contribucion* del *k*-ésimo eje a la inercia de las filas (o de las columnas):
   a) con relación a la inercia principal del eje;
   b) con relación a la inercia de la fila o de la columna.
5. La posibilidad (a) nos permite diagnosticar qué filas (o columnas) han tenido un mayor papel en la determinación de la orientación de los ejes principales. Estas contribuciones nos facilitan la interpretación de los ejes principales.
6. La posibilidad (b) nos permite diagnosticar la posición de los perfiles con relación a los ejes y si éstos están bien representados en el mapa. Si están bien representados los podemos interpretar con seguridad, en cambio, si están pobremente representados, debemos interpretar sus posiciones con más cautela. Estos valores de inercia son los cuadrados de cosenos de los ángulos formados por los perfiles y los ejes principales, también los podemos interpretar como correlaciones al cuadrado.
7. La suma de los cuadrados de las correlaciones de un perfil con las dimensiones de un determinado subespacio nos proporciona una medida de la *calidad* de la representación del perfil en dicho subespacio.
8. Las correlaciones de los perfiles con los ejes, y las calidades de la representación del AC equivalen, respectivamente, a los coeficientes de carga de los factores y a las comunalidades del análisis factorial.

CAPÍTULO 12

Puntos adicionales

Con frecuencia ocurre que tenemos filas y/o columnas de datos que no hemos considerado inicialmente, pero que, sin embargo, nos pueden ser útiles para interpretar características que hayamos descubierto en los datos originales. Siempre que tenga sentido comparar los perfiles de estas nuevas filas (o columnas) con los de las filas (o columnas) de la matriz de datos originales que configuraron el mapa, tendremos la posibilidad de añadirlos en el mapa. Llamamos *puntos adicionales* o *suplementarios* a las filas o columnas que añadimos en un mapa preexistente.

Contenido

Puntos activos

Hasta ahora hemos utilizado todas las filas y columnas de una determinada tabla de datos para configurar los ejes principales y, en consecuencia, el mapa: son las filas y columnas *activas* del análisis. No todos los puntos activos tienen la misma fuerza de atracción sobre los ejes principales. Esta fuerza de atracción depende de la posición del punto y de su masa. Los perfiles alejados de la media «influyen» principalmente en la orientación del mapa, mientras que los perfiles con mayor masa tienen más «fuerza de atracción» sobre los ejes.

Imagen 12.1: *Frecuencias de las categorías de financiación de 796 investigadores (imagen 10.1), con una columna adicional Y, correspondiente a una nueva categoría de «nuevos investigadores prometedores», una fila adicional correspondiente a los investigadores que trabajan en museos, y una nueva fila que contiene la suma de las frecuencias de* Estadística *y* Matemáticas, *etiquetada como Ciencias matemáticas*

	Categoría de financiación					
Disciplina científica	*A*	*B*	*C*	*D*	*E*	*Y*
Geología	3	19	39	14	10	*0*
Bioquímica	1	2	13	1	12	*1*
Química	6	25	49	21	29	*0*
Zoología	3	15	41	35	26	*0*
Física	10	22	47	9	26	*1*
Ingeniería	3	11	25	15	34	*1*
Microbiología	1	6	14	5	11	*1*
Botánica	0	12	34	17	23	*1*
Estadística	2	5	11	4	7	*0*
Matemáticas	2	11	37	8	20	*1*
Museos	*4*	*12*	*11*	*19*	*7*	
Ciencias matemáticas	*4*	*16*	*48*	*12*	*27*	

Definición de puntos adicionales

Sin embargo, hay situaciones en las que deseamos visualizar las proyecciones de determinados puntos que no queremos que intervengan en el cálculo de la configuración del mapa. Queremos que la configuración del mapa se ajuste sólo a los puntos activos. Lo más simple es considerarlos como puntos que tienen una posición en el mapa, pero no tienen masa. Es decir, son puntos que no contribuyen a la inercia, y, por tanto, no influyen en la configuración de los ejes principales. Los llamamos *puntos adicionales*, *pasivos* o *suplementarios* y así los distinguimos de los puntos activos que sí tienen masa. Existen tres situaciones en las que las filas o las columnas adicionales pueden ser útiles. Vamos a ilustrar cada una de ellas en el contexto de los datos sobre la financiación de la investigación científica que vimos en capítulos precedentes. En la imagen 12.1 mostramos una versión ampliada de los mencionados datos. Hemos añadido:

1. Una fila adicional, etiquetada como *Museos*, que contiene las frecuencias de investigadores que trabajan en museos (a diferencia de los restantes que trabajan en universidades).
2. Una columna adicional, etiquetada como *Y*. Se trata de una categoría especial de financiación para investigadores jóvenes, una categoría que se acaba de introducir en el sistema de financiación.
3. Y otra fila, etiquetada como *Ciencias matemáticas,* que es la suma de Estadística y Matemáticas.

Primer caso: un punto intrínsicamente diferente a los restantes

El estudio del que derivan estos datos estaba, en principio, centrado en investigadores universitarios. Sin embargo, los investigadores de los museos tienen niveles similares y obtienen recursos de las mismas organizaciones financieras, por tanto, las frecuencias de 53 investigadores que trabajan en museos están clasificados en

Imagen 12.2:
Mapa simétrico de los datos de la imagen 12.1 (la podemos comparar con la imagen 10.2) que, además, muestra la posición de la columna adicional Y, y las filas adicionales Museos y Ciencias matemáticas

las mismas categorías de financiación. Aunque sea necesario considerar a los investigadores que trabajan en los museos de forma separada de los que trabajan en universidades, sigue teniendo interés visualizar los perfiles de los primeros en el «espacio» de los investigadores universitarios. Lo podemos hacer declarando la fila *Museos* como un punto adicional. De esta manera, su perfil no participará en la configuración de los ejes principales. De todas maneras, su perfil puede proyectarse sobre el mapa. En la imagen 12.2 mostramos un mapa simétrico, como el de la imagen 10.2, pero con el punto adicional *Museos,* abajo a la izquierda del mapa. Este punto no contribuye a la inercia principal, pero permite examinar las contribuciones relativas de los ejes a este punto (es decir, los cosenos o correlaciones al cuadrado). Así, podemos ver que este punto queda bastante bien representado en el mapa, ya que la contribución relativa de los ejes es superior al 50%. Su posición indica que relativamente pocos de los investigadores que trabajan en

museos han visto rechazados sus proyectos, pese a que el nivel de financiación de sus proyectos es bajo. A un conjunto de datos activo le podemos añadir distintos tipos de información adicional. Información que podemos obtener del mismo estudio, como es el caso de los *Museos* que acabamos de ver. O información que proceda de estudios similares, por ejemplo, con el objetivo de seguir la evolución, en función del tiempo, de las posiciones de las disciplinas científicas con relación a las categorías de financiación. Así, podríamos añadir en el mapa filas adicionales correspondientes a los datos de una tabla de frecuencias similar obtenida de un estudio anterior sobre la financiación de los investigadores científicos. Otro ejemplo sería añadir en el mapa perfiles objetivo para las distintas disciplinas científicas. De esta manera podríamos valorar lo lejos que quedan las posiciones actuales de las posiciones objetivo. Este concepto de «punto ideal» se utiliza frecuentemente en estudios sobre el posicionamiento de productos en investigaciones de mercados.

Segundo caso: una observación atípica de poca masa

Supongamos que se acaba de introducir una categoría de financiación adicional *Y,* de la que hasta el momento se han beneficiado muy pocos investigadores; de hecho, sólo seis investigadores de seis disciplinas distintas. Esto significa que el perfil de esta columna es muy poco común: seis valores del perfil son iguales a $\frac{1}{6}$ = 0,167, y los restantes valores son 0. Ningún otro perfil columna tiene el más mínimo parecido con éste, por tanto, hay que esperar que su posición en el espacio multidimensional sea inusual. Efectivamente, como se aprecia en la imagen 12.2, este punto es una *observación atípica*. Si lo hubiéramos incluido en el análisis como punto activo, hubiera contribuido mucho su configuración. No sería una situación deseable, ya que la columna *Y* está constituida sólo por seis individuos —por tanto, aparte de una razón sustantiva, existe una razón de tipo técnico que nos aconseja considerar este punto como punto adicional—. En este caso particular, si incluyéramos a *Y* como punto activo, y a pesar de que su masa representa menos del 1% de la masa de las columnas, la inercia total de la tabla pasaría de 0,0829 a 0,0920, un incremento del 11%. Además, como vemos en la imagen 12.3, cambiaría sustancialmente la configuración del mapa: se produciría una rotación de 30º en comparación con el resultado anterior. La inclusión de *Y* ha hecho girar los ejes. Debemos ponernos en guardia ante este tipo de observaciones atípicas de poca masa que contribuyen mucho a la inercia de la solución. En algunos casos extremos, las observaciones atípicas dominan tanto el mapa que los contrastes más interesantes entre categorías con mayores frecuencias quedan completamente enmascarados. Si declaramos las observaciones atípicas puntos adicionales, podemos seguir visualizando sus posiciones, sin que influyan en la configuración final del espacio. Otra posibilidad es combinar, siempre que tenga sentido, las filas (o las columnas de poca masa) con otras filas (o columnas). Así, si tuviéramos una disciplina adicional, como por ejemplo «Infor-

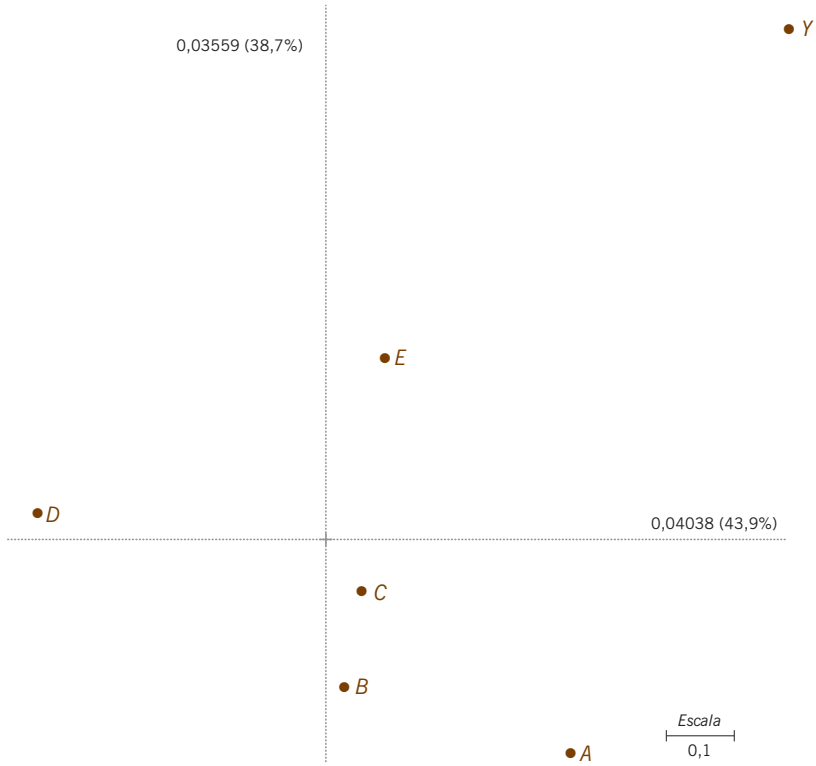


Imagen 12.3:
Mapa del AC de las columnas de la imagen 12.1 en el que hemos incluido Y como punto activo. Para facilitar las comparaciones, hemos expresado los ejes de los mapas de las imágenes 12.2 y 12.3 en la misma escala. Sin embargo, en la imagen 12.3 hemos rotado los ejes 30º respecto a los de la imagen 12.2

mática», con muy pocos investigadores y un posible perfil extraño, podríamos combinar esta disciplina con otra disciplina similar, como por ejemplo Matemáticas o Ingeniería. De todas formas, a pesar de lo que acabamos de comentar, es oportuno que señalemos que, en general, las observaciones atípicas de poca masa, no son un problema serio en el AC. Su influencia viene determinada por su masa multiplicada por una distancia al cuadrado, por tanto, la poca masa disminuye su influencia. El problema real es que las observaciones atípicas se hallen muy lejos de los restantes puntos (volveremos a este tema en el capítulo 13, cuando veamos distintas posibilidades de elección de las escalas de los mapas).

Tercer caso: representación de grupos o subdivisiones de puntos

También podemos utilizar los puntos adicionales para representar grupos de categorías o subdivisiones de una categoría. Por ejemplo, la fila adicional *Ciencias matemáticas* de la imagen 12.1, corresponde a la suma de las frecuencias de Matemáticas y Estadística, dos disciplinas que a menudo se agrupan. El perfil de esta nueva fila es el centroide de las dos filas que lo componen, ponderadas con sus respectivas masas. Dado que en Matemáticas y en Estadística hay 78 y 29 investigadores, respectivamente, el perfil de *Ciencias matemáticas* será:

$$\text{perfil de } \textit{Ciencias matemáticas} = \frac{78}{107} \times \text{perfil de Matemáticas} + \frac{29}{107} \times \text{perfil de Estadística}$$

de manera que el perfil de *Ciencias matemáticas* se parecerá más al perfil de Matemáticas que al de Estadística. Geométricamente, esto significa que el punto que representa el perfil de *Ciencias matemáticas* se halla en la línea que une Matemáticas con Estadística, pero más cerca de Matemáticas (compárese con la imagen 12.2). Vamos a representar *Ciencias matemáticas*, en la imagen 12.2, como un punto adicional. No lo podemos considerar un punto activo, ya que si así lo hiciéramos resultaría que en el análisis habríamos considerado dos veces los 107 investigadores de ambas disciplinas. En mapas de AC ya existentes representaremos de la misma manera las subdivisiones de categorías. Así, por ejemplo, supongamos que disponemos de datos que nos permiten subdividir Ingeniería en distintas ramas, como eléctrica, mecánica, civil, etc. Consideraremos estas nuevas filas como adicionales e investigaremos las posiciones de sus perfiles en el mapa. Igual que ocurría con *Ciencias matemáticas*, el punto activo Ingeniería es el centroide de las distintas ramas de la ingeniería que hemos considerado.

Cálculo de las posiciones de los puntos adicionales

Hasta ahora, hemos descrito los puntos adicionales como perfiles adicionales que proyectamos en un mapa calculado anteriormente. Una manera alternativa de obtener las posiciones de los puntos adicionales es situarlos con relación a los vértices de un mapa asimétrico. Así, en el capítulo 3 vimos que las posiciones de los perfiles fila resultaban del cálculo de la media ponderada de los vértices columna, ponderados con los elementos de los perfiles. Podemos situar los puntos adicionales exactamente de la misma forma. Una vez determinadas las posiciones de los ejes principales de los perfiles fila, conocemos, en cada eje principal, las posiciones de las coordenadas de los vértices que representan las columnas, es decir, las coordenadas estándares de las columnas. A partir de este momento, podemos situar los puntos adicionales calculando sus centroides con relación a los vértices ponderando con los elementos de sus perfiles. Por ejemplo, para calcular la posición del punto adicional *Museos*:

$$\text{posición de } \textit{Museos} = \frac{4}{53} \times \text{vértice } A + \frac{12}{53} \times \text{vértice } B + \ldots \text{ etc.}$$

es decir, calculamos la media ponderada de las coordenadas estándares de las columnas en cada uno de los ejes principales.

Contribuciones de los ejes a los puntos adicionales

Dado que los puntos adicionales tienen masa cero, su inercia es cero, por lo que no contribuyen a las inercias principales. Sin embargo, sigue siendo válida la interpretación de las contribuciones relativas de los ejes, en términos de los ángulos formados entre perfiles y ejes. Ello nos permite determinar si los puntos adicionales están bien representados. En el espacio bidimensional, las contribu-

ciones relativas de los ejes y las calidades de la representación de los tres puntos adicionales descritos anteriormente son las siguientes:

	Contribución relativa		*Calidad en dos dimensiones*
PUNTOS ADICIONALES	*Eje 1*	*Eje 2*	
Museos	225	331	556
Ciencias matemáticas	493	66	559
Y	4	587	641

Estos valores describen la bondad de la representación de los puntos adicionales. Por ejemplo, el coseno al cuadrado del ángulo del punto adicional *Y* con el primer eje es de 0,054, y con el segundo eje es de 0,587. Por tanto, la calidad de su representación en el plano es de 0,054 + 0,587 = 0,641, es decir, el 64,1% de su posición se halla en el mencionado plano, mientras que el 35,9% de su posición se halla en las restantes dimensiones. También podemos decir que la correlación de *Y* con el plano es $\sqrt{0,641} = 0,801$.

Los vértices son puntos adicionales

En realidad, ya nos encontramos anteriormente con puntos adicionales, ya que en el cálculo del mapa no tenemos en cuenta las posiciones de los vértices. Los vértices son puntos que proyectamos en los mapas con el objetivo de facilitar su interpretación; no intervienen en su configuración. Esta consideración nos sugiere una forma alternativa de cálculo de las posiciones de los vértices: en primer lugar, aumentaremos el número de filas en tantas filas como columnas tienen los datos; cada una de estas nuevas filas contendrá sólo un 1 y los restantes valores serán ceros. En cada una de las filas, el 1 se hallará en una columna distinta (imagen 12.4); en segundo lugar, declaramos las nuevas filas puntos adicionales. Las posiciones de estas filas adicionales son idénticas a las de los vértices de las columnas, es decir, sus coordenadas serán las coordenadas estándares de las columnas.

Variables categóricas adicionales y variables binarias

No debemos confundir el ejemplo de la columna adicional *Y* o el cálculo de las posiciones de vértices como filas adicionales de la imagen 12.4, con la codificación en «variables binarias», un tema que trataremos en detalle cuando, en los últimos capítulos, lleguemos al análisis de correspondencias múltiple. Supongamos, por

CATEGORÍA DE FINANCIACIÓN	*A*	*B*	*C*	*D*	*E*
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1

Imagen 12.4:
Agregación de filas adicionales a la tabla de la imagen 12.1: sus posiciones son idénticas a la de los vértices de las columnas

ejemplo, que clasificamos las disciplinas científicas en «Ciencias naturales» *(CN)* y en «Ciencias biológicas» *(CB)*, el último grupo incluiría Bioquímica, Zoología, Microbiología y Botánica, mientras que el primero contendría las restantes disciplinas científicas. En AC, una manera estándar de codificar estos datos sería con un par de variables binarias, *CN* y *CB*, es decir, como variables que toman los valores cero o uno. Así, para Geología (una ciencia natural), los valores de las variables binarias serían $CN = 1$ y $CB = 0$, mientras que para Bioquímica (una ciencia biológica), los valores serían $CN = 0$ y $CB = 1$, y así sucesivamente. Podríamos estar tentados en introducir estas variables binarias como columnas en la tabla y representarlas como puntos adicionales; sin embargo, esto no sería correcto. No estamos ante variables que expresen recuentos, a diferencia de la variable *Y*, que también tomaba los valores 0 y 1. En este caso, los valores de *Y* son verdaderos recuentos, que podían haber tomado otros valores enteros. La manera correcta de representar la información *CN*/*CB* sería mediante un par de filas, de manera similar a como representamos anteriormente *Ciencias matemáticas*. Sumaríamos las frecuencias de las filas *CN* y añadiríamos en la tabla una nueva fila que llamaríamos *CN*; haríamos lo mismo con las filas *CB*. De esta manera, los puntos *CN* y *CB* serían medias ponderadas de los puntos que representan los dos conjuntos de disciplinas científicas (en el capítulo 18 se retoma este tema).

Variables continuas adicionales

La información adicional en forma de variables continuas también requiere una especial consideración. Supongamos que tuviéramos información complementaria sobre las disciplinas científicas. Por ejemplo, el factor de impacto medio de los artículos publicados por investigadores de estas disciplinas en revistas internacionales. Podríamos situar esta información en una columna de datos y, dado que todos los valores son números positivos, podríamos estar tentados de representar el perfil de esta columna en forma de punto adicional. No obstante, debemos recordar que los perfiles columna representan números positivos que expresan proporciones de un total —no valores originales—, que además deben tener sentido en el contexto del estudio. ¿Qué haríamos si, por ejemplo, los datos expresaran cambios de la media del factor de impacto durante un determinado período, de manera que algunos valores fueran positivos y otros negativos? Es obvio que expresar estos valores con relación a su suma no tendría sentido alguno. En esta situación, podríamos representar esta variable continua de forma completamente distinta: por regresión. Veremos este tema con mucho más detalle en los capítulos 13 y 14, así como en el capítulo 24 cuando tratemos el análisis de correspondencias canónico, que consiste en una combinación del AC y de la regresión. Por el momento, únicamente pretendíamos alertar al lector sobre este problema.

RESUMEN: Puntos adicionales

1. Llamamos *puntos activos* a las filas y a las columnas de una tabla analizada por AC. Son los puntos que determinan la orientación de los ejes principales y, por tanto, contribuyen a la construcción de los mapas de baja dimensionalidad. Las filas y las columnas activas las proyectamos sobre el mapa.

2. Los puntos *adicionales* (o *pasivos)* son filas o columnas de la tabla que no han participado en la configuración del mapa, pero que tienen verdaderos perfiles. Son puntos que existen en los espacios completos de perfiles fila o de perfiles columna. Los podemos proyectar sobre un mapa de baja dimensionalidad con el objetivo interpretar sus posiciones con relación a los puntos activos.
3. Dado que los puntos adicionales tienen masa cero, el resultado de todos los cálculos en los que intervenga su masa será cero, como por ejemplo la inercia de los puntos o la contribución de los puntos a la inercia los ejes.
4. A pesar de que los puntos adicionales no contribuyen a la solución del AC, podemos calcular las contribuciones de los ejes principales (en términos de coseno o de correlación al cuadrado). Dichas contribuciones nos permiten valorar si los puntos adicionales se hallan bien representados en el mapa.
5. Hay que estar en guardia ante las observaciones atípicas de poca masa, cuya presencia en el análisis puede tener una gran influencia en la solución. Si es así, o bien los consideramos puntos adicionales o bien los combinamos —siempre que tenga sentido— con otras filas (u otras columnas).
6. Podemos crear una variable categórica adicional, por ejemplo una columna, para agrupar filas de acuerdo con las categorías de dicha variable. A continuación hallamos las frecuencias de dichas categorías y luego las añadimos como filas adicionales en la tabla.
7. Hay que ir con cuidado cuando añadamos una variable continua como punto adicional: sus valores no pueden ser negativos, además sus perfiles deben tener sentido en el contexto de los datos.

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:

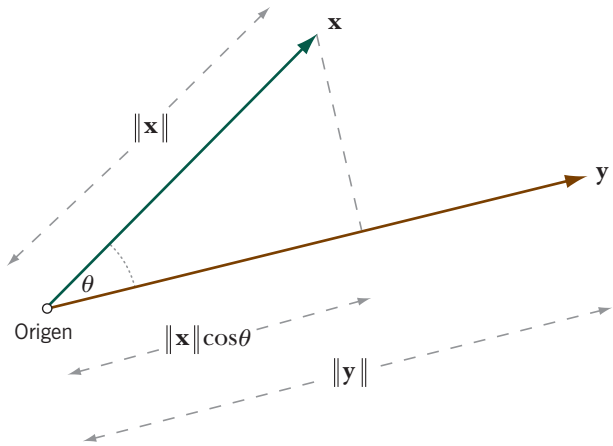
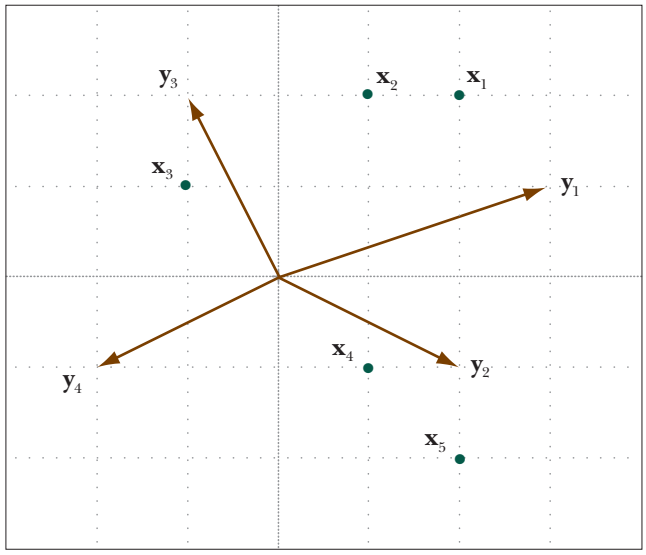


Imagen 13.1:
Ejemplo de dos puntos $\mathbf{x}$ *e* $\mathbf{y}$ *cuyos vectores forman un ángulo* θ *con relación al origen (en general, el centroide de un conjunto de puntos). El producto escalar entre los puntos es la longitud de la proyección de* $\mathbf{x}$ *sobre* $\mathbf{y}$*, multiplicada por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

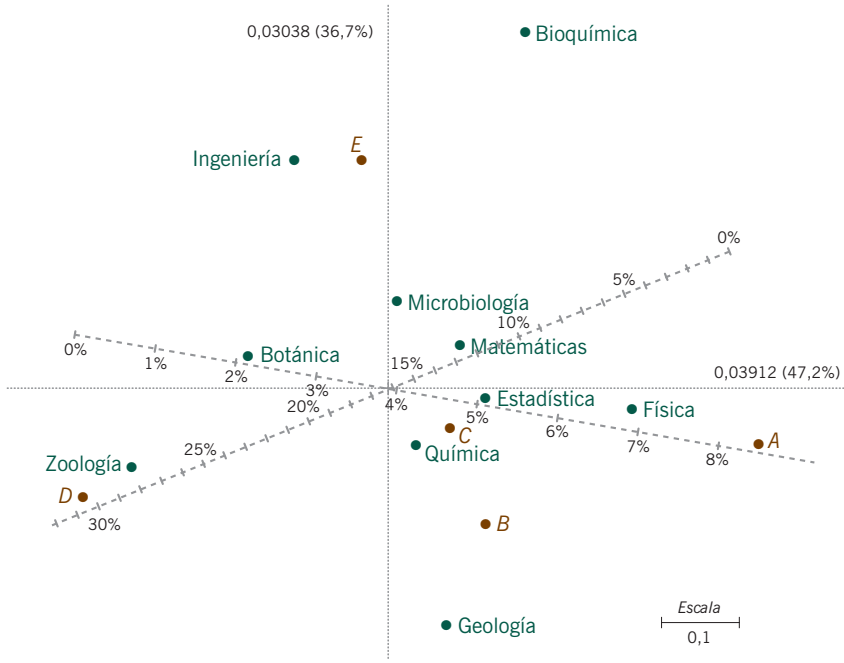


Imagen 13.4: *Mapa simétrico de la tabla 10.1 (datos sobre la financiación de la investigación científica) incluye los ejes de las categorías A y D calibrados. Fijémonos en que los ejes calibrados se hallan en la dirección de los vértices y en que no pasan exactamente por los puntos correspondientes 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.

CAPÍTULO 14

Relaciones de transición y regresión

En AC realizamos mapas en los que representamos conjuntamente filas y columnas como puntos cuya interpretación depende de las escalas escogidas para filas y columnas. Hemos visto cómo geométricamente las posiciones de las filas dependen de las posiciones de las columnas, y viceversa. En este capítulo nos centraremos en las relaciones matemáticas existentes entre las filas y las columnas; las ecuaciones de transición. Además, dado que el análisis de regresión es un método estadístico bien conocido, mostraremos cómo las coordenadas de las filas y las coordenadas de las columnas se pueden relacionar con los datos originales a través de modelos de regresión lineal. En realidad podríamos omitir este capítulo sin perder el hilo sobre la interpretación geométrica del AC.

Contenido

Las coordenadas en el primer eje del ejemplo sobre la financiación de la investigación científica

En este capítulo estamos interesados en las relaciones existentes entre las coordenadas principales y las coordenadas estándares de filas y de columnas, que emanan del AC, así como en su relación con los datos originales. Empecemos por fijarnos en las relaciones existentes en los ejes principales. En la tabla de la imagen 14.1 mostramos todos los resultados del primer eje principal del ejemplo

Imagen 14.1:
Coordenadas principales y coordenadas estándares de las disciplinas científicas y de las categorías de financiación en el primer eje principal del AC (datos originales en la imagen 10.1)

Disciplina científica	*Coordenada principal*	*Coordenada estándar*	Categoría de financiación	*Coordenada principal*	*Coordenada estándar*
Geología	0,076	0,386	*A*	0,478	2,417
Bioquímica	0,180	0,910	*B*	0,127	0,643
Química	0,038	0,190	*C*	0,083	0,417
Zoología	–0,327	–1,655	*D*	–0,390	–1,974
Física	0,316	1,595	*E*	–0,032	–0,161
Ingeniería	–0,117	–0,594			
Microbiología	0,013	0,065			
Botánica	–0,179	–0,904			
Estadística	0,125	0,630			
Matemáticas	0,107	0,540			

sobre la financiación de la investigación científica. Este eje tiene una inercia de $\lambda_1 = 0{,}03912$, siendo $\sqrt{\lambda_1} = 0{,}1978$. En el capítulo 8 vimos que este último valor es el factor de escala que relaciona coordenadas principales con coordenadas estándares. También vimos que lo podemos interpretar como un coeficiente de correlación entre las coordenadas de las filas y columnas en la primera dimensión. Dado que la correlación está relacionada con la regresión, en primer lugar veremos la regresión de las coordenadas de las filas sobre las de las columnas y viceversa.

Regresión entre coordenadas

En la imagen 8.5, utilizando los datos sobre la encuesta de la salud, mostramos el diagrama de dispersión de las coordenadas de las filas sobre las coordenadas de las columnas en el primer eje principal, de todos los individuos de la tabla de contingencia. En la imagen 14.2 mostramos el mismo tipo de representación gráfica para las coordenadas estándares de los datos sobre financiación de la investigación científica. En esta última imagen aparecen 50 cuadrados que corresponden a las 50 celdas de la tabla de contingencia de la imagen 10.1. En el diagrama, los cuadrados se sitúan en los correspondientes valores de filas y columnas de la tabla y tienen un área proporcional al número de individuos (científicos) de las celdas. Sabemos que la correlación, calculada para los 796 individuos representados por los 50 puntos de este diagrama, es de 0,1978. Ahora nos interesamos en la regresión de las disciplinas científicas sobre las categorías de financiación, y de las categorías de financiación sobre las disciplinas científicas. Para llevar a cabo este análisis de regresión haremos un listado en el que asignaremos, según los valores de la tabla de la imagen 14.1, a cada uno de los 796 científicos sus correspondientes pares de valores. Por ejemplo, las coordenadas estándares de un geólogo en la categoría *A*, son 0,386 (para la variable *y*) y 2,417 (variable *x*). Dado que sólo existen 50 pares distintos, una alternativa sería hacer un listado con únicamente los 50 pares en los que, junto con los valores de las coordenadas, aparezcan sus frecuencias y luego llevar a cabo una regresión ponderada tomando como pesos

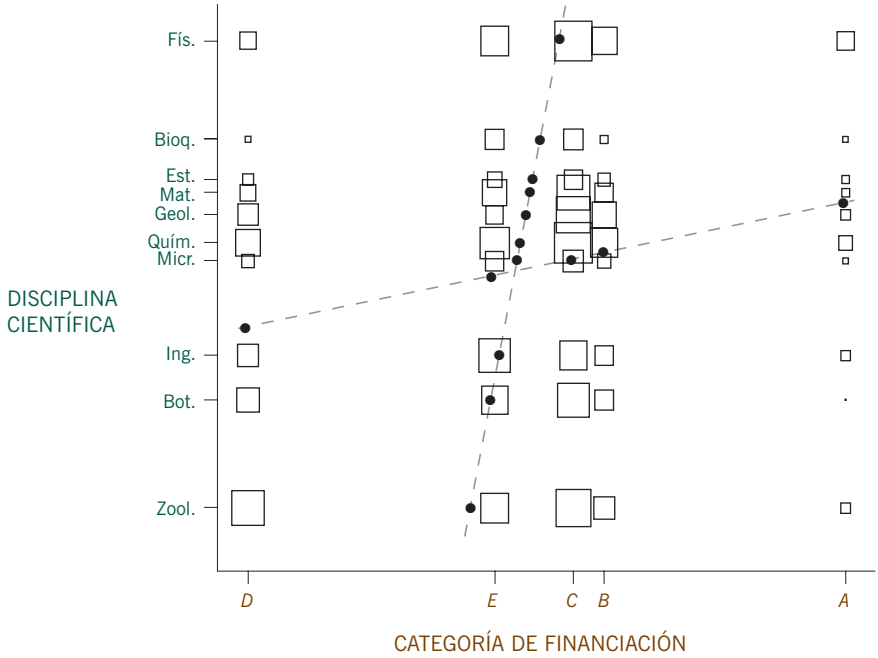


Imagen 14.2:
Diagrama de dispersión de las coordenadas estándares de las filas sobre las coordenadas estándares de las columnas en la primera dimensión del AC (imagen 14.1). Los cuadrados se sitúan en cada combinación de valores, con áreas proporcionales al número de individuos. Las dos rectas de regresión, de filas sobre columnas y de columnas sobre filas, tienen pendientes de 0,1978 y de 5,056, siendo cada una de ellas el valor inverso de la otra. Los puntos • indican medias condicionales (medias ponderadas), es decir, las coordenadas principales

las frecuencias (lo veremos en el apéndice de cálculo, B). Un resultado bien conocido de la regresión lineal simple es que la pendiente es igual a la correlación multiplicada por el cociente de la desviación típica de la variable *y* con la de la variable *x*. Las varianzas de las filas y de las columnas en coordenadas estándares son iguales a 1; por tanto, la pendiente de la recta de regresión de *y* sobre *x* será igual al coeficiente de correlación, concretamente igual a 0,1978 (imagen 14.2). De forma simétrica, la regresión de *x* sobre *y* tendrá también una pendiente de 0,1978, pero en este caso *x* estaría situada en el eje vertical y la *y* en el horizontal; sin embargo, dado que en el diagrama de la imagen 14.2 hemos situado *y* en el eje vertical, la pendiente entre *x* e *y* será 1/0,1978 = 5,056.

La relación entre perfiles y vértices

En el capítulo 3 vimos que la posición de los perfiles fila resulta de calcular las medias ponderadas de los vértices de las columnas, siendo los pesos los valores de los perfiles fila. Para los perfiles columna y los vértices fila se cumple la misma relación. Estas relaciones, basadas en el cálculo de medias ponderadas, también se cumplen para las proyecciones de filas y columnas sobre cualquier subespacio. En concreto, tal como vimos en el capítulo 8, se cumplen para las proyecciones de las coordenadas en los ejes principales. Es decir, en un eje principal *k*, las posiciones de las coordenadas principales de las filas son medias ponderadas de las coordenadas estándares de las columnas, y viceversa. En la imagen 14.2, ilustramos esta relación en el primer eje principal; hemos calculado las posiciones de

las filas —que mostramos como puntos, a partir de medias ponderadas de las coordenadas estándares de las columnas, y viceversa—. Podemos ver que las mencionadas coordenadas principales se sitúan en dos rectas.

En regresión, las coordenadas principales son medias condicionadas

Una regresión es un modelo de medias condicionales de la variable respuesta con relación a la variable explicativa. Los puntos negros de la imagen 14.2 no son más que medias condicionales de *y* sobre *x* (cinco medias en la recta de pendiente 0,1978) y de *x* sobre *y* (diez medias en la recta de pendiente 5,056). Estas medias condicionales son las coordenadas principales que, como podemos ver, en el diagrama definen dos funciones de regresión. Así, por ejemplo, la primera coordenada principal de Física, que mostramos como un punto negro en la parte superior del diagrama, es la media condicionada de las categorías de financiación expresadas en coordenadas estándares en el primer eje principal y ponderadas con las respectivas frecuencias de la matriz de datos de la imagen 10.1 (en este diagrama las hemos simbolizado por los cuadrados situados en la misma vertical que la de coordenada estándar de Física y, horizontalmente, según las coordenadas estándares de las columnas). De manera similar, hemos representado como un punto negro a la derecha la primera coordenada principal de *A*, que es la media condicionada de las categorías científicas expresadas en coordenadas estándares en el primer eje principal y ponderadas con las respectivas frecuencias de la matriz de datos de la imagen 10.1, que hemos simbolizado con los cuadrados situados en la misma horizontal que la coordenada estándar de *A* y, verticalmente, según las coordenadas estándares de las disciplinas científicas. Por tanto, en la imagen 14.2 mostramos simultáneamente coordenadas principales y coordenadas estándares. Las coordenadas principales de las filas son las coordenadas de los diez puntos situados sobre la recta de regresión en la escala de las coordenadas estándares de las columnas (escala horizontal) y viceversa.

Regresiones lineales simultáneas

El hecho de que en el AC las regresiones de *y* sobre *x* (filas sobre columnas) y de *x* sobre *y* (columnas sobre filas) sean rectas dio lugar a que inicialmente se presentara el AC como un sistema de *regresiones lineales simultáneas*. Si la correlación entre filas y columnas es alta, entonces las dos rectas de regresión serán muy parecidas y, en consecuencia, las coordenadas principales se hallarán más separadas; es decir, la inercia será mayor (recordemos que la inercia principal es igual al cuadrado de la correlación). Es decir, el AC se podría definir como un método que trata de buscar rectas de regresión simultáneas (como las mostradas en la imagen 14.2) que formen el menor ángulo posible entre ellas, lo que es equivalente a maximizar la correlación entre filas y columnas.

Ecuaciones de transición entre filas y columnas

Utilizando la notación que vimos anteriormente (págs. 51 y 139), y recordando que las coordenadas principales correspondan a perfiles y que las coordenadas estándares correspondan a vértices, podemos expresar la relación entre filas y columnas basada en el cálculo de medias ponderadas (o medias condicionales), de la manera siguiente:

perfil fila ← vértices de las columnas: $$f_{ik} = \sum_j \left(\frac{p_{ij}}{r_i} \right) \gamma_{jk} \qquad (14.1)$$

perfil columna ← vértices de las filas: $$g_{jk} = \sum_i \left(\frac{p_{ij}}{c_j} \right) \phi_{ik} \qquad (14.2)$$

(← simboliza «obtenido de», por ejemplo, «perfil fila ← vértices de las columnas» indica que hemos obtenido las coordenadas principales de una fila a partir las coordenadas estándares de todas las columnas utilizando la relación mostrada en la ecuación). Utilizamos la notación f y g para las coordenadas principales de las filas y de las columnas, respectivamente; y γ y ϕ para las coordenadas estándares de las filas y de las columnas, respectivamente. Para las filas utiizamos el subíndice i, para las columnas el subíndice j, y para dimensiones el subíndice k. Entre paréntesis mostramos los pesos, que son los perfiles de las filas en (14.1) y los de las columnas en (14.2). Llamamos *ecuaciones de transición* a las medias ponderadas de las expresiones (14.1) y (14.2). Recordemos que las relaciones existentes entre las coordenadas principales y las coordenadas estándares son:

perfil fila ← vértice fila: $$f_{ik} = \sqrt{\lambda_k}\, \phi_{ik} \qquad (14.3)$$

perfil columna ← vértice columna: $$g_{jk} = \sqrt{\lambda_k}\, \gamma_{jk} \qquad (14.4)$$

donde λ_k es la inercia principal (valor propio) del k-ésimo eje. Luego, las ecuaciones de transición entre las coordenadas principales de las filas y las coordenaas principales de las columnas, las podemos expresar de la siguiente manera:

perfil fila ← perfiles columna: $$f_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_j \left(\frac{p_{ij}}{r_i} \right) g_{jk} \qquad (14.5)$$

perfil columna ← perfiles fila: $$g_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_i \left(\frac{p_{ij}}{c_j} \right) f_{ik} \qquad (14.6)$$

y, de manera similar, podemos expresar las ecuaciones de transición entre las coordenadas estándares de filas y las coordenadas estándares de columnas como:

vértice fila ← vértices columna: $$\phi_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_j \left(\frac{p_{ij}}{r_i} \right) \gamma_{jk} \qquad (14.7)$$

vértice columna ← vértices fila: $$\gamma_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_i \left(\frac{p_{ij}}{c_j} \right) \phi_{ik} \qquad (14.8)$$

Regresión entre coordenadas usando ecuaciones de transición

Podemos utilizar cualquiera de las ecuaciones de transición anteriores, para «estimar» por regresión las coordenadas a partir de los perfiles (variables explicativas). Por ejemplo, supongamos que queremos utilizar la ecuación (14.1) para obtener las coordenadas estándares de las columnas. Las variables respuesta serán las primeras coordenadas principales de las filas (primera columna de la imagen 14.1), y las variables explicativas los diez perfiles de las filas de la matriz 10 × 5 de la imagen 10.1. El análisis de la regresión da los siguientes coeficientes de regresión:

Fuente de variación	*Coeficiente*
Ordenada en el origen	0,000
A	2,417
B	0,643
C	0,417
D	–1,974
E	–0,161

$R^2 = 1{,}000$

La varianza explicada es del 100% y los coeficientes de regresión son las coordenadas estándares de las columnas en el primer eje (última columna de la imagen 14.1).

Recordatorio del modelo bilineal del AC

Realizamos un análisis de regresión más interesante y más relevante, cuando predecimos los datos a partir de las coordenadas, como vimos de forma resumida en el capítulo 13 al tratar sobre el modelo del AC. Vamos a repetir aquí tres versiones del mencionado modelo; la «versión simétrica», utilizando sólo coordenadas estándares [véase (13.4)], y las dos versiones asimétricas con filas o columnas respectivamente en coordenadas principales:

$$\frac{p_{ij}}{r_i c_j} = 1 + \sum_{k=1}^{K^*} \sqrt{\lambda_k}\,\phi_{ik}\gamma_{jk} + e_{ij} \tag{14.9}$$

$$\left(\frac{p_{ij}}{r_i}\right) / c_j = 1 + \sum_{k=1}^{K^*} f_{ik}\gamma_{jk} + e_{ij} \tag{14.10}$$

$$\left(\frac{p_{ij}}{c_j}\right) / r_i = 1 + \sum_{k=1}^{K^*} \phi_{ik} g_{jk} + e_{ij} \tag{14.11}$$

Este modelo se llama *bilineal* porque es lineal con relación a los productos de dos parámetros. Tendremos que fijar los valores de las coordenadas estándares de las filas o de las columnas, para obtener las coordenadas principales mediante análisis de regresión múltiple.

Regresión ponderada

En los lados izquierdos de las ecuaciones (14.9), (14.10) y (14.11) aparecen los cocientes de contingencia que definimos en el capítulo 13, escritos de tres maneras equivalentes. Tomando como ejemplo (14.10), y suponiendo que conocemos

las coordenadas estándares γ_{jk} de las columnas, tenemos a la derecha un modelo de regresión que predice los valores de las filas de la izquierda. Supongamos que estamos interesados en la primera fila (Geología) y que queremos llevar a cabo una regresión para $K^* = 2$. Para ajustar el modelo de AC, tenemos que minimizar una suma ponderada de residuos en la que ponderamos las categorías (columnas) con sus masas. Otra forma de verlo es decir que en (14.10) normalizamos las «variables explicativas» γ_{jk} con las masas de las columnas de la siguiente manera: $\sum_j c_j \gamma_{jk}^2 = 1$. Además, como las variables explicativas son ortogonales, cuando ponderamos con las masas de las columnas: $\sum_j c_j \gamma_{jk} \gamma_{j'k} = 0$ si $j \neq j'$. Para llevar a cabo la regresión, acomodamos el vector respuesta como un vector de 5×1 con los cocientes de contingencia de Geología, y las variables explicativas la disponemos como una matriz de 5×2 con las coordenadas estándares de las columnas en los dos primeros ejes principales. Llevamos a cabo el análisis de la regresión ponderada, ponderando con las masas de las columnas c_j. Los datos (los cocientes de contingencia de Geología (fila 1) indicados como $p_{1j} / (r_1 c_j)$, las coordenadas estándares de las columnas en las dimensiones 1 y 2, indicadas como γ_1 y γ_2, y los pesos c_j) son los siguientes:

Categoría	Geología	γ_1	γ_2	*Masa*
A	0,9063	2,4175	–0,4147	0,0389
B	1,3901	0,6434	–0,9948	0,1608
C	1,1781	0,4171	–0,2858	0,3894
D	1,0163	–1,9741	–0,7991	0,1621
E	0,4730	–0,1613	1,6762	0,2487

Los resultados de la regresión son:

Fuente	*Coeficiente*	*Coeficiente estandarizado*
Ordenada en el origen	1,000	—
f_{11}	0,076	0,234
f_{12}	–0,303	–0,928

$R^2 = 0{,}916$

Los coeficientes son las coordenadas principales f_{11} y f_{12} de Geología (el primer valor lo encontramos en la imagen 14.1). La varianza explicada (R^2) es la calidad de la representación de Geología en el mapa bidimensional (imagen 11.8).

En la regresión ponderada, las correlaciones recuperan las contribuciones relativas

Dado que en la regresión ponderada las variables explicativas están estandarizadas y son ortogonales, los coeficientes de regresión estandarizados serán también las correlaciones parciales entre la variable respuesta y las variables explicativas. La matriz de correlaciones de las tres variables es la siguiente (recordemos que en los cálculos hemos tenido en cuenta los pesos):

Variables	Geología	γ_1	γ_2
Geología	1,000	0,234	–0,928
γ_1	0,234	1,000	0,000
γ_2	–0,928	0,000	1,000

Como esperábamos, las dos variables explicativas no están correlacionadas. Las correlaciones entre Geología y las dos variables explicativas son exactamente los coeficientes de regresión estandarizados. Los cuadrados de estas correlaciones, $0{,}234^2 = 0{,}055$ y $(-0{,}928)^2 = 0{,}861$, son los cosenos al cuadrado (contribuciones relativas) que vimos en la imagen 11.6. Los resultados que acabamos de ver, ilustran la propiedad de la regresión que establece que si las variables explicativas no están correlacionadas, la varianza explicada R^2 es igual a la suma de los cuadrados de las correlaciones parciales.

Cálculo recíproco de medias y mínimos cuadrados alternados

Las ecuaciones de transición (14.1) y (14.2) son la base de un conocido algoritmo para hallar la solución del AC, llamado *cálculo recíproco de medias*. Empezamos el algoritmo con unas coordenadas estándares de las columnas —que hemos centrado y normalizado con medias y sumas de cuadrados ponderadas—. Aplicando la fórmula (14.1) de cálculo de medias ponderadas, calculamos valores para las coordenadas de las filas, a continuación aplicando la fórmula (14.2) a los valores anteriores de las filas, calculamos nuevos valores para las coordenadas de las columnas. Seguidamente estandarizamos estos valores y repetimos el proceso desde el inicio hasta la convergencia de los resultados, es decir, hasta obtener las coordenadas principales en el primer eje principal (es necesario estandarizar los valores de las coordenadas de las columnas que obtenemos en cada iteración, en caso contrario en los sucesivos cálculos de medias llegaríamos al valor cero). Hallar el segundo conjunto de coordenadas es más complicado, ya que tenemos que asegurar la ortogonalidad con las primeras coordenadas, no obstante la idea es la misma. Anteriormente, hemos visto que el paso de coordenadas columna a coordenadas fila, y de coordenadas fila a coordenadas columna lo podemos hacer mediante regresión. Por este motivo, este algoritmo se conoce también como *mínimos cuadrados alternados*, o regresiones alternadas. De todas formas, numéricamente es mejor llevar a cabo los cálculos utilizando la DVS (veánse los apéndices A y B), pero conocer estos algoritmos alternativos nos ayuda a profundizar en la comprensión del AC.

RESUMEN: Relaciones de transición y regresión

1. Cualesquiera que sean los valores asignados a las categorías de filas y columnas, podemos calcular las medias condicionales (es decir, las regresiones) de las filas con relación a las columnas o de las columnas con relación a las filas.
2. Realizado el AC, las coordenadas estándares de filas y de columnas cumplen las siguientes propiedades:

— la regresión de filas sobre columnas, y viceversa, son lineales (de aquí el nombre de *regresiones lineales simultáneas):*
— se minimiza el ángulo entre las dos regresiones;
— las medias condicionales que se hallan en las dos rectas de regresión son las coordenadas principales.

3. Llamamos *ecuaciones de transición* a las medias ponderadas entre coordenadas de filas y columnas, ponderadas con los elementos de los perfiles (de filas o columnas según el caso). Llamamos *cálculo recíproco* de medias a un algoritmo que permite hallar la solución del AC mediante la aplicación sucesiva de un par de ecuaciones de transición.

4. Podemos definir el AC como un *modelo de regresión bilineal,* ya que podemos recuperar los datos originales a partir de un modelo lineal de productos de coordenadas de filas y columnas. Este modelo se convierte en lineal si contemplamos como fijos los valores de las coordenadas de filas o columnas, lo que conduce a un algoritmo para hallar la solución del AC llamado *regresión de mínimos cuadrados alternada* (que, en realidad, es idéntico al algoritmo del cálculo recíproco de medias).

CAPÍTULO 15

Agrupación de filas o de columnas

Hasta ahora hemos transformado matrices de datos en mapas en los que representamos filas y columnas como puntos en un espacio continuo, en general un espacio bidimensional. Una forma alternativa de representar una estructura consiste en realizar un análisis de grupos de los perfiles de filas o de columnas. Esta aproximación tiene muchas similitudes con el AC. En ambos análisis descomponemos la inercia de los perfiles, en grupos en el análisis de grupos y en ejes continuos en el AC. El análisis de grupos aplicado a tablas de contingencia conlleva poder disponer de una prueba estadística que nos permite contrastar si existen diferencias entre grupos de filas o de columnas.

Contenido

Agrupación de filas o de columnas

La idea de agrupar objetos es omnipresente en el análisis de datos. La agrupación puede venir dada por una determinada clasificación, o por algún criterio que agrupe objetos similares. En primer lugar, tratemos una agrupación establecida de acuerdo con una determinada variable categórica que clasifique las filas o las

Imagen 15.1: *Frecuencias de las categorías de financiación para 796 investigadores agrupados en cuatro categorías según disciplinas científicas*

	Categoría de financiación					
Disciplina científica	*A*	*B*	*C*	*D*	*E*	*Suma*
Geol/Fis/Est/Mat	17	57	134	35	63	*306*
Bioq/Quim	7	27	62	22	41	*159*
Zool/Micr/Biol	4	33	89	57	60	*243*
Ing	3	11	25	15	34	*88*
Suma	*31*	*128*	*310*	*129*	*198*	*796*

columnas de una tabla. Consideremos otra vez el ejemplo sobre la financiación de la investigación científica, y supongamos que existe una agrupación predeterminada de las disciplinas científicas en cuatro grupos, según las facultades de una determinada universidad: {Geología, Física, Estadística, Matemáticas}, {Bioquímica, Química}, {Zoología, Microbiología, Botánica} e {Ingeniería}. Como apuntamos en el capítulo 12, cuando definimos una variable categórica sobre las filas, como en este ejemplo, cada categoría de esta variable categórica define una fila adicional en la tabla que reúne las frecuencias afectadas por dicha categoría. Así, las diez filas de la imagen 10.1 se convierten en cuatro filas que corresponden a los cuatro grupos que mostramos en la imagen 15.1. El AC de los datos originales de la imagen 10.1 tenían una inercia total de 0,08288, mientras que si realizamos el AC de los datos de la imagen 15.1, la inercia total es de 0,04386. Cuando reunimos puntos, se produce una pérdida de inercia, de la misma manera que cuando dividimos filas o columnas, siguiendo algún criterio de subclasificación, se produce un incremento de inercia.

Inercia inter e intra grupos

La inercia de la tabla de grupos de la imagen 15.1 corresponde a la *inercia intergrupos*, ya que mide la variabilidad entre los cuatro grupos de filas de la tabla. Llamamos *inercia intragrupos*, a la diferencia entre la inercia total, 0,08288 y la inercia intergrupos, 0,04386. Esta diferencia mide la variabilidad que se pierde, dentro de los grupos, cuando unimos filas en grupos. Esta descomposición de la inercia es un resultado clásico del análisis de la varianza, en general aplicado a una sola variable, aunque también se puede aplicar a datos multivariantes. En AC, la inercia total de las filas viene dada por $\sum_i r_i d_i^2$ (fómula 4.7), siendo d_i la distancia χ^2 entre $\mathbf{a}_i$ y **c,** donde $\mathbf{a}_i$, es una fila que tiene asociada una masa r_i y **c** es el perfil fila medio (centroide) igual a las masas de las columnas. La inercia intergrupos se calcula de forma similar, mediante la ecuación $\sum_g \bar{r}_g \bar{d}_g^2$ que se aplica a las filas agrupadas, siendo $\bar{\mathbf{a}}_g$ los perfiles de las filas resultantes de la agrupación, donde $g = 1,..., G$ indica el grupo (aquí $G = 4$), $\bar{r}_g$ es la masa del g-ésimo grupo, resultante de la suma de las masas de los miembros del grupo. Los perfiles $\bar{\mathbf{a}}_g$ siguen teniendo el centroide en **c**, $\bar{d}_g$ son las distancias χ^2 al centroide. La inercia de cada grupo g a su propio centroide $\bar{\mathbf{a}}_g$ la calculamos mediante la expresión $\sum_{i \in g} r_i d_{ig}^2$ don-

Grupo	*Definición*	*Componente*	*Porcentaje sobre cada parte*	*Porcentaje sobre el total*
Inercia intergrupos				
Geol/Fis/Est/Mat	$\overline{r}_1\overline{d}_1^2$	0,01482	33,8%	17,9%
Bioq/Quim	$\overline{r}_2\overline{d}_2^2$	0,00099	2,3%	1,2%
Zool/Micr/Biol	$\overline{r}_3\overline{d}_3^2$	0,01548	35,3%	18,7%
Ing	$\overline{r}_4\overline{d}_4^2$	0,01256	28,6%	15,2%
Total	$\sum_g \overline{r}_g\overline{d}_g^2$	0,04386	100,0%	52,9%
Inercia intragrupos				
Geol/Fis/Est/Mat	$\sum_{i\in1} r_i d_{i1}^2$	0,01842	47,2%	22,2%
Bioq/Quim	$\sum_{i\in2} r_i d_{i2}^2$	0,01064	27,3%	12,8%
Zool/Micr/Biol	$\sum_{i\in3} r_i d_{i3}^2$	0,00996	25,5%	12,0%
Ing	$\sum_{i\in4} r_i d_{i4}^2$	0	0%	0%
Total	$\sum_g \sum_{i\in g} r_i d_{ig}^2$	0,03902	100,0%	47,1%

Imagen 15.2: *Descomposición de la inercia inter e intragrupos, que muestra los valores absolutos expresados como porcentajes con relación a la inercia de cada parte, y con relación a la inercia total. Las sumas de la inercia total intergrupos y la inercia total intragrupos es la inercia total, 0,08288 de la tabla de original (imagen 10.1)*

de d_{ig} es la distancia χ^2 de cada perfil i del grupo g al centroide $\overline{\mathbf{a}}_g$. Sumando estos valores para los cuatro grupos obtenemos la inercia intragrupos. Por tanto, la descomposición de la inercia es:

$$\text{inercia total} = \text{inercia intergrupos} + \text{inercia intragrupos}$$

$$\sum_i r_i d_i^2 = \sum_g \overline{r}_g \overline{d}_g^2 + \sum_g \sum_{i\in g} r_i d_{ig}^2 \qquad (15.1)$$

$$0{,}08288 = 0{,}04386 + 0{,}03902$$

Cálculo de la inercia dentro de un grupo

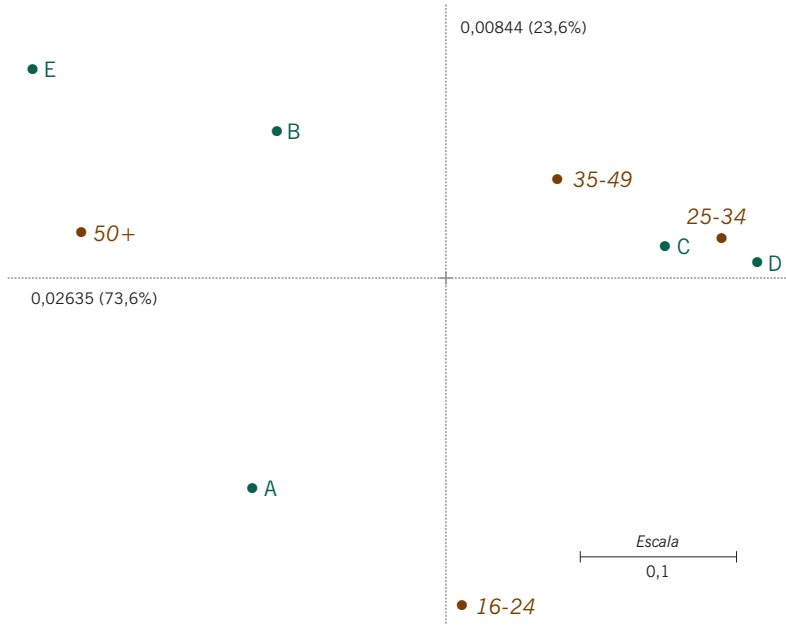
Según lo que acabamos de ver, la inercia intragrupos es igual a 0,03902, pero ¿cuál es la contribución de cada uno de los cuatro grupos? Lo podemos calcular directamente, recordando que, en todos los cálculos de distancias χ^2, debemos utilizar los mismos valores de **c**. Sin embargo, una manera más rápida de hallar esta contribución es aplicar el AC a las matrices resultantes de ir formando, uno a uno, los distintos grupos. Por ejemplo, si formamos el primer grupo reuniendo Geología, Física, Estadística y Matemáticas y analizamos este grupo juntamente con las restantes filas, sin reunir, es decir en total siete filas, la inercia total es 0,06446. Comparando este valor con el valor de la inercia total de los datos originales, 0,08288, la disminución de 0,01842, que corresponde a la inercia intragrupos perdida al formar este grupo. Si ahora reunimos Bioquímica y Química y llevamos a cabo de nuevo el AC de seis grupos, en esta ocasión el valor de la inercia total disminuye hasta 0,05382. Por tanto, la inercia intragrupos atribuible a este grupo es la diferencia, 0,06446 – 0,05382 = 0,01064, y así sucesivamente. En la imagen 15.2 mostramos la descomposición completa de la inercia en valores absolutos y en porcentajes. Fijémonos en que la inercia intragrupos del grupo formado por una sola fila, Ingeniería, es 0.

Conjunto de datos 8: distribución de edades en tiendas de comida

Hasta ahora, hemos realizado la agrupación de filas partiendo de información disponible. Vamos a considerar ahora la posibilidad de formar grupos utilizando un determinado criterio de análisis de grupos. Para ilustrar los cálculos vamos a utilizar una pequeña matriz de datos obtenida de una muestra real de compradores en cinco tiendas de comida distintas. En la imagen 15.3, mostramos la tabla de contingencia de 5 × 4 que hemos obtenido cruzando los datos según tiendas y grupos de edad. El estadístico χ^2 de esta tabla es 25,06, al que le corresponde un valor p de 0,015. Por tanto, existe una asociación significativa entre la edad y la elección de la tienda. Junto con la tabla mostramos el mapa simétrico del AC. Un investigador de mercados estaría interesado en conocer dónde se halla esta asociación significativa. Por ejemplo, estaría interesado en

Imagen 15.3: *Combinación de tiendas de comida con grupos de edad de una muestra de 700 consumidores, y mapa simétrico del AC, que explica el 97,2% de la inercia total de 0,03580*

	Grupo de edad (años)				
Tienda de comida	*16-24*	*25-34*	*35-49*	*50+*	*Suma*
A	37	39	45	64	*185*
B	13	23	33	38	*107*
C	33	69	67	56	*225*
D	16	31	34	22	*103*
E	8	16	21	35	*80*
Suma	*107*	*178*	*200*	*215*	*700*



saber qué tiendas o grupos de tiendas tienen un perfil de edad significativamente distinto de los otros. Observamos que el mayor contraste se halla entre el grupo de mayor edad a la izquierda y el segundo grupo más joven a la derecha (imagen 15.3). La tienda E es la que muestra una mayor asociación al mencionado grupo de mayor edad, mientras que las tiendas C y D tienden más hacia los grupos más jóvenes. El eje vertical contrasta el grupo de edad más joven con los otros. Vemos también que la tienda A se halla hacia el grupo de edad más joven, separada de las restantes tiendas.

Algoritmo de agrupación

Vamos a agrupar las filas y las columnas utilizando un algoritmo de agrupación que trata (al mismo tiempo) de maximizar la inercia intergrupos y de minimizar la inercia intragrupos. En la imagen 15.4 ilustramos, para las filas, dicho algorit-

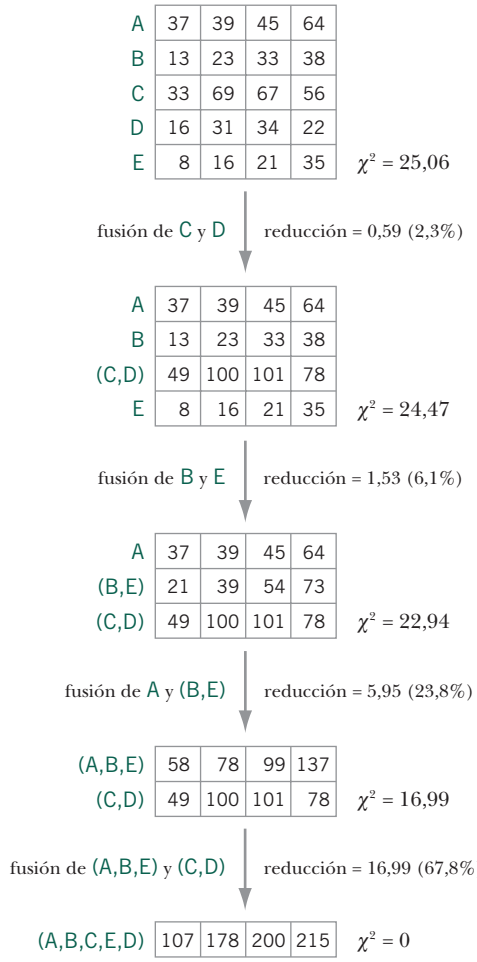


Imagen 15.4:
Pasos en la agrupación de las filas de la imagen 15.1: en cada paso se reúnen las dos filas que conducen a una menor reducción del valor del estadístico χ^2 o, de forma equivalente, a una menor reducción de la inercia intergrupos (para pasar de valores χ^2 a inercia, dividimos por el tamaño de la muestra, N = 700)

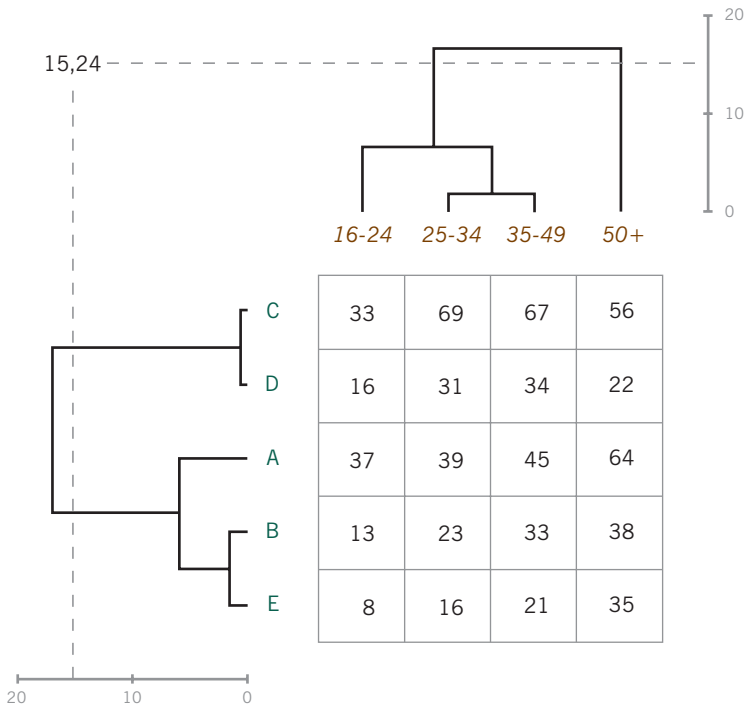
mo de agrupación. Al principio del proceso todas las filas están separadas entre sí, la inercia total intergrupos es igual a la inercia total de la tabla. Cualquier fusión de filas reducirá la inercia intergrupos. Por tanto, el primer paso consiste en identificar qué pares de filas (tiendas) se pueden reunir reduciendo al mínimo la pérdida de inercia. En este sentido, las filas más parecidas son las tiendas C y D. Cuando fusionamos estas dos filas para formar una nueva fila, etiquetada como (C,D), la inercia de la tabla de 4 × 4 resultante se reduce en 0,00084, alcanzándose el valor 0,03496, o en términos de χ^2 la reducción es de 0,59, llegándose al valor de 25,06 (en la imagen 15.4 mostramos los valores χ^2, son los valores de la inercia multiplicados por el tamaño de la muestra: $\chi^2 = 0{,}03496 \times 700 = 25{,}06$). En términos porcentuales, la reducción es del 2,3% tanto en la inercia como en el valor de χ^2. Luego, repetimos el procedimiento para hallar, en la nueva tabla, las filas, que son más parecidas en el sentido mencionado anteriormente. Son las tiendas B y E, lo que conduce a una disminución adicional de χ^2 de 1,53 (6,1%). Ahora la tabla tendrá tres filas etiquetadas como A, (B,E) y (C,D). Repetimos el procedimiento y vemos que la menor reducción se produce cuando la tienda A se une al par (B,E) para formar una nueva fila etiquetada como (A,B,E), χ^2 se reduce en 5,95 unidades adicionales (23,8%). Finalmente, se reúnen las dos filas (A,B,E) y (C,D) para formar una sola fila, que consta de las sumas marginales de las columnas de la tabla original, para la cual el valor de χ^2 es cero. Por tanto la reducción final es de 16,99 (67,8%), que es la inercia de la penúltima tabla de la imagen 15.4. Podemos repetir el procedimiento de la misma manera para las columnas de la tabla.

Representación en árbol de las agrupaciones

Podemos representar gráficamente la reunión sucesiva de filas, llamada *agrupación jerárquica*, como un *árbol binario* o *dendrograma* (se muestra en la imagen 15.5 junto con una agrupación jerárquica similar de las columnas). Fijémonos en que, generalmente, la ordenación original de filas y columnas exige modificaciones para adaptarlas a las representaciones en árbol. En este ejemplo sólo hemos reordenado las filas. Podemos ver en el árbol que las tiendas C y D han sido las primeras en fusionarse. Llamamos *nudo* al punto en el que ocurre esta unión, correspondiendo al mismo una determinada reducción del valor del estadístico χ^2.

Descomposición de la inercia (o del estadístico χ^2)

La descomposición del estadístico χ^2 hasta llegar a cero es la siguiente: 25,06 = 16,99 + 5,95 + 1,53 + 0,59. Dividiendo por 700, el tamaño de la muestra, obtenemos la correspondiente descomposición de la inercia: 0,03580 = 0,02427 + 0,00851 + 0,00218 + 0,00084. Si expresamos las dos descomposiciones anteriores como porcentajes, obtenemos los mismos valores: 67,8%, 23,8%, 6,1% y 2,3%. Hemos seguido un procedimiento de agrupación similar para las columnas, en esta ocasión, la descomposición de la inercia que nos señalan los nudos es la siguiente: 0,03580 = 0,02383 + 0,00938 + 0,00259, que en forma de porcentajes es: 66,6%, 26,2% y 7,2%.



C	33	69	67	56
D	16	31	34	22
A	37	39	45	64
B	13	23	33	38
E	8	16	21	35

Imagen 15.5:
Estructuras en árbol que representan la agrupación jerárquica de filas y de columnas. La agrupación se expresa en términos de χ^2, podemos convertirlo en inercias dividiendo por el tamaño de la muestra, 700. Indicamos el nivel crítico de χ^2, 15,24 (de filas y de columnas)

Decidiendo sobre la agrupación

En un análisis de grupos de este tipo es útil inspeccionar los árboles para decidir el número final de grupos con el que finalmente nos quedamos. Por ejemplo, si nos fijamos en la agrupación de las filas, vemos que existe una gran diferencia entre los dos grupos de tiendas (C,D) y (A,B,E), que viene indicada por el alto valor del nudo en el que se unen estos dos grupos. La descomposición de la inercia nos indica que la formación de estos dos grupos explica el 67,8% de la inercia de las filas. Si separamos la tienda A, como tercer grupo, entonces se explica un 23,8% de inercia adicional. Es decir, con estos tres grupos se explica el 91,6% de la inercia total. Así pues en este tipo de análisis de grupos, interpretamos los porcentajes de inercia asociados con los nudos igual que los porcentajes de inercia de los ejes principales en AC. La decisión sobre hasta qué porcentaje tenemos que llegar para detener la formación de grupos es, en general, informal, se basa en la secuencia de porcentajes y en la interpretación sustantiva de cada nudo o eje principal.

Contraste de hipótesis sobre los grupos de filas o columnas

El estadístico χ^2 de la tabla de contingencia original es significativo (p = 0,015); por tanto, en algún lugar de la tabla tienen que existir diferencias significativas entre perfiles. Precisar, desde un punto de vista estadístico, qué perfiles son significativamente distintos no es sencillo, ya que podemos contrastar la significación de muchos grupos de tiendas. Además, debemos tener en cuenta que, cuando queremos hacer muchas pruebas con unos mismos datos, es necesario ajustar

el nivel de significación. A todo ello hay que añadir que, en este caso, las agrupaciones de tiendas como por ejemplo la C con la D y de la B con la E, las sugieren los mismos datos, no se han establecido como hipótesis antes de la obtención de los datos.

Comparaciones múltiples

Estamos sobre la delicada línea existente entre el análisis exploratorio y el análisis confirmatorio de datos. Intentamos sacar conclusiones a partir de unos datos que hemos obtenido de manera exploratoria sin hipótesis establecidas *a priori*. Afortunadamente, se ha desarrollado un área de la estadística especialmente para este tipo de situaciones, son las llamadas *comparaciones múltiples*. Esta aproximación se utiliza más en el análisis de experimentos cuando se quieren comparar varios «tratamientos» entre sí, que en el análisis de experimentos clásicos sencillo en los que un tratamiento se compara con un control. El procedimiento de comparaciones múltiples permite que cualquier tratamiento (o grupos de tratamientos) se pueda contrastar con cualquier otro. Las decisiones estadísticas se pueden realizar a un nivel de significación preestablecido para proteger todas estas pruebas del llamado «error tipo I», es decir, de obtener un resultado que se deba completamente al azar.

Comparaciones múltiples para tablas de contingencia

Igual que en el caso de los diferentes tratamientos en una situación experimental, podríamos querer contrastar las diferencias entre cualquier par de filas de la tabla original o en las diferencias entre cualquier par de grupos de filas. Si sólo hiciéramos una prueba, calcularíamos la tabla reducida conteniendo dos filas (o grupos) y haríamos una prueba χ^2 de forma habitual. El procedimiento de comparación múltiple desarrollada para esta situación permite contrastar diferencias entre dos filas (o grupos de filas) cualquiera. Para ello, en primer lugar, calculamos el valor del estadístico χ^2 de la tabla reducida; a continuación, para conocer si la prueba es significativa o no, comparamos el valor del estadístico χ^2 calculado con el correspondiente valor crítico de la tabla que aparece en la imagen A.1 (pág. 277) que hemos incluido en el apéndice teórico, A. En esta tabla damos los valores críticos para tablas de contingencia de distinto tamaño a un nivel de significación del 5%. Así por ejemplo, a nuestra tabla de 5×4 le corresponde un valor crítico de 15,24. Por tanto, si el estadístico χ^2 es mayor que 15,24, decimos que las dos filas (o grupos de filas) son significativamente distintas.

Límites del valor de χ^2 para agrupaciones significativas

Podemos utilizar el valor crítico de la prueba de comparaciones múltiples para cualquier grupo de filas o de columnas de la tabla, en particular, para separar de forma estadísticamente significativa los grupos que mostramos en la agrupación jerárquica de la imagen 15.5. Así, con relación a los grupos de edad, vemos que el único contraste estadísticamente significativo se produce entre el grupo de mayor edad (50 o más años) y el resto de grupos; con relación a las tiendas de comida, la diferencia estadística se halla entre el grupo (A,B,E) y el grupo (C,D). Por

tanto, la separación observada en el segundo eje de la imagen 15.3 puede ser debida a la variabilidad aleatoria de los datos observados, ya que no existen diferencias significativas entre el grupo de edad más joven y los restantes grupos. Además, en el segundo eje, la distinción entre el grupo de edad *16-24* y el grupo *35-49*, también es difícil de justificar desde un punto de vista estadístico. Todo ello no significa que no podamos inspeccionar la información original en forma de mapa bidimensional como el la imagen 15.3 (prescindiendo de las consideraciones sobre la significación estadística, pues la información mostrada por los datos siempre es útil). En el capítulo 25, utilizaremos estos mismos valores críticos para llevar a cabo una prueba de significación sobre las inercias principales de una tabla de contingencia.

Agrupación de Ward

El algoritmo de agrupación que hemos descrito en este capítulo es un caso especial de la *agrupación de Ward*. En este tipo de agrupación, los grupos se reúnen según un criterio de distancia mínima que tienen en cuenta los pesos de los puntos que se agrupan. Por tanto, en vez de considerar en cada paso sólo la reducción de χ^2 (o de la inercia), utilizamos las distancias χ^2 entre perfiles y sus masas asociadas. Por ejemplo, la «distancia» entre dos grupos de filas g y h es:

$$\frac{\bar{r}_g \bar{r}_h}{\bar{r}_g + \bar{r}_h} \left\| \bar{\mathbf{a}}_g - \bar{\mathbf{a}}_h \right\|_c^2 \qquad (15.2)$$

donde $\bar{r}_g$ y $\bar{r}_h$ son las masas de los respectivos grupos, y $\left\| \bar{\mathbf{a}}_g - \bar{\mathbf{a}}_h \right\|_c$, la distancia χ^2 entre los perfiles de los grupos:

RESUMEN: Agrupación de filas o de columnas

1. El análisis de grupos de filas o de columnas, consistente en la fusión de filas (o columnas) similares en grupos discretos, proporciona una alternativa al examen de la estructura de los datos.
2. Los resultados de la agrupación se pueden representar gráficamente mediante una estructura en árbol *(dendrograma o árbol binario)*, en el que los nudos indican las uniones sucesivas de las filas (o columnas).
3. La inercia total (o de forma equivalente el estadístico χ^2) de la tabla se reduce lo menos posible en cada nivel sucesivo de agrupación de las filas (o columnas). Este procedimiento de *agrupación de Ward* proporciona una descomposición de la inercia con relación a los nudos del árbol, análogo a la descomposición de inercia con relación a los ejes principales en el análisis de correspondencias.
4. Gracias al procedimiento de *comparaciones múltiples,* podemos contrastar la significación de la inercia explicada por cada nudo, lo que nos permite hacer afirmaciones estadísticas sobre las diferencias intergrupos de filas (o de columnas). Esta prueba sólo se puede aplicar a verdaderas tablas de contingencia.

CAPÍTULO 16

Tablas de múltiples entradas

Hasta ahora, en los mapas de AC, hemos representado las frecuencias de concurrencia de dos variables, dispuestas en tablas de contingencia de dos entradas. En este capítulo vamos a considerar situaciones en las que tenemos datos de más de dos variables y veremos cómo explorar gráficamente este tipo de datos. Una posibilidad es acomodar las tablas de múltiples entradas como si fueran tablas de dos entradas y a continuación llevar a cabo el AC habitual.

Contenido

Introducción de una tercera variable en los datos sobre valoración de la salud

Volvamos a los datos sobre la autopercepción de la salud (conjunto de datos 3) con el que hemos trabajado en los capítulos 6 y 7. Se trata de una muestra, representativa de 6371 españoles que hemos clasificado según edad y autopercepción de su salud (imagen 6.1). La encuesta incluye otras variables como género, educación, región de residencia, etc. Como ejemplo de cómo introducir una tercera variable en AC, vamos a trabajar con la variable más sencilla, el género, una variable que sólo consta de dos categorías. Con esta nueva variable, podemos construir dos tablas de contingencia adicionales: género por grupo de edad y género por salud. La primera tabla puede ser interesante desde un punto de vista demográfico, sin embargo, la segunda es más relevante con relación al tema que estamos tratando (imagen 16.1). Para ver la estructura que presen-

Imagen 16.1:
Cruce de género por autopercepción de la salud, que muestra los perfiles de las filas como porcentajes

Género	*Muy buena*	*Buena*	*Regular*	*Mala*	*Muy mala*	*Suma*
Hombre	448	1789	636	177	39	*3089*
%	*14,5*	*57,9*	*20,6*	*5,7*	*1,3*	
Mujer	369	1753	859	237	64	*3282*
%	*11,2*	*53,4*	*26,2*	*7,2*	*2,0*	
Suma	*817*	*3542*	*1495*	*414*	*103*	*6371*
%	*12,8*	*55,6*	*23,5*	*6,5*	*1,6*	

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

tan los datos, de la primera tabla, no hay necesidad de llevar a cabo un AC; es una tabla de 2 × 5, de una sola dimensión, en la que podemos expresar todos los resultados en porcentajes. Así, vemos que, en general, los hombres tienen una opinión más optimista sobre su salud que las mujeres. En las categorías *muy buena* y *buena* los porcentajes de los hombres son mayores que los de las mujeres, mientras que las mujeres presentan mayores porcentajes en las categorías *regular*, *mala* y *muy mala*.

Interacción entre variables

Anteriormente, en el capítulo 6, vimos que la autopercepción de la salud empeoraba con la edad. En la tabla de la imagen 16.1 podemos ver el efecto del género. Vemos que en promedio los hombres son más optimistas sobre su salud que las mujeres. Queremos saber si el efecto del género se mantiene en todos los grupos de edad, o, si por el contrario, va variando. Puede ocurrir que para un determinado grupo de edad este efecto sea mayor, o incluso puede ocurrir que se dé el efecto contrario, es decir, existe una *interacción*; en este caso una interacción entre edad y género. La ausencia de interacción significaría que en todos los grupos de edad existe la misma disimilitud entre géneros.

Codificación interactiva

Para visualizar la posible interacción entre género y edad codificamos los datos de la siguiente manera: creamos una nueva variable a partir de todas las combinaciones posibles entre género y edad. En este caso tenemos dos géneros y siete grupos de edad, por tanto tenemos 2 × 7 = 14 combinaciones posibles: es lo que llamamos *codificación interactiva*. A continuación, cruzamos la variable codificada interactivamente con las categorías de salud, para, de esta manera, construir la tabla de contingencia que mostramos en la imagen 16.2.

AC de la tabulación codificada interactivamente

A pesar de que, como vimos en el capítulo 6, estos datos son más bien unidimensionales, en la imagen 16.3 mostramos el mapa simétrico bidimensional de la tabla de la imagen 16.2. En este mapa, los pares de puntos muestran las diferencias hombre-mujer de los distintos grupos de edad. Comparando los pares de

Género/Edad	*Muy buena*	*Buena*	*Regular*	*Mala*	*Muy mala*	*Suma*
h16-24	145	402	84	5	3	*639*
h25-34	112	414	74	13	2	*615*
h35-44	80	331	82	24	4	*521*
h45-54	54	231	102	22	6	*415*
h55-64	30	219	119	53	12	*433*
h65-74	18	125	110	35	4	*292*
h75+	9	67	65	25	8	*174*
m16-24	98	387	83	13	3	*584*
m25-34	108	395	90	22	4	*619*
m35-44	67	327	99	17	4	*514*
m45-54	36	238	134	28	10	*446*
m55-64	23	195	187	53	18	*476*
m65-74	26	142	174	63	16	*421*
m75+	11	69	92	41	9	*222*

Imagen 16.2:
*Cruce de la variable codificada interactivamente, género-edad, con la variable autopercepción de la salud (*h *= hombre,* m *= mujer, y siete grupos de edad como en la imagen 6.1). Hemos subdividido cada fila de la imagen 6.1 en dos filas, según su género*

puntos de cada grupo de edad, vemos, de forma consistente, que el punto de las mujeres se halla a la izquierda del correspondiente de los hombres. Ello ilustra el efecto que observamos en la tabla de la imagen 16.1, es decir, que en general, las mujeres son menos optimistas sobre su salud que los hombres. No observamos el fenómeno contrario en ningún grupo de edad, sin embargo, podemos apreciar algunas diferencias en las distancias hombre-mujer. Así, en las edades más tempranas, hasta el grupo de edad 35-44, las distancias hombre-mujer son relativamente pequeñas. En cambio, en el grupo de edad 45-54, podemos observar cambios importantes en la autopercepción de la salud (capítulo 6), que van acompañadas de una mayor diferencia entre hombres y mujeres. Este cambio se mantiene para los grupos de más edad. Vemos que las mujeres del grupo de edad 55-64, incluso son más pesimistas que los hombres del grupo de edad, 65-74 años. De forma similar, las mujeres del grupo de edad 65-74, son más pesimistas que los hombres del grupo de más edad, 75+. Esta diferencia cambiante entre hombres y mujeres de los distintos grupos de edad constituye una evidencia de que existe una interacción género-edad en cuanto a la autopercepción de la salud.

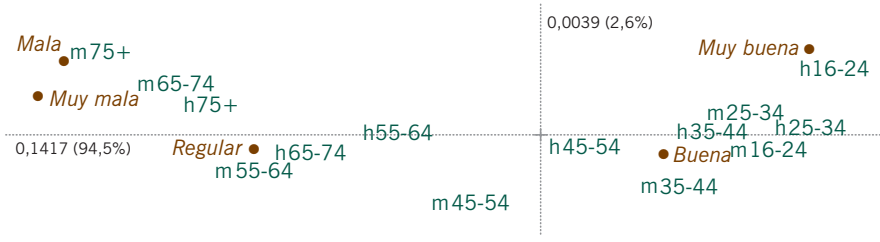


Imagen 16.3:
Mapa simétrico del AC correspondiente al cruce de género por edad, variable codificada interactivamente, con categoría de salud

Imagen 16.4: *Frecuencias de respuesta a la pregunta sobre el trabajo de las mujeres que tienen niños en edad escolar en casa en 24 países*

Países		*T*	*t*	*C*	*?*	*Suma*
AUS	Australia	256	1156	176	191	*1779*
AO	Alemania Occidental	101	1394	581	248	*2324*
AE	Alemania del Este	278	691	62	66	*1097*
GB	Gran Bretaña	161	646	70	107	*984*
IN	Irlanda del Norte	126	394	75	52	*647*
EU	Estados Unidos	482	686	107	172	*1447*
A	Austria	84	632	202	59	*977*
H	Hungría	285	736	447	32	*1500*
I	Italia	171	670	167	10	*1018*
IR	Irlanda	223	424	209	82	*938*
HO	Países Bajos	539	1205	143	81	*1968*
N	Noruega	487	1242	205	153	*2087*
S	Suecia	295	833	39	105	*1272*
CH	Checoslovaquia	228	585	198	13	*1024*
ES	Eslovenia	341	428	222	41	*1032*
P	Polonia	431	425	589	152	*1597*
B	Bulgaria	270	427	335	94	*1126*
RUS	Rusia	175	1154	550	119	*1998*
NZ	Nueva Zelanda	120	754	72	101	*1047*
CAN	Canadá	566	497	108	269	*1440*
F	Filipinas	243	448	484	25	*1200*
IS	Israel	468	664	92	63	*1287*
J	Japón	203	671	313	120	*1307*
E	España	738	1012	514	230	*2494*
Suma		*7271*	*17774*	*5960*	*2585*	*33590*
%		*21,6%*	*52,9%*	*17,7%*	*7,7%*	

Nota: Alemania del Este y Alemania Occidental todavía se tratan de forma separada, igual que Gran Bretaña e Irlanda del Norte). Expresamos los perfiles como porcentajes. Hemos utilizado las siguientes abreviaciones: *T*: tiempo completo, *t*: tiempo parcial, *C*: permanecer en casa, *?*: no sabe/no está seguro/no contesta.
Fuente: Encuesta ISSP sobre la familia y los cambios de rol de género, 1994.

Conjunto de datos 9: opiniones sobre el trabajo de las mujeres

Como ilustración adicional de codificación interactiva, vamos a introducir un nuevo conjunto de datos que utilizaremos varias veces en este y en los siguientes capítulos. Son datos obtenidos de una encuesta de 1994 del Programa Internacional de Investigación (ISSP, *International Social Survey Programme*) sobre la familia y los cambios de rol de género. El total de la encuesta consta de 33.590 individuos. Se llevó a cabo en 24 países (en las encuestas del ISSP, la antigua Alemania del Este y Alemania Occidental se tratan de forma separada, igual que Gran Bretaña e Irlanda del Norte). Vamos a analizar la relación entre variables demográficas y las respuestas a la siguiente pregunta relacionada con la participación de las mujeres en el mercado de trabajo: «Una mujer con un niño en edad escolar en casa, ¿debe trabajar a tiempo completo, a tiempo parcial, o debe permanecer en casa?». Como en todas las encuestas de este tipo, existe una opción adicional de

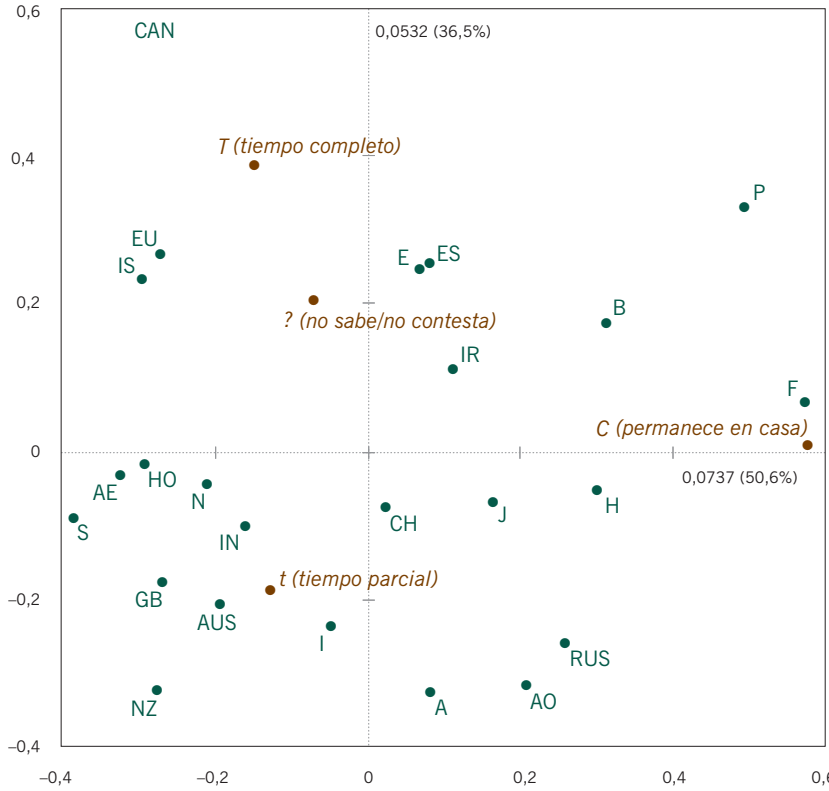


Imagen 16.5:
Mapa simétrico del AC correspondiente a 24 países y a 4 categorías de respuesta (tabla de la imagen 16.4)

respuesta: «no está seguro/no sabe», a la que hemos añadido también algunas, pocas, encuestas sin respuesta (en el capítulo 21, veremos en detalle las no respuestas). Aparte de las respuestas a la pregunta anterior, tenemos datos sobre algunas variables demográficas de cada encuestado. Las tres siguientes tienen, para nosotros, un interés especial: género (dos categorías), edad (6 categorías) y país (24 categorías). En la tabla de la imagen 16.4 mostramos las frecuencias de respuesta de cada país.

Mapa del AC de países por respuestas

En la imagen 16.5, mostramos el mapa de AC correspondiente a esta tabla (hemos cambiado el estilo de nuestros mapas de AC; además, al final del apéndice de cálculo (B), comentaremos algunas opciones de software para la creación de mapas). La interpretación de este mapa es bastante clara; de izquierda a derecha se produce un contraste entre las mujeres que trabajan (a la izquierda) y las que permanecen en casa (a la derecha), en vertical se produce un contraste entre las mujeres que trabajan a tiempo completo (arriba) *versus* las que trabajan a tiempo parcial (abajo). Con relación a este tema, los países más tradicionales son Filipinas y Polonia, mientras que países como Suecia, Alemania del Este, Israel, Nueva

Imagen 16.6: *Frecuencias de respuesta a la pregunta sobre el trabajo de las mujeres cuando tienen un niño en edad escolar en casa, son los datos de los 24 países que aparecen en la imagen 16.4, que se han subdividido según el género del encuestado (cualquier pequeña diferencia en los subtotales de un país y los totales de la imagen 16.4 se deben a unos pocos valores perdidos sobre el género)*

País	*T*	*t*	*C*	*?*	*Suma*
AUSh	117	596	114	82	*909*
AUSm	138	559	60	109	*866*
AOh	43	675	357	123	*1198*
AOm	58	719	224	125	*1126*
AEh	146	316	29	37	*528*
AEm	132	375	33	29	*569*
…	…	…	…	…	…
…	…	…	…	…	…
ISh	220	275	57	29	*581*
ISm	247	387	35	34	*703*
Jh	85	279	171	57	*592*
Jm	118	392	142	63	*715*
Eh	347	445	294	111	*1197*
Em	390	566	218	118	*1292*

Zelanda, Gran Bretaña, y Canadá son los más liberales. A la izquierda, la dirección vertical separa países como Canadá, más favorable a que la mujer trabaje a tiempo completo, de países como, por ejemplo, Nueva Zelanda, más favorable al empleo a tiempo parcial. Recordemos que el origen del mapa representa el perfil medio mostrado en la última fila de la tabla de la imagen 16.4. Por tanto, todos los países situados a la izquierda son más liberales que la media. Si dos países se hallan en la misma posición en el eje horizontal (por ejemplo, Estados Unidos y Gran Bretaña), el país más positivo con relación al eje vertical estará más a favor que el valor de la media, de que las mujeres trabajen a tiempo completo.

Codificación interactiva de género con país

Con el objetivo de visualizar las diferencias hombre-mujer en los distintos países, codificamos género con país. En la tabla de la imagen 16.6 mostramos las primeras y las últimas filas de la tabla de contingencia de 48 × 4. En el mapa de la imagen 16.7, podemos ver que no han cambiado mucho las posiciones de las categorías de respuesta. Sin embargo, es interesante comparar los pares de puntos de cada país. En casi todos los casos, el punto correspondiente a las mujeres queda más a la izquierda que el de su homólogo masculino (Bulgaria es la única excepción). Dentro de un mismo país, las respuestas son sorprendentemente homogéneas en comparación con las grandes diferencias entre países. Los países en los que existe la mayor distancia entre las opiniones de hombres y mujeres se hallan principalmente en el lado conservador del mapa: es el caso de Filipinas, Japón, Irlanda del Norte, Alemania Occidental y España. Sin embargo, en el lado izquierdo del mapa, Australia muestra una de las mayores diferencias hombre-mujer. En este análisis, la inercia es mayor que en el mapa de la imagen 16.5 debido a que la división de las muestras por géneros añade inercia. En realidad, la inercia total de este análisis es de 0,01546, mientras que la inercia del análisis anterior era

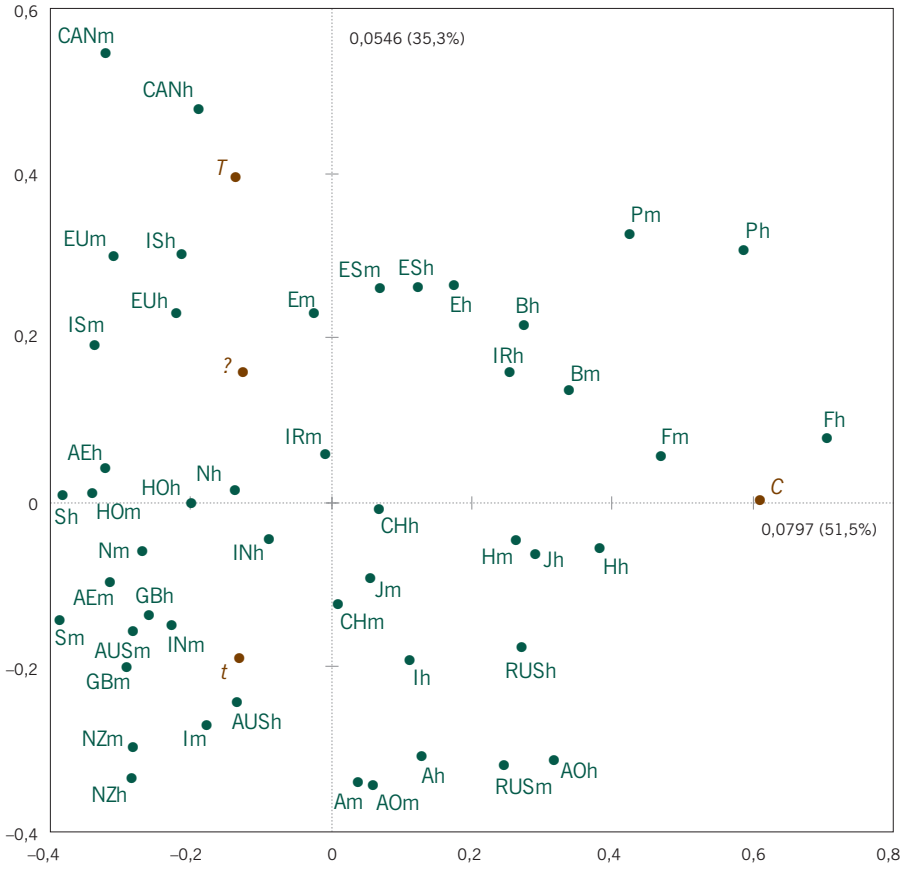


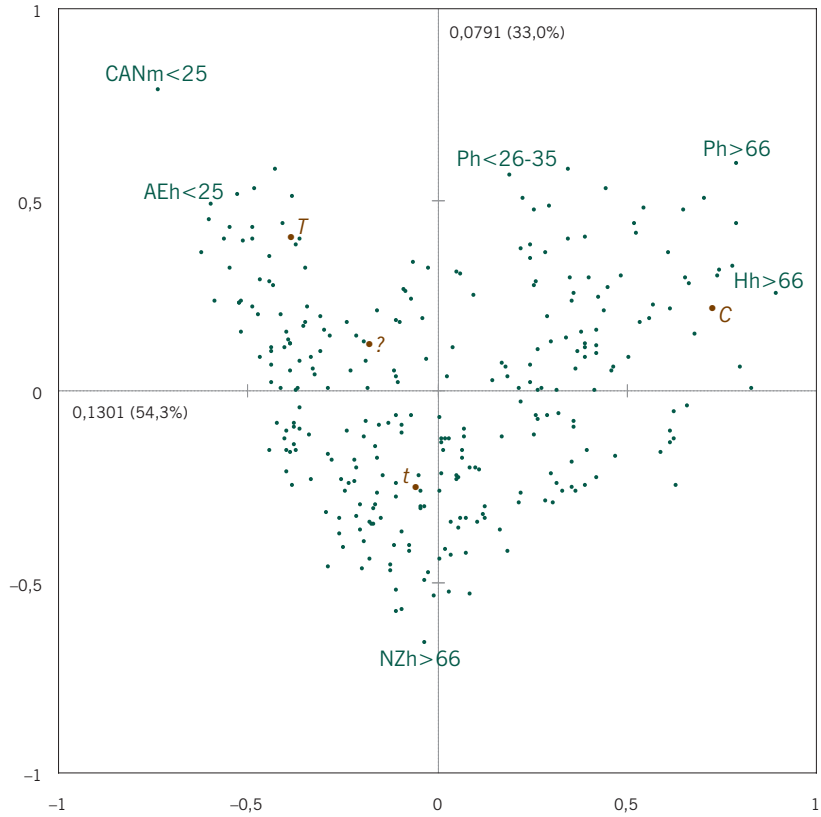
Imagen 16.7:
Mapa simétrico de datos codificados interactivamente (imagen 16.6). Los puntos correspondientes a los hombres se halla, de forma consistente, más a la derecha de los de sus contrapartes femeninas, con la sola excepción de Bulgaria (B)*, país en el que las mujeres son más conservadoras que los hombres*

de 0,01456. Por tanto, en el presente análisis, la inercia atribuible a la diferencia entre géneros es de 0,00090, es decir el 5,8% de la inercia. Como podemos apreciar en el mapa, la mayor diferencia se produce entre países, y no entre géneros.

Codificación interactiva de país, género y grupo de edad

Como las muestras son grandes, podemos seguir dividiéndolas sin problemas. Vamos a hacerlo por la edad, es decir, subdividimos cada grupo país-género en los seis grupos de edad siguientes: hasta 25 años, 26-35, 36-45, 46-55, 56-65, y 66 o más años. Ahora vamos a codificar interactivamente las tres variables en una nueva variable, con 24 × 2 × 6 = 288 categorías en total. En la imagen 16.8 mostramos el mapa de AC de la tabla resultante de 288 × 4. De nuevo vemos que el mapa no cambia demasiado con relación a la posición de las respuestas. Dado que es imposible etiquetar las 228 filas, las hemos representado mediante puntos. Solamente hemos etiquetado algunas observaciones atípicas. Por ejemplo, el grupo más liberal que se halla, lejos, arriba a la izquierda. Es el grupo más joven de mujeres canadienses, el grupo de hasta 25 años. En esta submuestra de 168 mujeres, 101

Imagen 16.8:
Mapa simétrico del AC correspondiente a una codificación de tres entradas. Vemos que los puntos del mapa, que representan los grupos país-género-edad, forman un patrón curvado que surge con frecuencia en los mapas de AC cuando los perfiles forman un gradiente de un extremo (T) a otro (C)



(60,1%) están a favor de que las mujeres con un niño en edad escolar en casa trabajen a tiempo completo, 32 (19,0%) a favor de que lo hagan a tiempo parcial, 3 (1,8%) dicen que las mujeres deben permanecer en casa, y 32 (19,0%) o no saben o no responden (en el capítulo 21 veremos que en la muestra de Canadá hay muchos «no saben»). El grupo de hombres más liberal es el grupo de hombres más joven de Alemania del Este. En el otro extremo, a la derecha, tenemos el grupo de más edad de hombres húngaros y polacos; así, de los 76 hombres polacos de 66 o más años, 16 (21,1%) responden que a tiempo completo, 13 (17,1%) que a tiempo parcial, 41 (53,9%) dicen que las mujeres deben permanecer en casa, las preguntas sin respuesta son 6 (7,9%). En la parte inferior tenemos el grupo de más edad de hombres neozelandeses —son los que están más a favor del trabajo de las mujeres a tiempo parcial.

Configuración en arco (herradura) del mapa

Finalmente, fijémonos en que la nube de puntos del mapa de la imagen 16.8 forma una curva, es lo que en AC llamamos *efecto arco* o *de herradura*. Este fenómeno, habitual, se debe a que el espacio de perfiles es un símplex, en el caso que nos

ocupa un tetraedro de tres dimensiones, ya que tenemos cuatro columnas. Cualquier gradiente de cambio de una esquina extrema del espacio *T* (trabajo a tiempo completo) a la otra *C* (permanecer en casa) seguirá, en este espacio limitado, más una trayectoria curva que recta. Los puntos que se hallan en el interior del arco, como por ejemplo el grupo de hombres polacos de 26-35 años, tienden a estar polarizados en el sentido de que presentan valores elevados en las dos respuestas extremas. Así, de los 141 individuos de este grupo, 45 (31,9%), responden que a tiempo completo, 31 (22,0%) tiempo parcial, 45 (31,9%) permanecer en casa y 20 no responden (14,2%) —en este grupo se observan respuestas por encima de la media en los dos extremos.

RESUMEN: Tablas de múltiples entradas

1. Podemos *codificar interactivamente* dos o más variables categóricas en una nueva variable formada por todas las combinaciones de categorías. Por ejemplo, podemos codificar dos variables con J_1 y J_2 categorías en una nueva variable con $J_1 J_2$ categorías.
2. Cruzamos la variable codificada interactivamente con otra variable y aplicamos el AC. El mapa resultante muestra la estructura de la interacción entre las variables que hemos codificado interactivamente.
3. En general, la codificación interactiva de tablas de múltiples entradas, no va más allá de tres variables, ya que aumenta mucho el número de categorías y por tanto la complejidad del mapa. El nivel de interacción que podemos investigar depende del tamaño de la muestra, ya que el codificado interactivo fragmenta la muestra en submuestras que no deberían ser demasiado pequeñas.

Capítulo 17

Tablas concatenadas

Habitualmente, la investigación en ciencias sociales, exige trabajar con multitud de variables. Es frecuente que en los cuestionarios de las encuestas encontremos respuestas a muchas preguntas, así como muchas características demográficas que, a menudo queremos relacionar con las opiniones de la gente. El AC tiene como ventaja su capacidad para visualizar simultáneamente muchas variables. Sin embargo, como mostramos en el capítulo anterior, debido al gran número de combinaciones de las categorías, existe un límite en el número de variables que podemos codificar interactivamente. Cuando nos encontramos en esta situación, un procedimiento alternativo consiste en codificar los datos en forma de tablas *concatenadas* o *compuestas*. Ello nos permite interpretar en un mapa conjunto las relaciones entre variables demográficas y variables de opinión. En este capítulo veremos ejemplos sobre esta aproximación, tanto para diversas características demográficas como para varias preguntas.

Contenido

Diversas variables demográficas, una pregunta

Vamos a ampliar el conjunto de datos del capítulo 16 relacionado con la opinión de la gente sobre el trabajo de las mujeres. Aparte de país (24 categorías, véanse las abreviaciones en la imagen 16.4), género (2 categorías, *h* y *m*) y grupo de edad (6 categorías, de A1 a A6), vamos a incluir el estado civil (5 catego-

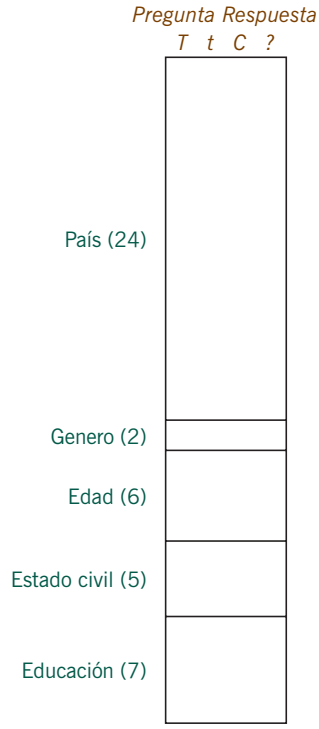


Imagen 17.1: *Tablas concatenadas de contingencia que hemos obtenido cruzando cinco variables demográficas por la respuesta a la pregunta sobre el trabajo de las mujeres (T = tiempo completo, t = tiempo parcial, C = permanecer en casa, ? = no sabe/no contesta)*

rías) y educación (7 categorías), tendremos en total cinco variables demográficas. Las definiciones y las abreviaciones de las dos variables adicionales son las siguientes:

— Estado civil: ca (casado), vi (viudo), di (divorciado), se (separado), so (soltero)

— Educación: E1 (sin educación formal), E2 (educación primaria incompleta), E3 (educación primaria), E4 (educación secundaria incompleta), E5 (educación secundaria), E6 (educación universitaria incompleta), E7 (educación universitaria)

Concatenación como alternativa al codificado interactivo

Codificar interactivamente las cinco variables es completamente imposible. ¡El número de combinaciones sería: 24 × 2 × 6 × 5 × 7 = 10080! Como alternativa, podemos cruzar cada variable demográfica con las respuestas a las preguntas y luego concatenar las tablas de contingencia resultantes, una encima de la otra, como mostramos en la imagen 17.1. La tabla de arriba es la correspondiente a la tabla de la imagen 16.4, con los países como filas. A continuación la tabla con los dos géneros como filas, luego las seis filas de los grupos de edad, y así sucesivamente. Este tipo de codificación no nos permitirá analizar interacciones, la podemos contemplar como una especie de AC medio de las cinco tablas individuales.

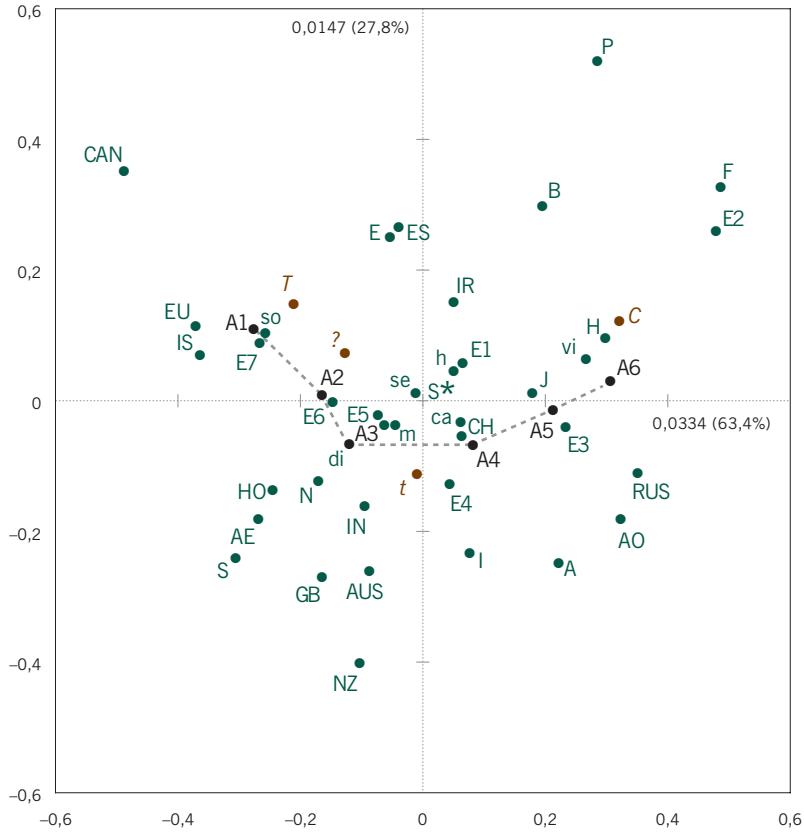


Imagen 17.2:
Mapa simétrico del AC correspondiente a la agrupación de las cinco tablas de contingencia que se muestran de forma esquemática en la imagen 17.1; inercia total = 0,05271, porcentaje de inercia del mapa: 91,2%

AC de tablas concatenadas

Aplicando el AC a la matriz de 44 × 4, correspondiente a las tablas concatenadas, obtenemos el mapa de la imagen 17.2. Las posiciones relativas de las cuatro respuestas categóricas, *T*, *t*, *C* y *?*, aparecen casi como en el mapa de la imagen 16.8. Comparando con los mapas de las imágenes 16.5 y 16.7 vemos que se ha producido una ligera rotación en las posiciones (en el epílogo trataremos sobre rotación). Cada categoría demográfica define un perfil que se sitúa en el mapa con relación a las cuatro respuestas categóricas. Son de especial interés las siguientes características del mapa:

- Podemos unir las categorías de las variables ordinales como, por ejemplo, edad. Vemos, como era de esperar, que edad sigue una trayectoria curva con relación a las respuestas *T*-*t*-*C*, de las más liberales a las más tradicionales.
- Educación sigue un esquema similar pero de derecha a izquierda, excepto para la categoría E1 (sin educación formal) que se halla cerca de la media.
- Las categorías correspondientes al estado civil, muestran a so (solteros) en el lado liberal, y a vi (viudos) en el lado tradicional, probablemente estas variables están correlacionadas con el grupo de edad.

- Los puntos correspondientes a género, h y m se hallan opuestos entre sí con relación a la media. En general, muestran las diferencias entre hombres y mujeres en los distintos países (vimos diferencias específicas en la imagen 16.7).
- De todas las variables demográficas, las diferencias entre países siguen siendo las más importantes.
- Países como España, Eslovenia, Irlanda y Bulgaria, que se hallan dentro del arco, son países polarizados con valores por encima de la media tanto para *T* (trabajo a tiempo completo) como para *C* (permanecer en casa).
- Vemos que el punto correspondiente a no respuesta *?* se halla en el lado liberal del mapa. Es decir, su perfil con relación a las variables demográficas es más similar a *T* que a *C* (en el capítulo 21 veremos que Canadá, tiene un alto porcentaje de no respuestas).

Limitaciones de la interpretación del análisis de tablas concatenadas

Es importante que nos demos cuenta de que el mapa de la imagen 17.2 únicamente muestra asociaciones entre variables demográficas y respuestas a las preguntas. No muestra relaciones entre las variables demográficas entre sí. Con las tablas concatenadas, no obtenemos información sobre la relación entre edad, educación y país. El hecho de que el grupo de edad más joven A1, el nivel de educación más alto E7, y países como Canadá, Estados Unidos e Israel se hallen todos en el lado izquierdo no significa que en estos países predominen jóvenes de nivel de educación más alto. Dado que las variables se relacionan de forma separada con las respuestas, la interpretación correcta es que los porcentajes de respuesta con relación a *T* (trabajo a tiempo completo) del grupo de edad más joven, el nivel de educación más elevado y los mencionados países se hallan por encima de la media. Para confirmar cualquier relación entre las variables demográficas entre sí, tendríamos que cruzarlas y luego analizar la tabla de contingencia resultante.

Descomposición de la inercia en tablas concatenadas

Un resultado muy útil para este capítulo y capítulos siguientes es que cuando cruzamos y agrupamos los mismos individuos, como podemos ver en la imagen 17.1, la inercia total del AC de la tabla concatenada es la media de las inercias de los AC de cada tabla. Lo podemos comprobar calculando las inercias de cada una de las cinco tablas de la imagen 17.1:

Tabla	*Inercia*
País	0,14558
Género	0,00452
Edad	0,04216
Estado civil	0,02675
Educación	0,04221
Media	*0,05224*

La inercia total del análisis de la tabla concatenada es de 0,05271, algo superior al valor medio que aparece en la tabla anterior. Ello es debido a que faltan datos de algunas variables demográficas. Los totales de las tablas varían de 30.471 para educación (a título indicativo señalemos que, en la muestra española, la respuesta a educación se codificaron como «no disponible» para todos los individuos de la muestra) a 33.590 (la muestra completa) para edad y país. El hecho de que los totales de las distintas tablas no sean iguales hace que aumente ligeramente la inercia total de la tabla compuesta, ya que existen pequeñas diferencias en los totales de las columnas de las distintas tablas. Para que la descomposición anterior sea exacta, el total de todas las tablas —y, en consecuencia, también los valores marginales— tienen que ser iguales. La tabla de inercias anterior también muestra que la mayor inercia se produce entre países. Por tanto, la relación entre las respuestas a las preguntas y los países debe dominar los resultados.

Concatenación de tablas por filas y por columnas

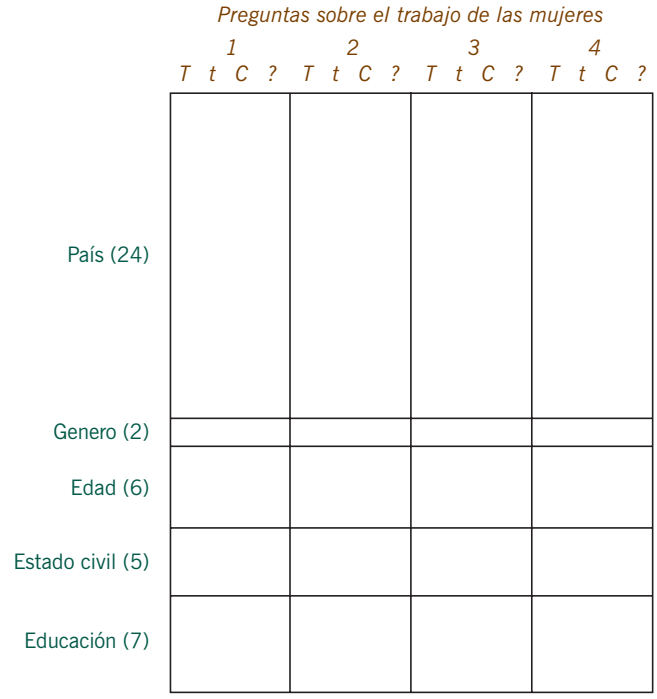
La concatenación de tablas se puede ampliar introduciendo nuevas preguntas que hayamos cruzado con los datos demográficos. En la encuesta ISSP, de la que hemos obtenido estos datos, había cuatro preguntas relacionadas con la opinión de los encuestados sobre el trabajo de las mujeres. Cada una de ellas con las mismas cuatro respuestas: trabajo a tiempo completo, trabajo a tiempo parcial, permanecer en casa y una categoría adicional que incluye todos los tipos de «no respuesta». En concreto, las preguntas hacían referencia a los siguientes supuestos sobre la situación de las mujeres: (1) casadas con hijos, (2) con un hijo en edad preescolar en casa, (3) con un hijo en edad escolar en casa (la pregunta que hemos estado analizando hasta ahora) y (4) con todos los hijos viviendo fuera de casa. Podemos cruzar cada una de las cinco variables demográficas con cada una de estas cuatro preguntas, con lo que obtenemos 20 tablas de contingencia, que asimismo podemos concatenar por filas y por columnas, como mostramos de forma esquemática en la imagen 17.3.

AC de tablas concatenadas por filas y por columnas

Aplicando el AC a las 20 tablas agrupadas en cinco filas y cuatro columnas obtenemos el mapa de la imagen 17.4. Hemos representado las categorías por puntos. Son de especial interés las siguientes características del mapa:

- Las 16 columnas forman una estructura en arco incluso más clara que la que vimos anteriormente, que se extiende desde *2T* y *3T*, arriba a la izquierda, hacia *3t*, *4t* y *2C*, en la parte baja, y luego sube hacia *4C* y *1C*, arriba a la derecha. Este es el resultado típico del AC cuando existe lo que los ecólogos llaman un *gradiente* en los datos. Aquí el gradiente se produce en la dispersión de opiniones, de liberales a tradicionales. A lo largo de este gradiente, podemos ordenar, de forma aproximada, las respuestas categóricas, de la siguiente manera (por el momento omitimos las no respuestas de la discusión *(?)*):

Imagen 17.3:
Tablas concatenadas de contingencia que se han obtenido cruzando cinco variables demográficas con las respuestas a las preguntas sobre el trabajo de las mujeres (T = tiempo completo, t = tiempo parcial, C = permanecer en casa, ? = no sabe/no contesta)



2T 3T 2t y 4T 1T 3t 4t 2C 1t 3C 4C 1C

lo que muestra cómo las columnas se alinean de las extremadamente liberales a la izquierda (las mujeres pueden trabajar a tiempo completo incluso con hijos en casa) a las extremadamente tradicionales a la derecha (las mujeres deben permanecer en casa incluso si los hijos no viven en ella).

- La mayor parte de las filas se sitúa a lo largo de esta curva. Sin embargo existe una sustancial dispersión en la segunda dimensión, que opone a países con una opinión polarizada en sus respuestas (parte superior del mapa, concretamente España) con países con la mayor parte de respuestas en las categorías intermedias del gradiente (parte inferior del mapa, como por ejemplo Austria y Alemania Occidental).
- Los cuatro puntos correspondientes a las no respuestas se hallan juntos formando un pequeño grupo, justo a la izquierda de la media —de hecho, estos puntos quedan mejor representados en la tercera dimensión de este análisis—. Es decir, hay que imaginarlos como si salieran verticalmente de la hoja, la tercera dimensión ordena los grupos demográficos con relación a los porcentajes de no respuesta en las cuatro preguntas.

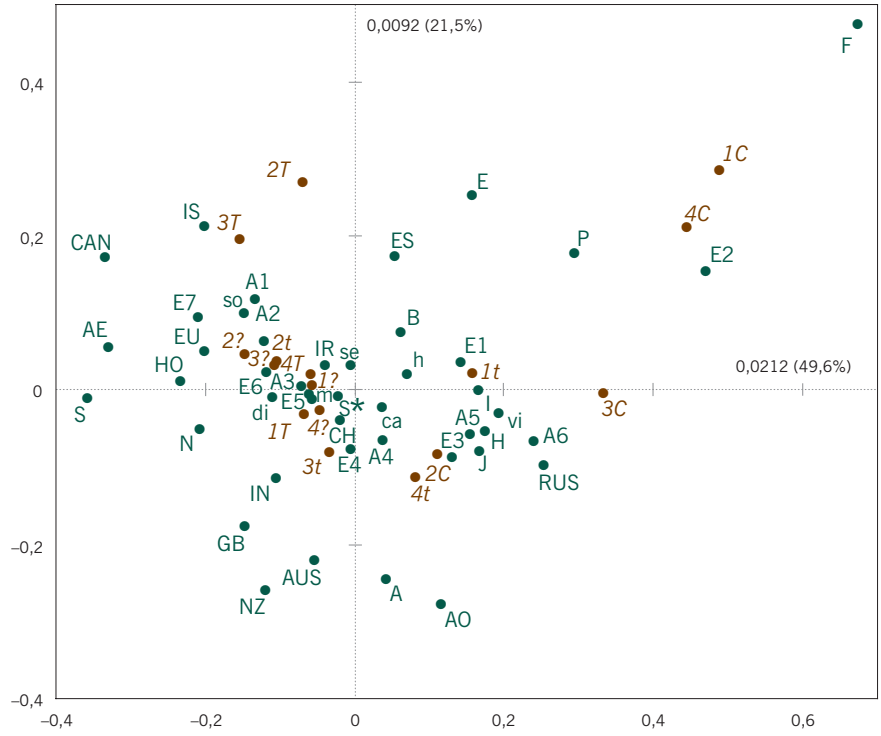


Imagen 17.4:
Mapa simétrico del AC correspondiente a la concatenación de las 20 tablas de contingencia que mostramos de forma esquemática en la imagen 17.3; inercia total = 0,0427; porcentaje de inercia del mapa = 71,1%

Descomposición de la inercia de la tabla concatenada

Aquí también podemos aplicar los resultados que vimos anteriormente sobre la descomposición de la inercia. En primer lugar, si todas las tablas de la tabla concatenada tuvieran exactamente el mismo número de individuos, la inercia de la tabla concatenada sería, exactamente, la media de las inercias de las distintas tablas de contingencia que la forman. Veámoslo de manera más formal. Supongamos que $\mathbf{N}_{qs}$, $q = 1,...,Q$, $s = 1,...,S$ son tablas de contingencia que cruzan, dos a dos, Q variables categóricas, con S variables categóricas, todas ellas con el mismo número n de individuos (en nuestro ejemplo, $Q = 5$ y $S = 4$). Sea $\mathbf{N}$ la tabla concatenada formada uniendo vertical y horizontalmente las Q × S tablas. Entonces:

$$\text{inercia}\,(\mathbf{N}) = \frac{1}{QS}\sum_{q=1}^{Q}\sum_{s=1}^{S}\text{inercia}\,(\mathbf{N}_{qs}) \qquad (17.1)$$

Este resultado es aproximado si hay pérdida de datos en alguna de las tablas de contingencia. En el ejemplo que nos ocupa hemos añadido los valores perdidos de las cuatro preguntas sobre el trabajo de las mujeres a las categorías *?* respectivas, que incluyen respuestas como, por ejemplo, «no sabe». Existen muy pocos valores perdidos correspondientes a variables demográficas de los diferentes países. La falta de algunos valores hará que el resultado de (17.1) no se cumpla exacta-

mente. En efecto, la inercia de la tabla concatenada **N** (lado derecho de (17.1)) aumentará en una pequeña cantidad ε, debido a diferencias en las frecuencias marginales, de manera que (17.1) se convierte en:

$$\text{inercia}(\mathbf{N}) = \frac{1}{QS}\sum_{q=1}^{Q}\sum_{s=1}^{S}\text{inercia}(\mathbf{N}_{qs}) + \varepsilon \qquad (17.2)$$

En la tabla de la imagen 17.5 tenemos los valores de las inercias de todas las tablas de contingencia, así como las medias de las filas, las columnas y la media total. Como era de esperar, la inercia total de la tabla concatenada es ligeramente superior (0,04273) a la media de las inercias de las tablas que la componen (0,04239), la diferencia es del 0,8%. En la tabla de la imagen 17.6 mostramos las inercias de la tabla de la imagen 17.5 expresadas en tantos por mil con relación al valor de 0,04273 × 20 (el valor a la izquierda de (17.2), multiplicado por $QS = 20$) para hacer más cómodo el análisis, igual que hicimos cuando interpretamos las contribuciones a la inercia en el capítulo 11. Estos resultados muestran que en promedio los países explican el 65,6% de la inercia del análisis de la tabla concatenada, seguida por educación (13,3%) y edad (11,6%). La pregunta 3 es la que, en general, presenta mayores inercias (30,6% de la inercia total) —es decir, para esta pregunta existen más diferencias entre los grupos demográficos—, mientras que, para la pregunta 4, las inercias son generalmente menores (21,6% de la inercia total). El total de esta tabla, 992, indica, que la contribución de ε a la inercia es del 0,8% [ecuación (17.2)]. Podemos explicar este valor por las pequeñas diferencias existentes entre las sumas marginales de las 20 tablas de contingencia, diferencias ocasionadas por la pérdida de datos.

Los mapas sólo muestran las asociaciones entre las respuestas y los grupos demográficos, no entre las respuestas entre sí

Una vez más hacemos hincapié en los límites de la interpretación de mapas como el de la imagen 17.4. Con relación a las cuatro preguntas, debemos recordar que no estamos analizando las asociaciones *entre* estas preguntas, sino que lo que analizamos son las asociaciones entre estas variables y las variables demográficas. El análisis de las asociaciones entre variables respuesta es el tema del próximo capítulo, que trata sobre análisis de correspondencias múltiples.

Imagen 17.5:
Inercias de las 20 tablas de contingencia que integran la tabla concatenada de la imagen 17.4; se dan los valores de las medias de las filas y de las columnas, así como el de la media global

VARIABLE	*Pregunta 1*	*Pregunta 2*	*Pregunta 3*	*Pregunta 4*	*Media*
País	0,15268	0,12834	0,14558	0,13410	*0,14018*
Género	0,00821	0,00336	0,00452	0,00484	*0,00523*
Edad	0,01033	0,03359	0,04216	0,01266	*0,02469*
Estado civil	0,00529	0,01341	0,02675	0,00869	*0,01354*
Educación	0,02306	0,02380	0,04221	0,02430	*0,02834*
Media	*0,03991*	*0,04050*	*0,05224*	*0,03692*	*0,04239*

Variable	*Pregunta 1*	*Pregunta 2*	*Pregunta 3*	*Pregunta 4*	*Total*
País	179	150	170	157	*656*
Género	10	4	5	6	*24*
Edad	12	39	49	15	*116*
Estado civil	6	16	31	10	*63*
Educación	27	28	49	28	*133*
Total	*234*	*237*	*306*	*216*	*992*

Imagen 17.6:
Contribuciones a la inercia, expresadas en tantos por mil, de las 20 tablas de contingencia que integran la tabla concatenada; el 0,8% restante es la inercia adicional debido a la diferencia en los valores marginales de las columnas ocasionadas por los valores perdidos

RESUMEN:
Tablas concatenadas

1. Una posibillidad para analizar la relación entre variables demográficas y las respuestas a varias preguntas es agrupar todas las tablas que cruzan los dos conjuntos de variables, y analizar la *tabla concatenada* resultante mediante AC.
2. Debemos interpretar el mapa del AC de la tabla compuesta teniendo presente que la información que analizamos deriva de un conjunto de relaciones entre las respuestas y cada una de las variables demográficas. En el mapa no existe información específica sobre las relaciones existentes entre las respuestas entre sí o entre las variables demográficas entre sí.
3. Debemos contemplar el mapa del AC de una tabla concatenada como un mapa promedio resultante del AC de cada una de las tablas individuales.
4. Cuando los valores marginales de filas y columnas de todas las tablas que integran la tabla concatenada son idénticos, la inercia total de la tabla concatenada es igual a la media de las inercias de cada una de las tablas (ello se cumple cuando todas las tablas tienen el mismo número de individuos). Sin embargo, cuando en algunas tablas faltan individuos, el resultado anterior es sólo aproximado, ya que la inercia total de la tabla concatenada será ligeramente mayor que la media de las inercias de las tablas que la integran.

CAPÍTULO 18

Análisis de correspondencias múltiples

Hasta ahora hemos analizado la relación entre dos variables categóricas, o entre dos conjuntos de variables categóricas, en las que las variables fila eran diferentes de las variables columna. En cambio, en este capítulo y en los dos siguientes, analizaremos la relación existente entre variables similares, mediante el *análisis de correspondencias múltiples*, de forma abreviada ACM. Investigaremos el tipo de asociación existente entre variables y su intensidad. Podemos llevar a cabo el ACM sobre una matriz que contenga los datos codificados de forma binaria, la *matriz binaria*, o bien sobre una matriz formada por todos los cruzamientos posibles entre las variables, la *matriz de Burt*. Ambas posibilidades, muy relacionadas entre sí, presentan algunos inconvenientes que intentaremos solucionar en el capítulo 19, en el que presentamos versiones mejoradas del ACM.

Contenido

Un conjunto de variables categóricas «homogéneas»

En este capítulo analizaremos la relación existente entre más de dos variables, generalmente en el contexto de un solo fenómeno de interés. Por ejemplo, las cuatro variables que vimos en el capítulo 17, sobre el trabajo de las mujeres, podrían ser nuestros datos de interés. También podrían ser las respuestas a preguntas relacionadas con la opinión de la gente sobre la ciencia, o datos que describan

las condiciones ambientales de una serie de localidades. Lo importante es que las variables sean «homogéneas», es decir, que sean sustantivamente similares. Por ejemplo, no debemos mezclar variables de opinión con variables demográficas.

Matriz binaria

Consideremos las mismas cuatro variables que analizamos en el capítulo 17. Para evitar grandes diferencias culturales entre países, utilizaremos sólo datos de Alemania, incluyendo las muestras de Alemania del Este y Alemania Occidental, y así llegamos a un total de 3418 encuestados ($N = 3418$). (Hemos omitido, de las muestras originales, tres casos para los que faltaban algunos datos demográficos: véase el apéndice de cálculo, B.) Nos centraremos en las cuatro preguntas etiquetadas de 1 a 4 sobre el trabajo de las mujeres. Cada pregunta puede tomar cuatro valores categóricos, que al igual que antes etiquetamos de la manera siguiente: *T* (trabajo a tiempo completo), *t* (trabajo a tiempo parcial), *C* (permanecer en casa) y *?* (no sabe/no contesta). La *matriz binaria* resultante será una matriz de 3418 × 16 en la que hemos codificado todas respuestas de forma binaria. En la tabla de la imagen 18.1 mostramos la codificación para las seis primeras filas. Las 16 columnas de la derecha corresponden a la codificación binaria de las 16 posibles respuestas. Para el primer individuo, por ejemplo, las respuestas a las cuatro primeras preguntas son: 1, 3, 2 y 2, es decir *T, t, C* y *?*, que hemos codificado como 1 0 0 0, 0 0 1 0, 0 1 0 0 y 0 1 0 0, respectivamente.

Definición 1 del ACM: AC de la matriz binaria

Podemos definir el ACM como el AC de la matriz binaria. Este análisis proporciona coordenadas para las 3418 filas y las 16 columnas. En el mapa de la imagen 18.2 mostramos las posiciones de las 16 categorías. El primer eje principal muestra que las cuatro categorías de «no respuesta» se hallan juntas, oponiéndose a todas las variables sustantivas. En el análisis anterior de estas preguntas (mapa de la imagen 17.4), en el que relacionamos las respuestas con variables demográficas, las no respuestas no ejercían un papel importante en ninguno de los dos primeros ejes. Sin embargo ahora, debido a que estamos interesados en la relación

Imagen 18.1: *Datos originales y codificación binaria correspondiente a los seis primeros encuestados de* $N = 3418$

Preguntas				*Pregunta 1*				*Pregunta 2*				*Pregunta 3*				*Pregunta 4*			
1	*2*	*3*	*4*	*T*	*t*	*C*	*?*	*T*	*t*	*C*	*?*	*T*	*t*	*C*	*?*	*T*	*t*	*C*	*?*
1	3	2	2	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
2	3	3	2	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0
4	3	3	2	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0
4	4	4	4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
4	4	4	4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
1	3	2	1	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0
⋮	⋮	⋮	⋮		⋮				⋮				⋮				⋮		

… y así sucesivamente para 3418 filas

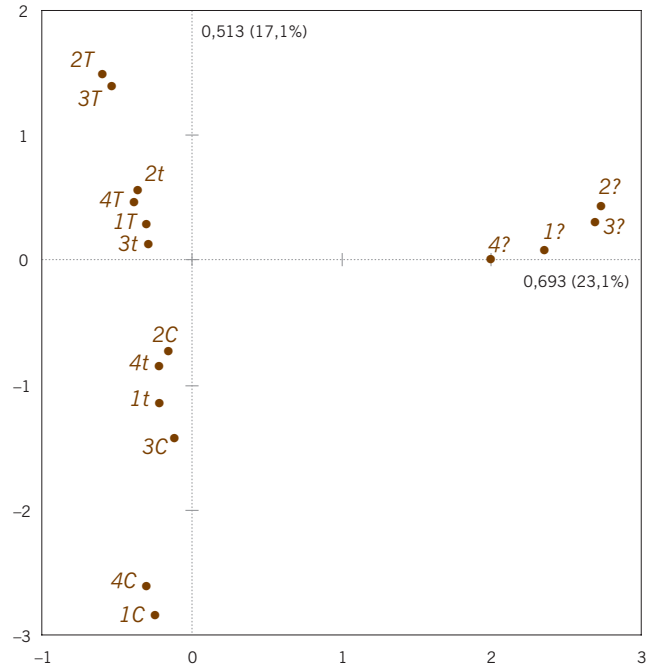


Imagen 18.2:
Mapa del ACM correspondiente a las cuatro preguntas sobre el trabajo de las mujeres; inercia total = 3, porcentaje de inercia del mapa: 40,2%

entre las variables respuesta entre sí, las no respuestas son el hecho más destacable: las personas que no responden una pregunta tienden a hacer lo mismo con las otras (por ejemplo, entre los seis primeros individuos de la tabla de la imagen 18.1 ya hay dos personas con no respuesta para las cuatro preguntas). En el segundo eje del mapa de la imagen 18.2 aparecen alineadas las categorías sustantivas; de las opiniones tradicionales, abajo, a las más liberales, arriba. En el mapa de la imagen 18.3 mostramos la segunda y la tercera dimensión del mapa, que dejan fuera la mayor parte del efecto de las no respuesta. Podemos ver que aquí las posiciones de los puntos son muy similares a las del mapa de la imagen 17.4. El hecho de que el lado liberal de la dimensión horizontal se halle ahora a la derecha no tiene consecuencia alguna sobre la interpretación; siempre es posible invertir un eje (multiplicando todas las coordenadas por –1).

Inercia de la matriz binaria

El cálculo de la inercia total de la matriz binaria es muy simple. Depende sólo del número de preguntas y del número de respuestas categóricas. No depende de sus valores concretos. Supongamos que tenemos Q variables y que cada variable q, tiene J_q categorías, J indica el número total de categorías: $J = \sum_q J_q$ (en nuestro ejemplo, $Q = 4$, $J_q = 4$, $q = 1, \ldots, Q$ y $J = 16$). La matriz binaria, simbolizada por $\mathbf{Z}$, con J columnas, es una matriz compuesta formada por tablas $\mathbf{Z}_q$ agrupadas lateralmente, una para cada variable. En cada tabla, los valores marginales de las filas

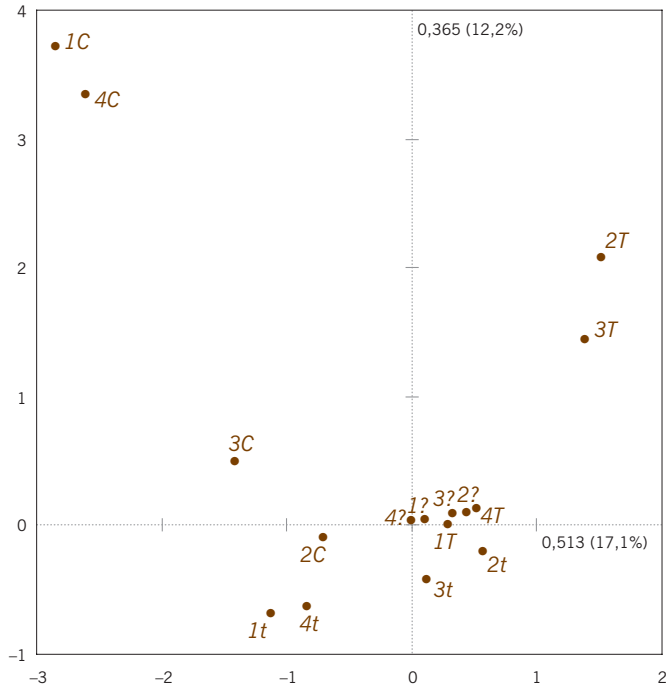


Imagen 18.3: *Mapa del ACM correspondiente a las cuatro preguntas sobre el trabajo de las mujeres, mostrando las dimensiones tercera y cuarta; inercia total = 3, porcentaje de inercia del mapa: 29,3%*

son iguales a una columna de unos. Por tanto, podemos aplicar el resultado de (17.1) del capítulo 17: la inercia total de la matriz binaria es igual a la media de la inercia de las tablas que la componen. Cada tabla $\mathbf{Z}_q$ tiene un solo uno en cada fila, los restantes valores son ceros. Por tanto, estamos ante un ejemplo de matriz para la que todos los perfiles fila se hallan en los vértices, las asociaciones más extremas posibles entre filas y columnas. En consecuencia, en todas las tablas, las inercias de todos los ejes principales serán iguales a 1. Y, por tanto, la inercia total de la tabla $\mathbf{Z}_q$ sera igual a su dimensionalidad, es decir, igual a $J_q - 1$. La inercia de $\mathbf{Z}$ será la media de las inercias de las tablas que la componen:

$$\text{inercia}(\mathbf{Z}) = \frac{1}{Q}\sum_q \text{inercia}(\mathbf{Z}_q) = \frac{1}{Q}\sum_q (J_q - 1) = \frac{J-Q}{Q} \qquad (18.1)$$

Dado que $J - Q$ es la dimensionalidad de $\mathbf{Z}$, la inercia media por dimensión será $1/Q$. Fijémonos en que, en los mapas de las imágenes 18.2 y 18.3, las primeras tres dimensiones que hemos interpretado tienen inercias principales iguales a 0,693, 0,513 y 0,365, respectivamente, todas por encima de la media de 1/4 = 0,25. Utilizamos el valor $1/Q$ como umbral para decidir para qué ejes es interesante interpretar el ACM (similar al valor umbral de 1 de los valores propios en el análisis de componentes principales).

1T	*1t*	*1C*	*1?*	*2T*	*2t*	*2C*	*2?*	*3T*	*3t*	*3C*	*3?*	*4T*	*4t*	*4C*	*4?*
2501	0	0	0	172	1107	1131	91	355	1710	345	91	1766	538	40	157
0	476	0	0	7	129	335	5	16	261	181	18	128	293	17	38
0	0	79	0	1	6	72	0	1	17	61	0	14	21	38	6
0	0	0	362	1	57	108	196	7	96	55	204	51	45	2	264
172	7	1	1	181	0	0	0	127	48	4	2	165	15	0	1
1107	129	6	57	0	1299	0	0	219	997	61	22	972	239	13	75
1131	335	72	108	0	0	1646	0	24	989	573	60	760	616	84	186
91	5	0	196	0	0	0	292	9	50	4	229	62	27	0	203
355	16	1	7	127	219	24	9	379	0	0	0	360	14	1	4
1710	261	17	96	48	997	989	50	0	2084	0	0	1348	567	23	146
345	181	61	55	4	61	573	4	0	0	642	0	202	286	73	81
91	18	0	204	2	22	60	229	0	0	0	313	49	30	0	234
1766	128	14	51	165	972	760	62	360	1348	202	49	1959	0	0	0
538	293	21	45	15	239	616	27	14	567	286	30	0	897	0	0
40	17	38	2	0	13	84	0	1	23	73	0	0	0	97	0
157	38	6	264	1	75	186	203	4	146	81	234	0	0	0	465

Imagen 18.4:
Matriz de Burt que contiene todos los cruces posibles de las cuatro variables del ejemplo sobre la opinión de la gente sobre el trabajo de las mujeres. En la diagonal se hallan los cruces de las variables por ellas mismas

Matriz de Burt

Una estructura alternativa de datos para el ACM es la matriz compuesta por todas las tablas resultantes de cruzar todas las variables de interés dos a dos, la *matriz de Burt,* que mostramos en la tabla de la imagen 18.4 para los datos del ejemplo que estamos considerando. En este caso, la matriz de Burt es una matriz compuesta de 4 × 4, formada por 16 tablas. Con excepción de las tablas de la diagonal, las restantes 12 se obtienen cruzando los valores de dos variables de los 3418 encuestados. La matriz de Burt es simétrica, por tanto, fuera de la diagonal, sólo hay seis cruzamientos distintos que se transponen a ambos lados de la diagonal de la matriz compuesta. Las tablas de la diagonal corresponden a los cruces de las variables por ellas mismas, son matrices diagonales con las frecuencias marginales de la variable en su diagonal. Por ejemplo, las frecuencias marginales de la pregunta 1 son: 2501 para *T*, 476 para *t*, 79 para *C* y 362 para *?*. La matriz de Burt, **B**, se relaciona, de forma sencilla, con la matriz binaria **Z** de la manera siguiente:

$$\mathbf{B} = \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \tag{18.2}$$

Definición 2 del ACM: AC de la matriz de Burt

La otra forma «clásica» de definir el ACM es el AC de la matriz de Burt **B**. Dado que **B** es una matriz simétrica, las soluciones de filas y de columnas son idénticas, por tanto sólo mostramos una de ellas (mapa de la imagen 18.5). Debido a la relación directa (18.2), no es ninguna sorpresa que los resultados de los dos análisis sean similares. De hecho, a primera vista, el mapa de la imagen 18.5 tiene el

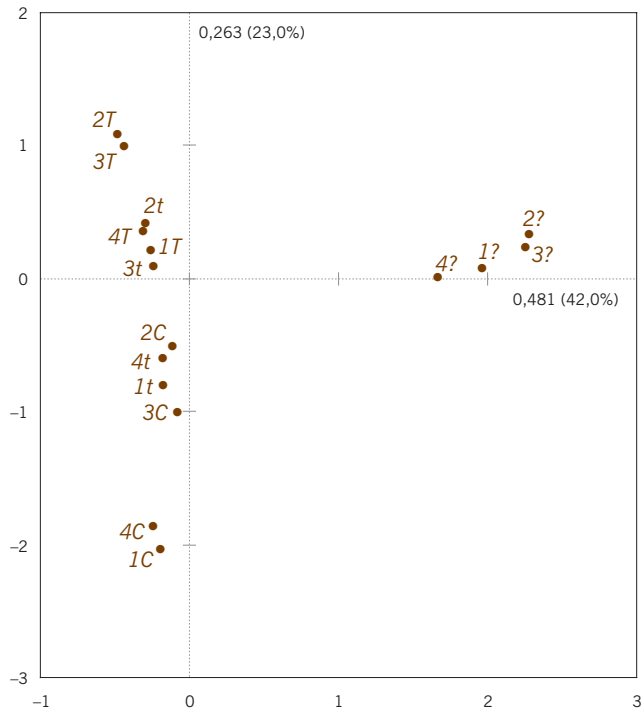


Imagen 18.5:
Mapa del ACM de la matriz de Burt correspondiente a las cuatro preguntas sobre el trabajo de las mujeres, que muestra la primera y la segunda dimensión; inercia total = 1,145, porcentaje de inercia del mapa: 65,0%

mismo aspecto que el mapa de la imagen 18.2, sólo observamos un ligero cambio de escala en los dos ejes. Es la única diferencia entre ambos análisis: la versión Burt del ACM genera coordenadas principales en una escala reducida en comparación con las de la versión binaria. La reducción es relativamente mayor en el segundo eje en comparación con el primero.

Comparación del ACM de las matrices binaria y de Burt

Las dos formas de definir el ACM se relacionan de la siguiente manera:

- En los dos análisis, las coordenadas estándares de las respuestas categóricas son idénticas: es una resultado directo de la relación (18.2).
- También, como resultado de (18.2), las inercias principales del análisis de Burt son los cuadrados de los de la matriz binaria.
- Dado que las inercias principales son menores de 1, sus cuadrados darán valores más pequeños (y en consecuencia las inercias principales más pequeñas tendrán cuadrados relativamente más pequeños). Las coordenadas principales son las coordenadas estándares multiplicadas por la raíz cuadrada de las inercias principales, lo que explica la reducción de escala del mapa de la imagen 18.5 con relación al de la imagen 18.2.

Preguntas	*Pregunta 1*	*Pregunta 2*	*Pregunta 3*	*Pregunta 4*	*Media*
Pregunta 1	3,0000	0,3657	0,4262	0,6457	*1,1094*
Pregunta 2	0,3657	3,0000	0,8942	0,3477	*1,1519*
Pregunta 3	0,4262	0,8942	3,0000	0,4823	*1,2007*
Pregunta 4	0,6457	0,3477	0,4823	3,0000	*1,1189*
Media	*1,1094*	*1,1519*	*1,2007*	*1,1189*	*1,1452*

Imagen 18.6: *Inercias, obtenidas aplicando el AC de cada una de las 16 subtablas, de la matriz de Burt*

- En consecuencia, los porcentajes de inercia serán siempre mayores en el análisis de Burt.

Inercia de la matriz de Burt

Todas las subtablas que componen la matriz de Burt tienen los mismos valores marginales totales de filas y columnas. Así, se cumple de forma exacta el resultado (17.1): la inercia de **B** será la media de las inercias de las subtablas $\mathbf{B}_{qs}$ que lo componen. En la tabla de la imagen 18.6 mostramos las 16 inercias individuales de la matriz de Burt, así como las medias de sus filas y de sus columnas. La media global es igual a la inercia total de **B**, es decir 1,145. En esta tabla, las inercias de las matrices de la diagonal son exactamente igual a 3. Las inercias cumplen lo que vimos en (18.1) para las inercias de las tablas de la matriz binaria —hay $J_q \times J_q$ tablas de dimensionalidad $J_q - 1$ con una asociación perfecta fila-columna—; por tanto, la inercia máxima es igual al número de dimensiones. Los altos valores de las inercias de las matrices de la diagonal de la tabla de la imagen 18.6 explican porqué la inercia total de la matriz de Burt es tan alta, ello también explica los bajos porcentajes de inercia de los ejes. En el próximo capítulo retomaremos este tema.

Situación de variables adicionales en el mapa

Supongamos que queremos relacionar las variables demográficas (género, edad, etc.) con las asociaciones observadas en los mapas de ACM. Existen dos maneras similares de hacerlo. La primera posibilidad consiste en codificar estas variables adicionales como variables binarias y añadirlas como columnas adicionales a la matriz binaria. La segunda posibilidad es cruzar las variables demográficas con las cuatro preguntas, como hicimos en el análisis de matrices compuestas del capítulo 17, y añadir estas tablas de contingencia como filas adicionales de la matriz binaria o como filas (o columnas) adicionales a la matriz de Burt. En el próximo capítulo veremos que, en la práctica, la segunda opción es mejor, ya que nos permite utilizar las versiones mejoradas del ACM. Ambas opciones proporcionan las mismas posiciones de los puntos adicionales, además de tener la misma interpretación como posiciones medias de los casos pertenecientes a una determinada categoría demográfica. En el mapa de la imagen 18.7 mostramos las posiciones de las cinco variables demográficas que vimos anteriormente, las podríamos sobreponer a los mapas de las imágenes 18.2 y 18.5.

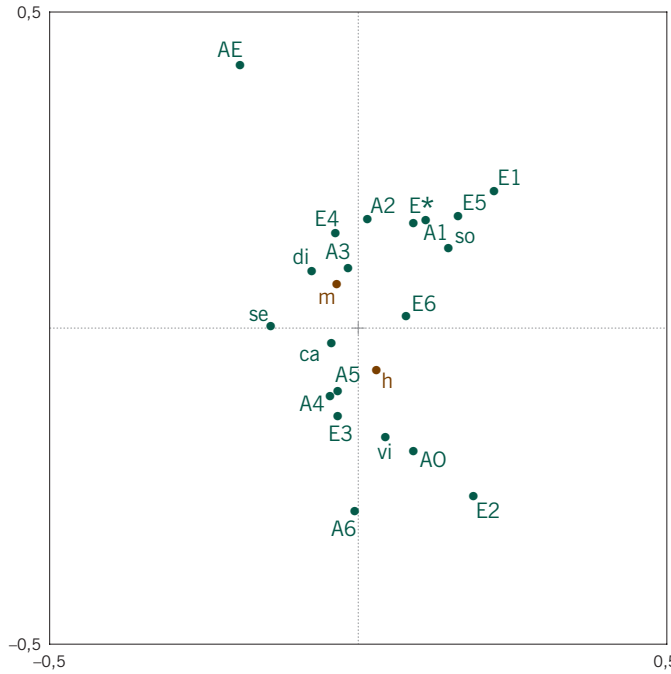


Imagen 18.7:
Variables adicionales con relación a los dos primeros ejes principales, que podríamos superponer a los mapas de la imágenes 18.2 o 18.5. Estos puntos ocupan una pequeña área del mapa (fijémonos en la escala); de todas formas, presentarán una mayor dispersión en el mapa de la matriz de Burt que en el de la matriz binaria

Interpretación de los puntos adicionales

Las posiciones de las respuestas en las dos primeras dimensiones del mapa de la imagen 18.2 (igual que en el de la imagen 18.5) indican que cuanto más a la derecha se halle una categoría demográfica, mayor será la frecuencia de no respuestas. Cuanto más arriba se halle una categoría, más liberales serán las opiniones, y cuanto más abajo, más tradicionales serán las opiniones. En consecuencia, Alemania Occidental es más tradicional y tiene una mayor proporción de no respuestas que Alemania del Este. Situación prácticamente idéntica al contraste hombre-mujer (h-m), pero no tan pronunciado como la diferencia entre las dos regiones alemanas. Los grupos de edad muestran la misma tendencia que vimos anteriormente con los jóvenes (A1) en la parte de arriba (liberal) y los de mayor edad (A6) abajo (tradicionales). Los niveles de educación más bajos tienen frecuencias de no respuesta más elevadas, mientras que los niveles educativos más altos tienden a tener opiniones más liberales, lo que no ocurre con los niveles educativos más bajos E1 y E2. Con relación a los estados civiles, los solteros (so) están por encima de la media con relación a la no respuesta y a las actitudes liberales, oponiéndose a los separados (se) que presentan una ocurrencia de no respuestas. Sin embargo, se hallan en la media respecto a la dimensión liberal-tradicional.

RESUMEN: Análisis de correspondencias múltiples

1. El ACM se ocupa de las relaciones entre un conjunto de variables; en general, variables homogéneas en cuanto hacen referencia a un mismo tema, siendo además las escalas de respuesta iguales.
2. Podemos recodificar las variables en la matriz binaria, que tiene tantas filas como casos y tantas columnas como categorías de respuesta. En las filas (es decir, los casos), los valores son todos 0 con excepción de un 1 que indica una categoría en particular de las variables.
3. La *matriz de Burt*, una matriz cuadrada simétrica, está formada por tablas de contingencia de dos entradas resultantes del cruce de todos los pares de variables. En la diagonal se hallan los cruces de las variables por ellas mismas.
4. Las dos definiciones alternativas de ACM, resultantes de la aplicación del AC a la matriz binaria o a la matriz de Burt, son prácticamente equivalentes. Las dos dan coordenadas estándares idénticas para los puntos correspondientes a las categorías.
5. La diferencia entre las dos definiciones se halla en las inercias principales: las de la matriz de Burt son los cuadrados de las de la matriz binaria. En consecuencia, los porcentajes de inercia del análisis de Burt son siempre más elevados que los del análisis binario.
6. Sin embargo, la codificación hace que los porcentajes de inercia de las dimensiones de los mapas sean artificialmente bajos, lo que conlleva una subestimación de la calidad de los mapas.

CAPÍTULO 19

Análisis de correspondencias conjunto

Ampliar el AC simple de una tabla de contingencia de dos entradas a muchas variables no es tan sencillo. Como hemos visto en el capítulo anterior, la estrategia habitual es aplicar el AC a la matriz binaria o la matriz de Burt. Sin embargo, la geometría de estas aproximaciones ya no es tan clara; además, aparte de que la inercia total del AC tiene poco sentido, los porcentajes de inercia explicados por el mapa son bajos. En el mapa de ACM basado en la matriz de Burt, intentamos visualizar la matriz completa y, sin embargo, en realidad estamos interesados sólo en las tablas de contingencia que se hallan fuera de la diagonal. Es decir, en las tablas que cruzan pares de variables distintas. En este capítulo veremos que el análisis de correspondencias conjunto (ACCo) intenta superar este inconveniente, centrándose en las tablas que se hallan fuera de la diagonal, lo que lleva a una mejor determinación de la inercia total y a una mejor representación de los datos en los mapas.

Contenido

El ACM se ajusta mal debido a que se ha exagerado la inercia total

La tabla de la imagen 18.6 proporciona las inercias de cada una de las tablas que componen la matriz de Burt, así como la de su media, que es la inercia total de la matriz de Burt del ejemplo sobre el trabajo de las mujeres. El valor de esta media viene dado en gran medida por las inercias de las tablas de la diagonal, cuyos valores son iguales

Imagen 19.1: *Porcentaje de inercia de cada una de las 16 tablas de la matriz de Burt explicado por el mapa bidimensional del ACM*

Preguntas	*Pregunta 1*	*Pregunta 2*	*Pregunta 3*	*Pregunta 4*	*Porcentaje pregunta*
Pregunta 1	51,9	78,4	82,5	80,4	*61,2*
Pregunta 2	78,4	55,5	88,2	76,6	*65,3*
Pregunta 3	82,5	88,2	59,6	86,6	*69,7*
Pregunta 4	80,4	76,6	86,6	54,6	*63,5*

al número de categorías de la variable correspondiente menos 1 (4 – 1 = 3, para las variables del ejemplo que consideramos). Como el análisis intenta explicar la inercia de toda la tabla, las elevadas inercias de la diagonal afectarán gravemente el ajuste de la tabla. Por ejemplo, en la tabla de la imagen 19.1 mostramos el porcentaje de inercia de cada tabla explicado por el mapa bidimensional del ACM. Dicho mapa llega a explicar el 65% de la inercia total, observamos que las tablas situadas fuera de la diagonal se explican mucho mejor que las situadas en la diagonal. Sumando las inercias explicadas de las tablas situadas fuera de la diagonal y expresando esta suma con relación a la suma de total, vemos que el mapa bidimensional del ACM explica el 83,2% de la inercia de estas tablas (podemos obtener los valores de inercia de cada tabla a partir de los porcentajes de la tabla de la imagen 19.1 y de los valores de las inercias totales de la tabla de la imagen 18.6). De forma similar, podemos ver que el mapa bidimensional del ACM explica el 55,4% de la inercia de las tablas situadas en la diagonal. Si sólo estamos interesados en las tablas situadas fuera de la diagonal, deberíamos considerar que la inercia explicada por el mapa bidimensional del ACM, es del 83,2%, y no del 65,0%. Como veremos más adelante existe la posibilidad de mejorar este valor del 83,2%.

Omisión de las tablas de la diagonal: AC conjunto

Queda claro que, en la matriz de Burt, la inclusión de las tablas de la diagonal degrada los resultados globales del ACM, ya que intentamos visualizar de forma innecesaria las tablas de la diagonal que, además, presentan inercias muy elevadas. De hecho, las tablas de la diagonal tienen los máximos valores de inercia que se pueden alcanzar. Ignorando las tablas de la diagonal podríamos mejorar el cálculo de la inercia explicada por el mapa. El *análisis de correspondencias conjunto* (ACCo) es un algoritmo iterativo que lleva a cabo el AC de la matriz de Burt, buscando el ajuste sólo para las tablas situadas fuera de la diagonal. En este procedimiento, partimos de los resultados del ACM para, a continuación, sustituir las tablas de la diagonal por valores estimados a partir de estos resultados mediante la fórmula de reconstitución (13.4). Dado que la matriz de Burt es simétrica, las coordenadas y las masas de filas y de columnas son iguales. La fórmula toma la siguiente expresión para la solución bidimensional:

$$\hat{p}_{jj'} = c_j c_{j'} \left(1 + \sqrt{\lambda_1}\gamma_{j1}\gamma_{j'1} + \sqrt{\lambda_2}\gamma_{j2}\gamma_{j'2}\right) \qquad (19.1)$$

donde $\hat{p}_{jj'}$ es el valor estimado de la frecuencia relativa de la (j, j')-ésima celda de la matriz de Burt. Utilizando esta fórmula, podemos sustituir las tablas de la dia-

Preguntas	*Pregunta 1*	*Pregunta 2*	*Pregunta 3*	*Pregunta 4*	*Porcentaje pregunta*
Pregunta 1	—	97,8	95,8	77,8	*88,2*
Pregunta 2	97,8	—	87,4	97,0	*91,8*
Pregunta 3	95,8	87,4	—	96,7	*91,9*
Pregunta 4	77,8	97,0	96,7	—	*88,5*

Imagen 19.2:
Porcentaje de inercia de cada una de las 12 tablas fuera de la diagonal de la matriz de Burt explicado por el mapa bidimensional del ACCo

gonal de la matriz de Burt por sus valores estimados, obteniendo así una *matriz de Burt modificada*. A continuación, llevamos a cabo un nuevo AC de la matriz de Burt modificada para obtener una nueva solución. Una vez obtenidos los resultados, volvemos a reemplazar las tablas de la diagonal por sus estimaciones de (19.1) y obtenemos una nueva matriz de Burt modificada. Y así sucesivamente hasta llegar a la convergencia; en cada iteración mejoramos el ajuste de las tablas situadas fuera de la diagonal.

Resultados del ACCo

La aplicación del ACCo a los datos correspondientes a las cuatro variables sobre el trabajo de las mujeres nos lleva a los siguientes resultados: 90,2% de inercia explicada, y los porcentajes de inercia explicada de cada tabla que mostramos en la tabla de la imagen 19.2. Estos resultados son claramente mejores que los anteriores. Todas las tablas quedan muy bien representadas, la peor es la correspondiente al cruce de las preguntas 1 y 4, para la que la inercia explicada es del 77,8%. La imagen 19.3 muestra el mapa del ACCo, en el que la escala se ha mantenido de forma intencionada idéntica a la de la tabla 18.5 para poderla comparar. El resultado es prácticamente idéntico, la única diferencia observada es la disminución de la escala. En la imagen 18.5, las inercias de los dos ejes principales eran de 0,481 y 0,263, respectivamente, mientras que ahora son de 0,353 y de 0,128. Por tanto, una vez más se ha producido una disminución de las inercias principales, especialmente del segundo eje. Ello también ocurrió cuando en el ACM pasamos de la matriz binaria (imagen 18.2) a la matriz de Burt (imagen 18.5). Sin embargo, en aquella ocasión las coordenadas estándares de los dos análisis eran idénticas; aquí el resultado del ACCo es distinto del resultado del ACM, como podemos ver examinando en detalle el mapa de la imagen 19.3 y comparándolo con los mapas del ACM del capítulo 18.

Los resultados del ACCo no se hallan anidados

En el ACCo, como ocurre en el ACM, los ejes principales no se hallan anidados (es decir, el resultado de dos dimensiones no contiene exactamente la mejor solución unidimensional como eje principal). No obstante, en la práctica se produce un anidamiento aproximado. Este es el motivo por el cual en el mapa de la imagen 19.3 no damos los porcentajes de inercia de los ejes —sólo podemos dar el porcentaje de inercia del resultado global, en este caso el 90,2%—. Este hecho también impide dar las contribuciones de los ejes a la inercia de un punto. A pesar de que cada punto tiene una cierta calidad de re-

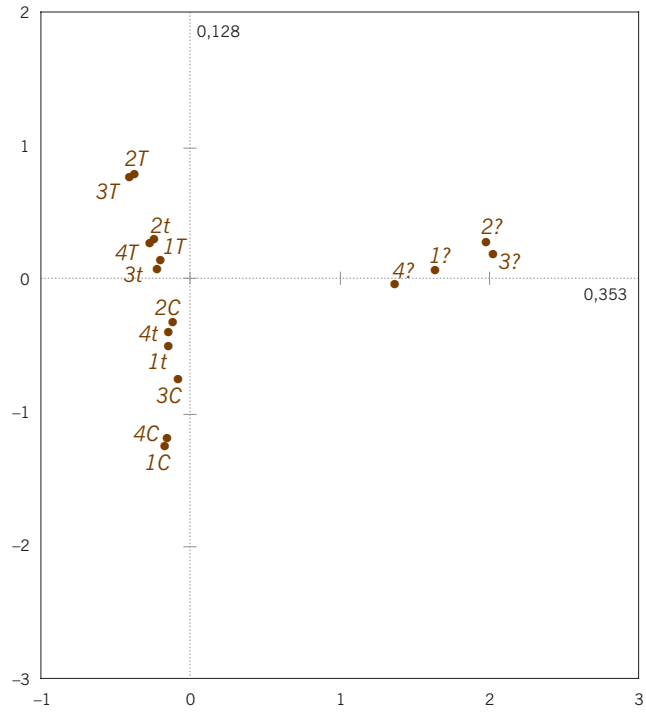


Imagen 19.3:
Mapa del ACCo de la matriz de Burt correspondiente a las cuatro preguntas sobre el trabajo de las mujeres; porcentajes de inercia del mapa: 90,2%. El porcentaje de inercia es la suma de las inercias explicadas de cada tabla (obtenidos de las imágenes 19.2 y 18.6) y expresados como un porcentaje de la suma de los valores de las inercias de las tablas situadas fuera de la diagonal (consúltese el apéndice teórico, A)

presentación en el mapa, no la podemos dividir en las partes correspondientes a cada eje.

Ajuste de los resultados del ACM para ajustar las tablas de fuera de la diagonal

Nuestra experiencia nos ha enseñado que casi siempre existe una gran semejanza entre los resultados del ACCo y del ACM. Este hecho sugiere que, en realidad, lo que distingue los resultados del ACCo de los del ACM es sólo un cambio de escala. Por tanto, podríamos investigar este cambio de escala como alternativa para mejorar el ACM. Dadas las coordenadas estándares del resultado del ACM, ¿cómo podríamos cambiar de escala (es decir, definir las coordenadas principales) para reconstruir de forma óptima los datos de las tablas de la matriz de Burt que se hallan fuera de la diagonal? Se trata de un problema de regresión, utilizando, una vez más, la fórmula de reconstitución (19.1), pero considerando como factores de escala (es decir, las raíces cuadradas de las inercias principales) los coeficientes desconocidos del modelo de regresión β_1 y β_2 (para una solución bidimensional) obtenemos:

$$\frac{p_{jj'}}{c_j c_{j'}} - 1 = \beta_1 \gamma_{j1} \gamma_{j'1} + \beta_2 \gamma_{j2} \gamma_{j'2} + e_{jj'} \qquad (19.2)$$

Llevamos a cabo la regresión de manera que obtengamos los valores de la «variable respuesta», disponiendo en forma de vector todos los valores situados a la

izquierda de (19.2), sólo para las celdas de las tablas situadas fuera de la diagonal —en nuestro ejemplo de cuatro variables, con seis tablas de 4 × 4 situadas fuera de la diagonal, tendremos, en el vector, 6 × 16 = 96 valores—. Como «variable explicativa» tenemos los correspondientes productos de $\gamma_{j1}\gamma_{j'1}$ y $\gamma_{j2}\gamma_{j'2}$. Llevamos a cabo una regresión de mínimos cuadrados ponderada, sin ordenada en el origen, con pesos iguales de valores respectivos $c_j c_{j'}$ —en nuestro ejemplo, obtenemos unos coeficientes estimados $\hat{\beta}_1 = 0{,}5922$ y $\hat{\beta}_2 = 0{,}3532$—. Los cuadrados de estos valores proporcionan, como «inercias principales», los valores óptimos 0,351 y 0,125, respectivamente, para los que la inercia explicada es del 89,9% (el coeficiente de determinación R^2 de la regresión). Es lo mejor que podemos hacer con los resultados del ACM; fijémonos en la similitud entre estos valores y los de las inercias principales del mapa de ACCo de la imagen 19.3, cuyos valores eran 0,353 y 0,128. Para representar las categorías en el mapa, calculamos las coordenadas principales multiplicando las coordenadas estándares del ACM de los dos primeros ejes por los factores de escala $\hat{\beta}_1$ y $\hat{\beta}_2$. De nuevo, la solución obtenida no es anidada, los resultados dependen de la dimensionalidad de la solución; si llevamos a cabo el mismo cálculo para una solución tridimensional, los dos primeros coeficientes de regresión no serán exactamente los que acabamos de obtener. El anidamiento se mantendrá sólo cuando las «variables explicativas» de (19.2) no estén correlacionadas. Ignorando las correlaciones, podríamos obtener de manera más simple un ajuste anidado (pero subóptimo), como describimos a continuación.

Ajuste simple del ACM

Vamos a describir un ajuste más simple de las inercias principales que cumple la condición de anidamiento, fácil de calcular y que implica los siguientes pasos: 1) recalcular la inercia total sólo para las tablas que se hallan fuera de la diagonal, y 2) un ajuste simple de las inercias principales derivado del ACM. Según nuestra experiencia, en general con este ajuste simplificado obtenemos resultados muy similares a los del ajuste óptimo que nos permiten expresar las inercias principales de la forma habitual, como porcentajes de inercia.

Inercia ajustada = inercia media de las tablas situadas fuera de la diagonal

En el ACM de la matriz de Burt **B**, la inercia total es la media de las inercias de todas las tablas que la componen, incluyendo las de la diagonal. Sin embargo con el ACCo, la inercia total es la media de las inercias de las tablas situadas fuera de la diagonal. Lo podemos calcular fácilmente a partir de la inercia total de **B**, ya que conocemos de forma exacta cuáles son los valores de las inercias de las tablas de la diagonal: $J_q - 1$, donde J_q es el número de categorías de la variable q-ésima. Por tanto,

$$\text{suma de las inercias de las } Q \text{ tablas de la diagonal} = J - Q \qquad (19.3)$$

mientras que

$$\text{suma de las inercias de todas las tablas de doble entrada} = Q^2 \times \text{inercia}(\mathbf{B}) \qquad (19.4)$$

Restando (19.3) de (19.4) obtenemos la suma de las inercias de las tablas situadas fuera de la diagonal, y luego dividiendo por $Q(Q-1)$ obtenemos la media, lo que lleva a:

$$\text{media de la inercia fuera de la diagonal} = \frac{Q}{Q-1} \times \left(\text{inercia}(\mathbf{B}) - \frac{J-Q}{Q^2} \right) \qquad (19.5)$$

Utilizando como ejemplo nuestros datos sobre el trabajo de las mujeres:

$$\text{inercia media fuera de la diagonal} = \frac{4}{3} \times \left(1{,}1452 - \frac{12}{16} \right) = 0{,}5269$$

Otra manera de calcular este valor es promediar directamente la media de las inercias de todas las tablas proporcionadas por la tabla de la imagen 18.6. Lo calculamos sólo en uno de los triángulos de la tabla, ya que solamente hay $\frac{1}{2}Q(Q-1) = 6$ pares de tablas distintas:

$$\frac{1}{6}(0{,}3657 + 0{,}4262 + 0{,}6457 + 0{,}8942 + 0{,}3477 + 0{,}4823) = 0{,}5269$$

Ajuste de cada inercia principal

Supongamos que en el ACM de la matriz de Burt **B** indicamos como λ_k, para $k = 1, 2$, etc., las inercias principales (valores propios). Calcularemos las inercias principales ajustadas λ_k^{adj} de la siguiente manera:

$$\lambda_k^{\text{adj}} = \left(\frac{Q}{Q-1} \right)^2 \times \left(\sqrt{\lambda_k} - \frac{1}{Q} \right)^2, \quad k = 1,2,\ldots \qquad (19.6)$$

En nuestro ejemplo, las dos primeras inercias principales ajustadas son:

$$\frac{16}{9} \times \left(0{,}6934 - \frac{1}{4} \right)^2 = 0{,}3495 \quad \text{y} \quad \frac{16}{9} \times \left(0{,}5132 - \frac{1}{4} \right)^2 = 0{,}1232$$

(fijémonos en la similitud de los valores óptimos 0,351 y 0,125, que vimos anteriormente).

Porcentajes de inercia del ajuste simple del ACM

Para obtener los porcentajes de inercia de cada eje principal, expresamos las inercias ajustadas con relación a la inercia total ajustada:

$$100 \times \frac{0{,}3495}{0{,}5269} = 66{,}3\% \quad \text{y} \quad 100 \times \frac{0{,}1232}{0{,}5269} = 23{,}4\%$$

En consecuencia, el porcentaje de inercia de la solución bidimensional del ajuste simple del ACM es del 89,7%, sólo el 0,2% menor que la del ajuste óptimo (no anidado) y el 0,5% menor que la del ACCo. Con los ejemplos anteriores, hemos visto,

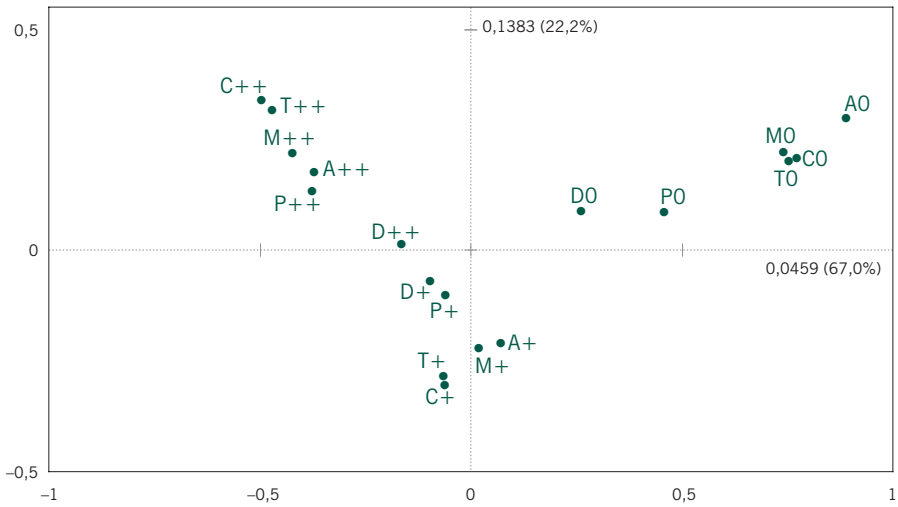


Imagen 19.4:
Mapa del ACM ajustado correspondiente a los datos sobre los intereses de los europeos. Porcentaje de inercia del mapa: 89,2% (si se hubiera llevado a cabo el ACM con la matriz binaria, la inercia explicada sería sólo el 41,1%)

pues, que los porcentajes calculados a partir de este ajuste simple proporcionan un porcentaje de inercia global que es un límite inferior del porcentaje óptimo obtenido con el ACCo. Por tanto, cuando demos el resultado de un ACM, lo mejor es expresar el ajuste de la manera que acabamos de ver. Y en consecuencia, para obtener las coordenadas principales a partir de las coordenadas estándares utilizaremos como factores de escala las raíces cuadradas de las inercias principales ajustadas de forma simple. No vamos a representar otra vez el mapa, ya que las posiciones relativas de los puntos son las mismas que vimos en los mapas de las imágenes 18.2 y 18.5; solamente la escala es distinta, más parecida a la del mapa de la imagen 18.2.

Conjunto de datos 10: interés por las noticias en Europa

Como ejemplo adicional y también para ilustrar otros aspectos del ACCo, consideremos un conjunto de datos derivado de la encuesta del Eurobarómetro de 2005. Entre otros aspectos, se preguntó a los encuestados sobre su interés por las noticias sobre: deportes (D), política (P), descubrimientos médicos (M), contaminación ambiental (A), innovación tecnológica (T) y descubrimientos científicos (C). Las posibilidades de respuesta eran: «muy interesados» (++), «moderadamente interesados» (+) y «nada interesados» (0). Siguiendo esta notación, vamos a simbolizar las respuestas por: A+, por ejemplo para «moderadamente interesados en la contaminación atmosférica» y P0 para «nada interesados en política». Con el objetivo de evitar el efecto de las no respuestas, que, como vimos en el ejemplo anterior, tiene un fuerte efecto sobre los resultados, hemos omitido los encuestados con respuestas tipo «no sabe» o con respuestas «no contesta». Ello ha significado una reducción de la muestra de 33.190 a 29.652, es decir, una reducción del 10,7% (en el capítulo 21 trataremos específicamente sobre las no respuestas). En la imagen 19.4, mostramos el mapa de ACM ajustado correspondiente a estos datos. El mapa muestra que los «nada interesados» forman su propia diagonal de dispersión a la derecha

Imagen 19.5:
Países europeos representados como puntos adicionales en el mapa del ACM ajustado correspondiente a datos sobre los intereses de los europeos. (Se muestran los nombres de los países como aparecen en el Eurobarómetro.)

de los «interesados», que asimismo van de «moderadamente interesados» abajo, a «muy interesados» arriba a la izquierda. Estamos ante un ejemplo de mapa que podría mejorar con una rotación de ejes, si queremos que estas dos líneas de dispersión coincidan más con los ejes principales. (En el epílogo hacemos algunos comentarios sobre las rotaciones.) El primer eje explica el 67,0% de la inercia y define una escala de interés general sobre las noticias. El segundo eje (22,0%) muestra el interés por los descubrimientos científicos y la innovación tecnológica en los extremos, lo que indica una elevada correlación entre ambos. Sin embargo, los dos puntos correspondientes al interés por los deportes se hallan cerca del centro de dispersión, lo que indica que, por ejemplo, un elevado interés por los deportes (D++) va asociado con un interés moderado en los otros temas, y viceversa. Recordemos que lo que estamos visualizando son las asociaciones de las categorías de una determinada variable con las categorías de las restantes variables.

Puntos adicionales en ACCo y en ACM ajustado

A pesar de que no mostramos las posiciones de los 29.652 casos de los datos, los podemos imaginar como puntos adicionales. Es decir, si añadiéramos como filas adicionales la gran matriz binaria de 29.652 × 18 a la matriz de Burt, cada encuestado tendría una posición en el mapa del ACM (sin embargo, fijémonos en que sólo existen $3^6 = 729$ respuestas distintas, por tanto, las respuestas de los encuestados situarán en los puntos que representan cada tipo de respuesta). Dado que en las tres ver-

siones distintas del ACM (binario, Burt y ajustado), las coordenadas estándares son iguales, las posiciones de las coordenadas principales de los encuestados serán las mismas en los tres. Como vimos en el capítulo 18, podemos mostrar las categorías adicionales añadiendo las tablas correspondientes a sus cruzamientos por las variables activas como filas adicionales de la matriz de Burt. Por ejemplo, con los datos que nos ocupan, tenemos muestras de 34 países europeos. Los encuestados de los distintos países tienen una posición en el mapa, la de cada país se halla en la posición media de los puntos correspondientes a los encuestados de ese país. El mapa de la imagen 19.5 muestra las posiciones de los países etiquetadas con los nombres locales (podemos imaginar este mapa superpuesto al mapa de la imagen 19.4). TURKIYE (Turquía) es, de todos los países, el que se halla más en la posición de «nada de interés»: un 40% de los encuestados turcos no mostró interés alguno en ningún tema, excepto la contaminación ambiental (22%); sin embargo KYPROS (Chipre), ELLADA (Grecia) y MALTA parecen ser los países más interesados; así, por ejemplo, Chipre tenía los mayores porcentajes en «muy interesado» en temas como contaminación ambiental (75%), descubrimientos médicos (62%), innovación tecnológica (53%) y descubrimientos científicos (55%).

RESUMEN: Análisis de correspondencias conjunto

1. Podemos definir el ACM como el AC de la matriz de Burt, formada por todas las tablas de contingencia de dos entradas derivadas del cruzamiento de un conjunto de variables, incluyendo las de una variable con ella misma, lo que conlleva una sobreestimación de la inercia total.
2. El *análisis de correspondencias conjunto* (ACCo) busca el mapa que mejor explica las tablas de contingencia correspondientes a los cruces de todos los pares de variables, ignorando las tablas que se hallan en la diagonal de la matriz de Burt. Esta aproximación implica un nuevo algoritmo iterativo que conduce a una solución óptima no anidada.
3. En el ACCo, la inercia total corresponde a la media de las inercias de todas las tablas que se hallan fuera de la diagonal de la matriz de Burt.
4. Una solución intermedia consiste en hallar las coordenadas estándares del ACM, mediante regresión de mínimos cuadrados ponderada de las tablas de contingencia de interés. Sin embargo, de nuevo, esta solución no es anidada.
5. El ACM *ajustado* es una solución simple, anidada, y que, por tanto, mantiene todas las buenas propiedades del AC, al mismo tiempo que resuelve el problema de la baja inercia. Consiste en aplicar algunos ajustes a las inercias principales y a la inercia total del ACM.
6. En todos los tipos de ACM, representamos las variables adicionales de la misma manera. Concretamente, cruzamos las variables adicionales con las variables activas, y añadimos las tablas de contingencia resultantes como filas adicionales a la matriz de Burt (o en la matriz de Burt modificada en ACCo).

CAPÍTULO 20

Propiedades del escalado óptimo del ACM

En los capítulos 7 y 8 vimos que existen distintas definiciones de AC, así como diversas maneras de abordar este método. En este libro hemos enfatizado la aproximación geométrica de Benzécri que lleva a la visualización de datos. En los capítulos 18 y 19 quedó claro que el paso del AC simple, de dos variables, a las formas multivariadas no es directo, especialmente si tratamos de generalizar la interpretación geométrica. Una aproximación alternativa al caso multivariado, con exactamente el mismo aparato matemático que el ACM, consiste en acercarse a este método como una manera de cuantificar datos categóricos, generalizando de esta manera las ideas sobre escalado óptimo que vimos en el capítulo 7. También aquí veremos que existen distintas formas de presentar el ACM como una técnica de escalado. El estudio de estas aproximaciones alternativas enriquecerán nuestra comprensión sobre las propiedades de este método. En la literatura, la aproximación al ACM como un método de escalado óptimo se llama *análisis de homogeneidad*.

Contenido

Conjunto de datos 11: actitudes hacia la ciencia y el medio ambiente

Este conjunto de datos lo hemos obtenido de la encuesta multinacional del ISSP de 1993 sobre el medio ambiente. Nos centraremos en $Q = 4$ preguntas sobre la opinión de la gente acerca del papel de la ciencia. Se preguntó a los encuestados si estaban o no de acuerdo con las afirmaciones siguientes:

A Creemos demasiado en la ciencia, y demasiado poco en los sentimientos y en la fe.

B En general, la ciencia moderna provoca más daños que beneficios.

C Cualquier cambio que el hombre cause en la naturaleza, a pesar de que tenga un base científica, es probable que empeore las cosas.

D La ciencia moderna solucionará los problemas medioambientales sin modificar sustancialmente nuestro modo de vida.

Tenemos cinco posibles respuestas categóricas:

1 Muy de acuerdo.

2 Algo de acuerdo.

3 Ni de acuerdo ni en desacuerdo.

4 Algo en desacuerdo.

5 Muy en desacuerdo.

Para simplificar solamente hemos utilizado datos de Alemania Occidental. Hemos omitido los casos con valores perdidos en cualquiera de las cuatro preguntas, lo que nos ha llevado a una muestra de $N = 871$. (Estos datos se hallan incluidos en nuestro paquete **ca** para R, que se ofrece en el apéndice B.)

La cuantificación de categorías como objetivo

En el capítulo 7 definíamos el AC como un método de cuantificación de las categorías de la variable columna que nos lleva a la mayor diferenciación, o discriminación, posible entre las categorías de la variable fila, o viceversa. Es lo que llamaríamos definición «asimétrica», ya que las filas y las columnas desempeñan papeles distintos en la definición, lo que también se refleja en los resultados. Así, expresamos los resultados de las columnas en coordenadas estándares, mientras que los de las filas los expresamos en coordenadas principales. En el capítulo 8 definimos el AC de forma «simétrica» como un método de cuantificación de las categorías que nos lleva a la mayor correlación entre filas y columnas. En esta definición, el papel de filas y columnas es el mismo. Esta cuantificación de las categorías no incluye ningún concepto geométrico específico; en concreto, no hace mención alguna a un espacio en el que podamos imaginar situados los datos, lo que, por el contrario, es muy importante en la aproximación geométrica para poder medir la inercia total y los porcentajes de inercia en los subespacios de baja dimensionalidad.

El ACM como el análisis de componentes principales de la matiz binaria

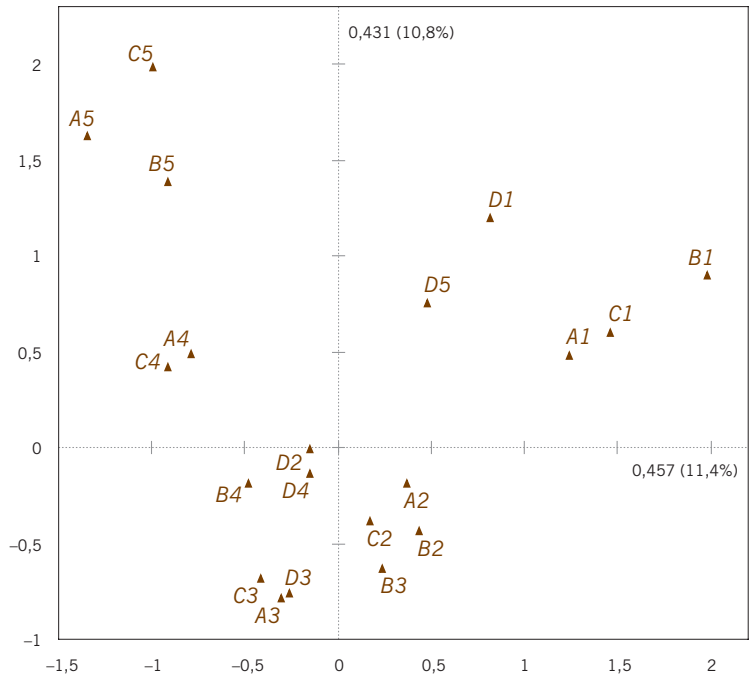
La metodología de cuantificación de categorías utilizada en el AC asimétrico sobre la matriz binaria se parece mucho al análisis de componentes principales (ACP). En general, aplicamos el ACP a datos derivados de una escala continua, no obstante, el ACP está muy relacionado con la teoría y el cálculo del AC (en realidad, podríamos decir que el AC es una variante del ACP aplicado a datos categóricos). En el ACP,

dado un conjunto de datos, en el que las filas son los casos y las columnas las variables (m variables, $x_1,..., x_m$), asignamos a las columnas unos coeficientes $\alpha_1,...,\alpha_m$ (que tendremos que estimar) que conducen a combinaciones lineales para las filas (casos) de la forma $\alpha_1 x_1 + ... + \alpha_m x_m$, las *puntuaciones*. Calculamos los coeficientes de manera que se maximice la varianza de las puntuaciones de las filas. Como vimos anteriormente, para hallar la solución tenemos que definir unas condiciones de identificación. En el ACP, estas condiciones son, en general, que la suma de los cuadrados de los coeficientes sea 1: $\sum_j \alpha_j^2 = 1$. Aplicar estas ideas a la matriz binaria, que sólo consta de ceros y unos, y asignar coeficientes $\alpha_1,..., \alpha_J$ a las variables binarias, para calcular luego las combinaciones lineales de las filas, simplemente significa sumar los coeficientes α (es decir los valores de la escala) de cada caso. Por tanto, la maximización de la varianza de los casos recuerda el procedimiento de escalado óptimo que vimos en el capítulo 7 (maximización de la discriminación entre filas). De hecho se trata de un concepto casi idéntico, con la excepción de las condiciones de identificación. En el escalado óptimo, las condiciones de identificación serían que la varianza ponderada (inercia) de los coeficientes (no la simple suma de cuadrados) fuera 1: $\sum_j c_j \alpha_j^2 = 1$. Aquí las c_j son las masas de las columnas, es decir, la suma de las columnas de la matriz binaria divididas por la suma total NQ de la matriz binaria; así, para cada variable categórica, la suma de los c_j es $1/Q$. Por tanto, con este cambio en las condiciones de identificación, podríamos llamar al ACM, ACP de datos categóricos, que maximiza la varianza de los casos. Los coeficientes son las coordenadas estándares de las categorías de las columnas, mientras que las coordenadas principales del ACM de los casos son las medias de los valores de éstos. Es decir $1/Q$ veces la suma de lo que hemos llamado antes «puntuaciones». La primera dimensión del ACM maximiza la varianza (primera inercia principal), la segunda dimensión maximiza la varianza con la condición de que las puntuaciones no estén correlacionadas con las de la primera dimensión y así sucesivamente.

Maximización de la correlación entre ítems

El *análisis de homogeneidad*, visto como una técnica de escalado óptimo del ACM, se contempla, habitualmente, como una generalización de la correlación según se expuso en el capítulo 8. En concreto, vimos la ecuación (8.1) como una manera alternativa de optimizar la correlación entre dos variables categóricas, que podemos fácilmente generalizar a más de dos variables. Para ilustrar este hecho utilizaremos una notación correspondiente a cuatro variables, sin embargo, podemos extenderlo fácilmente a Q variables con cualquier número de categorías (en nuestro ejemplo $Q = 4$, y el número total de categorías es $J = 20$). Supongamos que las cuatro variables toman los valores (desconocidos) de a_1 a a_5, de b_1 a b_5, de c_1 a c_5 y de d_1 a d_5. Asignaremos a los encuestados cuatro de estos valores a_i, b_j, c_k y d_l de acuerdo con sus respuestas, y de esta manera cuantificaremos las respuestas de toda la muestra, que simbolizamos como a, b, c y d (es decir, a indica todas las 871 respuesta cuantificadas a la pregunta A, etc.). Cada encuestado tendrá como puntuación la suma estos valores, $a_i + b_j + c_k + d_l$. Simbolizaremos las puntuaciones

Imagen 20.1:
Mapa del ACM (versión matriz binomial) sobre la actitud hacia la ciencia, que muestra los puntos correspondientes a las categorías en coordenadas principales. Dado que las inercias principales difieren sólo ligeramente (e incluso menos en forma de raíces cuadradas), en ambos ejes, las coordenadas principales presentan casi la misma contracción que las coordenadas estándares



de toda la muestra como $a + b + c + d$. En este contexto, llamamos *ítems* a las variables, *puntuaciones de los ítems* a los valores de a a d, y *suma de puntuaciones* a la suma $a + b + c + d$. Expresaremos, el criterio de búsqueda de los valores óptimos de la escala, como la maximización de la media de las correlaciones al cuadrado entre las puntuaciones de los ítems y la suma de puntuaciones:

$$\text{correlaciones al cuadrado} = \frac{1}{4}[\text{cor}^2(a, a+b+c+d) + \text{cor}^2(b, a+b+c+d) + \text{cor}^2(c, a+b+c+d) + \text{cor}^2(d, a+b+c+d)] \quad (20.1)$$

(Comparemos con el caso de dos variables en (8.1).) De nuevo necesitamos las condiciones de identificación. Es conveniente aplicar a la suma de puntuaciones la condición de media 0 y varianza 1: media $(a + b + c + d) = 0$, $\text{var}(a + b + c + d) = 1$. Obtenemos, de forma exacta, la solución a este problema de maximización, con las coordenadas estándares de las categorías de los ítems en el primer eje principal del ACM, la media de las correlaciones al cuadrado (20.1) maximizada es exactamente la primera inercia principal (del ACM de la matriz binaria).

ACM del ejemplo de la opinión sobre la ciencia

En el mapa bidimensional de la matriz binomial de la imagen 20.1 vemos, de nuevo, porcentajes de inercia muy bajos (los porcentajes basados en las inercias ajustadas son el 44,9% y el 34,2%, respectivamente). Sin embargo, en este caso, dado

Categorías	Preguntas				
	A	*B*	*C*	*D*	*Suma*
1 «Muy de acuerdo»	115	174	203	25	*518*
2 «Algo de acuerdo»	28	21	6	3	*57*
3 «Ni de acuerdo ni en desacuerdo»	12	7	22	9	*49*
4 «Algo en desacuerdo»	69	41	80	3	*194*
5 «Muy en desacuerdo»	55	74	32	22	*182*
Suma	*279*	*317*	*343*	*62*	*1000*

Imagen 20.2:
Contribuciones a la inercia en tantos por mil (‰) del primer eje principal (versión matriz binaria) de los datos sobre ciencia y medio ambiente

que los valores de las inercias principales son medias de correlaciones al cuadrado, debemos ignorar los porcentajes, ya que los valores de las inercias principales tienen interés *per se*. El valor máximo de (20.1) es 0,457. La segunda inercia principal, 0,431, se halla buscando un nuevo conjunto de valores que nos lleven a unas puntuaciones que no estén correlacionadas con los que se obtuvieron anteriormente, y que además maximicen (20.1); este valor máximo es 0,431. Y continuaríamos de esta manera para hallar los resultados de los restantes ejes, siempre no correlacionados con los hallados anteriormente. En el mapa de la imagen 20.1, vemos que las preguntas *A*, *B* y *C* presentan una distribución muy similar, como una cuña en forma de herradura, que va de profundos desacuerdos, a la izquierda, a fuertes acuerdos, a la derecha. Sin embargo, la pregunta *D* sigue una trayectoria completamente distinta con los dos valores extremos muy próximos. Las primeras tres preguntas presentaban un redactado negativo hacia la ciencia, mientras que la pregunta *D* tenía un redactado mucho más positivo; por tanto, habríamos esperado que *D5* se hallara hacia *A1*, *B1* y *C1*, y *D1* se hallara al lado de *A5*, *B5* y *C5*. Sin embargo, el hecho de que *D1* y *D5* se hallen tan cerca y dentro de la herradura indica que ambas están asociadas con respuestas extremas de las restantes tres preguntas: la explicación más plausible es que algunos encuestados hayan interpretado mal el cambio de sentido del redactado de la cuarta pregunta.

Correlaciones individuales al cuadrado

También es interesante conocer los valores de cada una de las correlaciones al cuadrado que componen (20.1). Podemos obtener estos valores directamente sumando la contribución de cada pregunta a la inercia del primer eje principal. Habitualmente, los resultados del ACM proporcionan esta información en forma de proporciones o en tanto por mil. En la imagen 20.2 detallamos estos valores en esta última forma para ilustrar cómo recuperar estas correlaciones. Las preguntas de *A* a *D* contribuyen, en las proporciones 0,279, 0,317, 0,343 y 0,062 de la inercia principal de 0,457. Dado que 0,457 es la media de las cuatro correlaciones al cuadrado, las correlaciones al cuadrado y, en consecuencia, las correlaciones son:

$$A: 0{,}279 \times 0{,}457 \times 4 = 0{,}510 \quad \text{correlación} = \sqrt{0{,}510} = 0{,}714$$
$$B: 0{,}317 \times 0{,}457 \times 4 = 0{,}579 \quad \text{correlación} = \sqrt{0{,}579} = 0{,}761$$
$$C: 0{,}343 \times 0{,}457 \times 4 = 0{,}627 \quad \text{correlación} = \sqrt{0{,}627} = 0{,}792$$
$$D: 0{,}062 \times 0{,}457 \times 4 = 0{,}113 \quad \text{correlación} = \sqrt{0{,}113} = 0{,}337$$

Estos cálculos muestran el bajo valor de la correlación de la pregunta *D* con relación a la puntuación total. Fijémonos en que, a pesar de que el ACM de la matriz binomial era la peor, desde un punto de vista geométrico habitual de distancias χ^2, inercia total, etc., las inercias principales y las contribuciones a las inercias principales tienen una interpretación muy interesante por sí mismas. En el *análisis de homogeneidad*, que teóricamente es equivalente al ACM de la matriz binaria, pero que interpreta el método desde el punto de vista de cuantificación de las categorías, denominamos *valores de discriminación* a las correlaciones al cuadrado 0,510; 0,579; 0,627 y 0,113.

Pérdida de homogeneidad

El análisis de homogeneidad generaliza la función objetivo (8.3) a muchas variables. Utilizando la notación anterior para el ejemplo que nos ocupa con cuatro variables, calcularíamos la puntuación media $\frac{1}{4}(a_i + b_j + c_k + d_l)$ de las puntuaciones de los ítems de cada encuestado y luego calcularíamos la varianza del encuestado dentro de su grupo de respuestas cuantificadas:

$$\begin{aligned}\text{varianza (para un caso)} = \frac{1}{4}\Big(& [a_i - \frac{1}{4}(a_i + b_j + c_k + d_l)]^2 \\ & + [b_j - \frac{1}{4}(a_i + b_j + c_k + d_l)]^2 \\ & + [c_k - \frac{1}{4}(a_i + b_j + c_k + d_l)]^2 \\ & + [d_l - \frac{1}{4}(a_i + b_j + c_k + d_l)]^2 \Big)\end{aligned} \qquad (20.2)$$

A continuación, calcularíamos la *pérdida de homogeneidad* como la media de todos estos valores con relación a los *N* casos. El objetivo es minimizar esta pérdida. De nuevo el ACM (versión matriz binaria) resuelve este problema. La minimización de la pérdida es 1 menos la primera inercia principal, es decir, 1 – 0,457 = 0,543. Minimizar la pérdida es equivalente a maximizar la correlación que hemos definido anteriormente.

Geometría de la función de pérdida en el análisis de homogeneidad

El objetivo de minimizar la pérdida de homogeneidad tiene una interpretación geométrica muy atractiva, muy relacionada con la definición de distancia entre filas y columnas del AC que vimos en el capítulo 7. En realidad, el cálculo de la pérdida de homogeneidad es exactamente igual al cálculo de la distancia ponderada (7.6), aplicada a la matriz binaria. En la imagen 20.3 mostramos el mapa asimétrico de ACM de todos los *N* = 871 encuestados (en coordenadas principales) y las

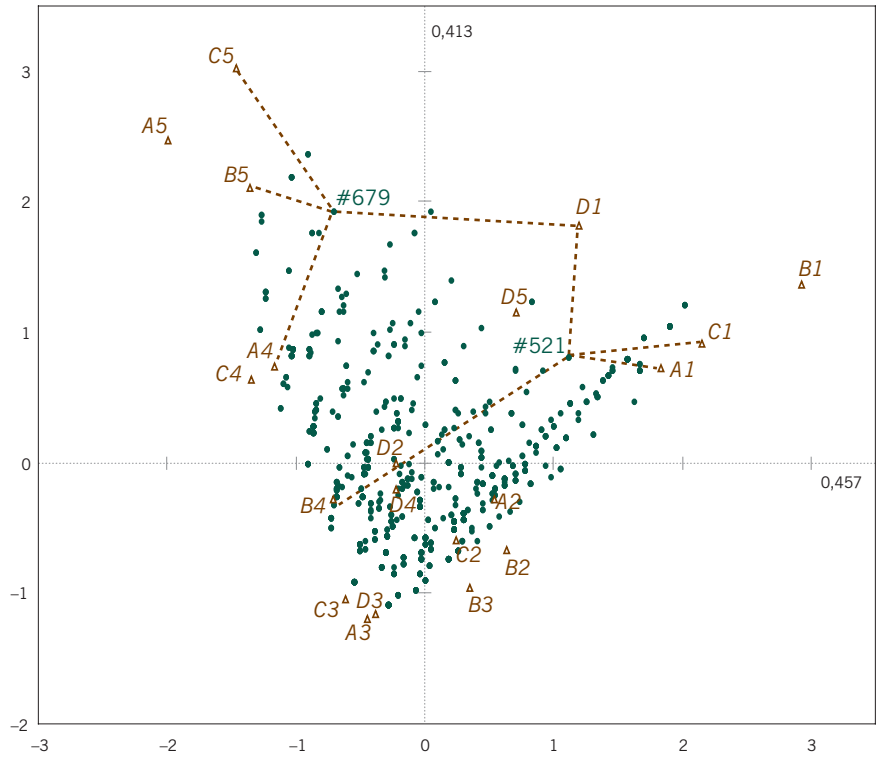


Imagen 20.3:
Mapa asimétrico (versión matriz binaria) de la opinión sobre la ciencia, que muestra los encuestados en coordenadas principales y las categorías en coordenadas estándares. Cada encuestado se halla en la media de sus cuatro respuestas. El ACM minimiza la suma de las distancias al cuadrado entre los puntos correspondientes a los individuos y sus respuestas

$J = 20$ categorías (en coordenadas estándares). Esto significa que los encuestados se hallan en los centroides de las categorías, siendo los pesos los valores relativos de las filas de la matriz binaria. Cada encuestado tiene un perfil que consta de ceros, y valores de $\frac{1}{4}$ en las posiciones correspondientes a las cuatro respuestas. Por tanto, el punto correspondiente a cada encuestado se halla en la posición de la media ordinaria de sus respuestas. En el mapa (imagen 20.3) hemos etiquetado los encuestados #679 y #521. Las respuestas del individuo #679 son: (*A4*, *B5*, *C5*, *D1*). Es decir, está en desacuerdo con las tres primeras preguntas y de acuerdo con la cuarta —en el mapa hemos unido mediante trazo discontinuo este individuo, situado a la izquierda, con las categorías correspondientes a sus respuestas—. Se trata de una fuerte y consistente opinión a favor de la ciencia. En contraste, las respuestas del individuo #521 son más variadas: (*A1*, *B4*, *C1*, *D1*). Este último individuo opina que creemos demasiado en la ciencia y que la interferencia de los humanos en la naturaleza empeorará las cosas. Sin embargo, también está muy de acuerdo en que la ciencia solucionará nuestros problemas medioambientales, al mismo tiempo que opina que la ciencia hace más daño que bien. Este tipo de respuestas explica el hecho de que *D1* se haya acercado hacia el centro del mapa, entre las dos opiniones extremas. Cada encuestado se halla en la media de sus

cuatro respuestas. Para cualquier configuración de respuestas categóricas, los encuestados se hallarán en la posición media. El mapa que mostramos en la imagen 20.3 es óptimo en el sentido de que las líneas que unen los encuestados con las categorías son las más cortas posibles (en términos de sumas de distancias al cuadrado). Llamamos *diagrama de estrellas* a los diagramas resultantes de unir los puntos correspondientes a los individuos con los de sus respuestas. Podríamos decir que el objetivo del ACM es la obtención de diagramas de estrellas con las menores distancias entre los individuos y sus respuestas en el sentido mínimo-cuadrático. El número de uniones entre los puntos correspondientes a los N encuestados y los correspondientes a sus Q respuestas es NQ. La pérdida de homogeneidad es la media de los cuadrados de las uniones (por ejemplo, en (20.2) donde $Q = 4$, dividimos la suma de los cuatro cuadrados por 4; para los N individuos dividimos la suma de cuadrados por $4N$). Por tanto, la media de la suma de las uniones al cuadrado en la primera dimensión es 1 – 0,457 = 0,513 y en la segunda dimensión es 1 – 0,413 = 0,587. Por el teorema de Pitágoras, la media de la suma de las uniones al cuadrado en el mapa bidimensional de la imagen 20.3 es 0,513 + 0,587 = 1,100.

Fiabilidad y alfa de Cronbach

En el ejemplo que nos ocupa con datos sobre la ciencia y el medio ambiente, vimos que la pregunta *D* no está muy correlacionada con las restantes (pág. 210). En este contexto, si hubiéramos querido obtener un indicador global de la opinión de la gente sobre la ciencia, hubiéramos dicho que estos resultados nos muestran que la pregunta *D* empeoraba la *fiabilidad* de la puntuación total, y que lo mejor habría sido eliminarla. En teoría de fiabilidad suponemos que las Q preguntas o ítems miden una estructura subyacente. La *alfa de Cronbach* es una medida estándar de fiabilidad definida como:

$$\alpha = \frac{Q}{Q-1}\left(1 - \frac{\sum_q s_q^2}{s^2}\right) \tag{20.3}$$

donde s_q^2 es la varianza de la puntuación del ítem q-ésimo, $q = 1,..., Q$ (por ejemplo, las varianzas de *a, b, c* y *d*) y s^2 es la varianza de la suma de las puntuaciones media (por ejemplo, la varianza de $(a + b + c + d)$). Aplicando esta definición a la primera dimensión del resultado del ACM, vemos que la alfa de Cronbach se reduce a:

$$\alpha = \frac{Q}{Q-1}\left(1 - \frac{1}{Q\lambda_1}\right) \tag{20.4}$$

donde λ_1 es la primera inercia principal de la matriz binaria. Por tanto, cuanto mayor sea la inercia principal, mayor será la fiabilidad. Utilizando $Q = 4$ y $\lambda_1 = 0,4574$ (cuatro dígitos significativos para aumentar un poco la exactitud) obtenemos:

$$\alpha = \frac{4}{3}\left(1 - \frac{1}{4 \times 0,4574}\right) = 0,605$$

Una vez visto el comportamiento de la pregunta D, una posibilidad es eliminarla y hacer de nuevo los cálculos con las tres preguntas que están altamente intercorrelacionadas. No mostramos aquí estos resultados, solamente señalaremos que la primera inercia principal correspondiente a estas tres variables es $\lambda_1 = 0{,}6018$, con un incremento de fiabilidad hasta $\alpha = 0{,}669$ (utilizando (20.4), con $Q = 3$). Como comentario final es interesante que nos demos cuenta de que la media de las correlaciones al cuadrado de un conjunto de variables aleatorias, cuyas correlaciones dos a dos no son iguales a cero, es igual a $1/Q$, lo que corresponde a una alfa de Cronbach igual a 0. El valor $1/Q$ es exactamente el umbral que hemos utilizado en (19.7) para ajustar las inercias principales (valores propios), y es también la inercia principal media del ACM de la matriz binaria que hemos mencionado en el capítulo 18.

RESUMEN: Propiedades del escalado óptimo del ACM

1. En el contexto de dos variables, definimos el escalado óptimo como la búsqueda de valores numéricos para las categorías de una variable que lleven a la separación máxima de los grupos definidos por la otra variable. Este problema es equivalente a hallar los valores numéricos, para las categorías, que conduzcan a la máxima correlación entre las variables fila y las variables columna.
2. En un contexto multivariado, el escalado óptimo consiste en la búsqueda de valores numéricos para las categorías de todas las variables que optimicen la correlación entre las variables y su suma (o promedio). En concreto, maximizamos la media de las correlaciones al cuadrado de los valores numéricos de cada variable, *puntuaciones de los ítems*, con su suma (o promedio), es decir, su *puntuación*.
3. De forma equivalente, el *análisis de homogeneidad* consiste en minimizar el promedio de las varianzas de las puntuaciones de los ítems de cada encuestado de la muestra.
4. En general, la aproximación al ACM como un escalado óptimo, ejemplificado por el análisis de homogeneidad, es el mejor marco para la interpretación de los resultados del ACM de una matriz binaria. Es mejor interpretar las inercias principales y su descomposición como correlaciones que de forma geométrica, tal como hacíamos en el AC simple.
5. La primera inercia principal del ACM, versión binaria, presenta una relación monotónica con la fiabilidad expresada con la alfa de Cronbach: cuanto mayor sea la inercia principal, mayor será la fiabilidad.
6. Dado que las coordenadas estándares de la matriz binaria, de la matriz de Burt y de la forma ajustada son idénticas, podemos aplicar las propiedades de escalado óptimo a las tres versiones del ACM.

CAPÍTULO 9

Análisis de correspondencias de subgrupos

A menudo conviene analizar sólo una parte de la matriz de datos, dejando fuera algunas filas y/o columnas. Así, por ejemplo, puede que nos convenga analizar de forma separada las columnas en grupos que formemos siguiendo algún criterio sustantivo. O puede ocurrir, por ejemplo, que sea conveniente excluir del análisis categorías con valores perdidos. En estas situaciones podríamos aplicar directamente el AC a la submatriz de interés; pero, al proceder de esta manera, puede ocurrir que los valores marginales de uno o ambos márgenes de la submatriz sean distintos de los de la matriz original y, en consecuencia, cambien los perfiles, las masas y las distancias. Sin embargo, en el *análisis de correspondencias de subgrupos*, el procedimiento de análisis que presentamos en este capítulo para la determinación de masas y distancias χ^2 de cualquier submatriz, utilizamos los valores marginales originales de la matriz completa. Este tipo de análisis tiene muchas ventajas; así, por ejemplo, nos permite descomponer la inercia total de la matriz original en las distintas submatrices, de manera que la información, contenida en la matriz de datos original queda recogida en las submatrices analizadas.

Contenido

Análisis de consonantes y vocales en los datos sobre autores

Los datos sobre autores en lengua inglesa de la tabla de la imagen 10.6 son un buen ejemplo de tabla que podemos dividir de forma natural (el alfabeto inglés está formado por 21 consonantes y 5 vocales). En el capítulo 10, vimos que la inercia total de esta tabla era muy baja, 0,01873 pero, sin embargo existía una clara estructura en las filas (los 12 textos de seis autores). Podría ser interesante ver el resultado del AC de vocales y de consonantes de forma separada. Una posibilidad sería analizar sin más las dos submatrices; la submatriz de 12 × 21 de frecuencias de consonantes y la submatriz de 12 × 5 de frecuencias de vocales. No obstante, proceder de esta manera implicaría recalcular los valores de los perfiles de cada texto en relación con los valores marginales de las nuevas submatrices. Por ejemplo, en el análisis de consonantes, calcularíamos las frecuencias relativas de *b*, *c*, *d*, *f*, ..., etc. en relación con el número total de consonantes del texto, y no en relación con el número total de letras. De esta manera, la masa de cada texto sería proporcional al recuento de consonantes, y no al número total de letras. Los perfiles de las consonantes se mantendrían invariables, sin embargo, las distancias χ^2 entre ellos serían distintas ya que éstas dependen de las masas de las filas que, como hemos comentado, han cambiado.

El análisis de subgrupos mantiene invariables los valores marginales originales

El análisis de correspondencias de subgrupos es una aproximación alternativa, con muchas ventajas. En dicho análisis para el cálculo de masas y de distancias utilizamos los valores marginales de la matriz original. Ello conlleva introducir una modificación muy simple en el algoritmo de cálculo del AC: todo lo que tenemos que hacer es suprimir los cálculos de las sumas marginales «locales» de la submatriz seleccionada, y mantener los valores de los cálculos realizados con la matriz original.

AC de subtablas: análisis de consonantes, biplot estándar

Aplicando el AC de subgrupos a la tabla de frecuencias de consonantes (págs. 315-316), obtenemos el mapa de la imagen 21.1. Aquí en vez de mostrar el mapa simétrico o el asimétrico, mostramos el biplot estándar del AC (cap. 13). Dado que los textos están en coordenadas principales, las distancias entre puntos son aproximadamente distancias χ^2. En el cálculo de las distancias χ^2 sólo tenemos en cuenta las consonantes; dejamos fuera las vocales. Expresamos las consonantes en coordenadas estándares multiplicadas por las raíces cuadradas de las correspondientes frecuencias relativas de las consonantes (es decir, las frecuencias relativas de las 26 letras; recordemos que las sumas marginales son siempre las de la tabla original). En cada eje, las raíces cuadradas de las longitudes de los vectores de consonantes son proporcionales a sus contribuciones al eje. Razón por la cual la letra *y* es tan importante en el segundo eje (más del 50%). Los biplots estándares funcionan igual de bien para tablas con inercias bajas o altas, y en este ejemplo, con una inercia extremadamente baja, es particularmente útil. Comparando este mapa con el mapa asimétrico de la imagen 10.7, vemos que las letras apuntan más o menos hacia las mismas direcciones. También vemos que la configuración de los textos es bastante similar. La inercia total del mapa es de 0,01637, valor exactamente igual a la suma de las inercias de las consonantes del análisis completo. En la página 111

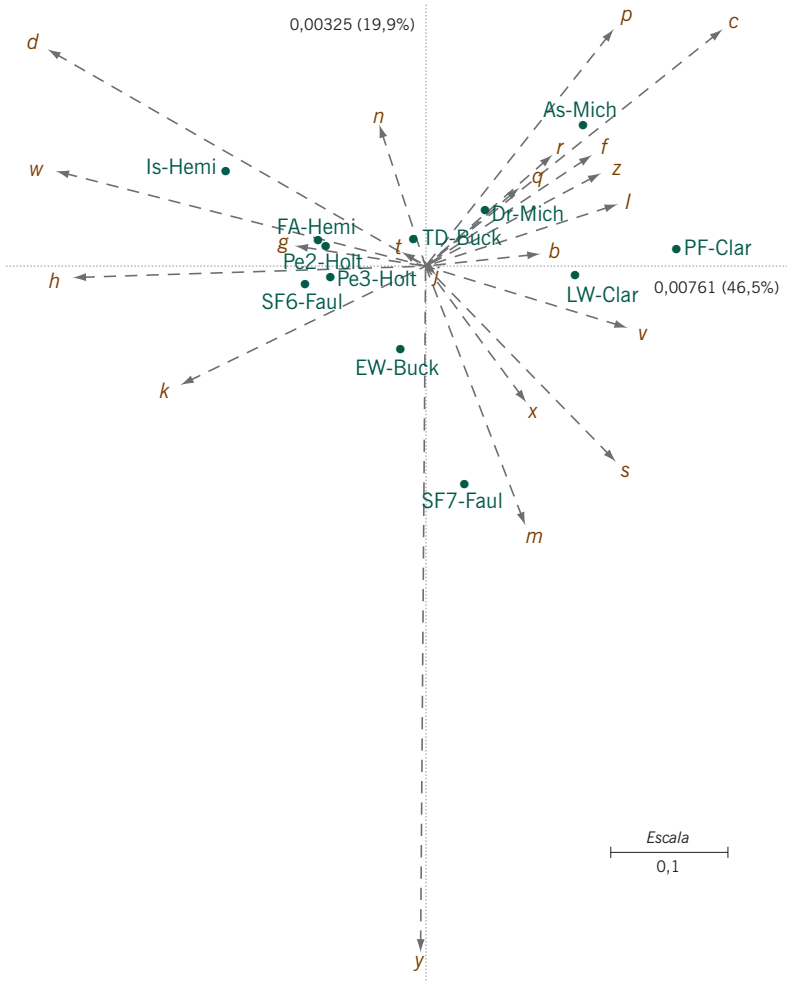


Imagen 21.1:
AC del subgrupo de consonantes del ejemplo de los autores; biplot estándar de filas, es decir, filas (textos) en coordenadas principales y columnas (letras) en coordenadas estándares multiplicadas por la raíces cuadradas de las masas de las columnas

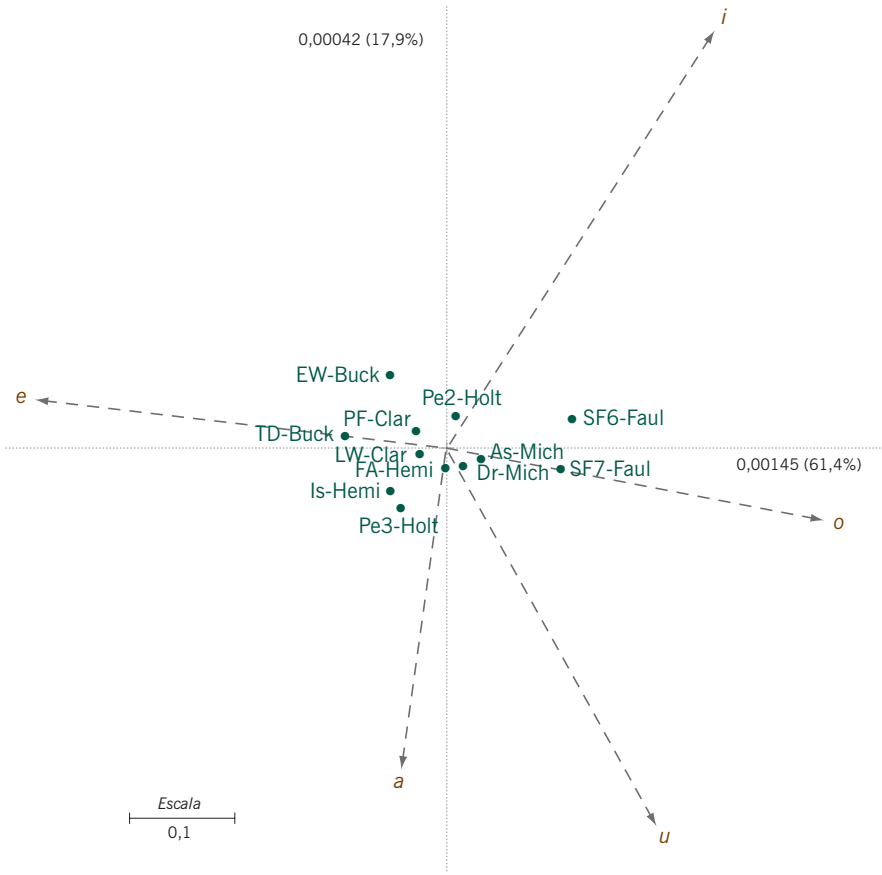
vimos que la inercia total de la tabla completa era de 0,01873; por tanto, la inercia atribuible a las consonantes es del 87,4% (0,01637 con relación a 0,01873) de la inercia total. Así, pues, una vez visto que la mayor parte de la inercia es atribuible a las consonantes, no debe sorprendernos que los mapas del análisis completo de la imagen 10.7 y en el análisis de subgrupos de la imagen 21.1 sean muy similares.

AC de subgrupos: análisis de vocales, biplot estándar

Hemos visto que podemos descomponer la inercia total de la tabla original en inercia de consonantes e inercia de vocales de la siguiente manera:

inercia total = inercia de consonantes + inercia de vocales
0,01873 = 0,01637 + 0,00236
(87,4%) (12,6%)

Imagen 21.2:
Análisis de subgrupos de vocales en el ejemplo sobre los autores; biplot estándar de filas, es decir, filas (textos) en coordenadas principales y columnas (letras) en coordenadas estándares multiplicadas por las raíces cuadradas de las masas de las columnas



A pesar de que las vocales son letras relativamente más frecuentes (las cinco vocales representan el 38,3% del total de letras, mientras que las 21 consonantes representan el 61,7%), la inercia de la submatriz de vocales es mucho menor, sólo el 12,6% de la inercia total original. En el mapa de la imagen 21.2 mostramos, igual que antes para las consonantes, el biplot estándar de las vocales. La menor dispersión de los textos con relación a los vectores de las letras es muy aparente y contrasta con el mapa de la imagen 21.1. Sin embargo, bastantes pares de textos siguen hallándose muy cerca. Podemos observar que las letras *e* y *o* se hallan en posiciones opuestas, a la izquierda y a la derecha, respectivamente. Asimismo se hallan en posiciones opuestas los textos de Buck y los de Faulkner. De los seis autores, los textos de Holt parecen ser los más distintos. En el capítulo 25 veremos las pruebas de permutaciones que nos permitirán contrastar la significación de estos resultados. Anticipándonos un poco, podemos indicar que el emparejado de textos en los mapas es altamente significativo tanto para las consonantes como para las vocales.

ACM de subgrupos

Cuando trabajemos con datos categóricos multivariantes, la idea de dividir la matriz original en submatrices y luego aplicar el ACM es muy útil para investigar si existen estructuras en determinadas submatrices. Así, en datos procedentes de encuestas puede ser interesante, desde un punto de vista sustantivo, centrarnos en un determinado subgrupo de respuestas. Por ejemplo, centrarnos sólo en las categorías de acuerdo, en una escala de acuerdo/desacuerdo con cinco respuestas posibles, o centrarnos únicamente en las respuestas intermedias («ni de acuerdo ni en desacuerdo»), o en respuestas no sustantivas («no sabe», «no contesta», «otras», etc.). O, simplemente, podría ser que quisiéramos excluir las respuestas no sustantivas y concentrarnos sólo en las que sí lo son. En todos estos casos, el análisis de subgrupos nos permitirá ver de forma más clara la relación entre este tipo especial de respuestas y las variables demográficas, lo que posiblemente no ocurriría si analizáramos todas las respuestas conjuntamente. La posibilidad de hacer submatrices nos permite, para diferentes grupos de categorías, dividir la variabilidad de los datos en partes que, luego, podemos visualizar separadamente. La manera de realizar el ACM de subgrupos consiste en llevar a cabo el AC de subgrupos a las partes adecuadas de la matriz binaria o de la matriz de Burt, como veremos a continuación.

Análisis de subgrupos de la matriz binaria

Volvamos a los datos sobre el trabajo de las mujeres que introducimos en el capítulo 17 y analizamos en el capítulo 18 utilizando el ACM. Las cuatro preguntas tienen prevista una categoría, etiquetada en los mapas con el símbolo *?*, para las respuestas del tipo «no sabe» y las respuestas perdidas. Estas categorías tienen un papel muy prominente en el primer eje principal del ACM (imagen 18.2). Vamos realizar un análisis de subgrupos restringido a las respuestas sustantivas de las cuatro variables, etiquetadas como *T* (trabajo a tiempo completo), *t* (trabajo a tiempo parcial) y *C* (permanecer en casa), prescindiendo así de las columnas de la matriz binaria correspondientes a respuestas no sustantivas *?*; en el análisis utilizaremos los valores marginales de filas y de columnas de la matriz binaria original. Dado que en este caso en particular, las sumas de las filas de la matriz binaria son iguales a 4, en el análisis de subgrupos para la ponderación de las filas (encuestados) mantendremos este valor. Los valores de los perfiles seguirán siendo ceros o $\frac{1}{4}$. Es decir, los encuestados con cuatro respuestas sustantivas tendrán cuatro valores $\frac{1}{4}$ en sus perfiles, mientras que los que tengan tres respuestas sustantivas tendrán tres valores $\frac{1}{4}$ y así sucesivamente. Por el contrario, si simplemente prescindiéramos de las columnas no sustantivas y lleváramos a cabo el AC ordinario sobre la matriz binaria, tendríamos valores de $\frac{1}{3}$ para los encuestados con tres respuestas sustantivas, $\frac{1}{2}$ con dos, y 1 con sólo una. Además, sería imposible calcular los perfiles de los casos con cuatro respuestas no sustantivas. Cosa que no ocurre en el AC de subgrupos. Este tipo de casos con sólo ceros se sitúan en el origen en el mapa. La inercia total del AC de subgrupos, con 12 categorías, es de 2,1047. Dado que la inercia total la matriz binaria completa es de 3, vemos que

Imagen 21.3:
Matriz de Burt de las cuatro variables categóricas de la imagen 18.4, arreglada de manera que todas las categorías correspondientes a respuestas no sustantivas (?) se hallan en la últimas filas y columnas. Todas las respuestas sustantivas (T, t y C), 12 × 12, se hallan en la parte superior izquierda, mientras que la esquina inferior izquierda de 4 × 4 contiene la concurrencia de las respuestas no sustantivas («no sabe/valores perdidos»)

1T	1t	1C	2T	2t	2C	3T	3t	3C	4T	4t	4C	1?	2?	3?	4?
2501	0	0	172	1107	1131	355	1710	345	1766	538	40	0	91	91	157
0	476	0	7	129	335	16	261	181	128	293	17	0	5	18	38
0	0	79	1	6	72	1	17	61	14	21	38	0	0	0	6
172	7	1	181	0	0	127	48	4	165	15	0	1	0	2	1
1107	129	6	0	1299	0	219	997	61	972	239	13	57	0	22	75
1131	335	72	0	0	1646	24	989	573	760	616	84	108	0	60	186
355	16	1	127	219	24	379	0	0	360	14	1	7	9	0	4
1710	261	17	48	997	989	0	2084	0	1348	567	23	96	50	0	146
345	181	61	4	61	573	0	0	642	202	286	73	55	4	0	81
1766	128	14	165	972	760	360	1348	202	1959	0	0	51	62	49	0
538	293	21	15	239	616	14	567	286	0	897	0	45	27	30	0
40	17	38	0	13	84	1	23	73	0	0	97	2	0	0	0
0	0	0	1	57	108	7	96	55	51	45	2	362	196	204	264
91	5	0	0	0	0	9	50	4	62	27	0	196	292	229	203
91	18	0	2	22	60	0	0	0	49	30	0	204	229	313	234
157	38	6	1	75	186	4	146	81	0	0	0	264	203	234	465

la inercia se ha descompuesto en 2,1047 (70,2%) para las categorías sustantivas y 0,8953 (29,8%) para las no sustantivas. Las inercias principales y los porcentajes de inercia de las dos primeras dimensiones de la submatriz analizada son de 0,5133 (24,4% del total de 2,1047) y de 0,3652 (17,4%) sobre este mismo total. Por tanto, el porcentaje global de inercia explicada por la solución bidimensional es del 41,8%. Igual que ocurría en el ACM, estos porcentajes son artificialmente bajos. Como vimos en el capítulo 19, y como veremos a continuación, podemos mejorar el mapa implementando un ajuste de los factores de escala de los ejes.

Análisis de subgrupos de la matriz de Burt

Igual que vimos en el ACM, podemos mejorar el mapa del AC de subgrupos llevando a cabo el análisis en la parte apropiada de la matriz de Burt. Para ilustrar este procedimiento, consideremos la matriz de Burt que vimos en la imagen 18.4 del capítulo 18. Podemos reacomodar dicha matriz de manera que todas las categorías de la submatriz de interés se hallen, como mostramos en la imagen 21.3, en la parte superior izquierda de la tabla concatenada. Así, la submatriz de interés de 12 × 12 está formada por cuatro tablas con tres respuestas sustantivas en cada una de ellas. Ahora, las cuatro respuestas no sustantivas se hallan en las últimas cuatro filas y cuatro columnas de la matriz. El AC de subgrupos da una inercia total de 0,6358, y unas inercias principales (y porcentajes) de 0,26354 (41,4%) y de 0,1333 (21,0%) para las dos primeras dimensiones. Igual que ocurría con el ACM,

obtenemos una mejora del mapa con relación al AC de subgrupos llevado a cabo en la matriz binaria. Ahora llegamos a explicar el 62,4% de la inercia, mientras que con el análisis anterior llegábamos a explicar el 41,8% de la inercia. Fijémonos también en que la relación entre el AC de subgrupos en la matriz binaria y en la matriz de Burt es la misma que vimos para el ACM habitual: las inercias principales en el AC de subgrupos en la matriz de Burt son los cuadrados de las de la matriz binaria, así, por ejemplo, $0{,}2635 = 0{,}5133^2$.

Análisis de subgrupos con una solución e inercias ajustadas

El problema de las bajas inercias es el mismo que vimos con el ACM. Efectivamente, en la diagonal de la tabla concatenada de la imagen 21.3, podemos ver matrices diagonales de 3×3 que interfieren en los resultados del AC del subgrupo de interés. Al igual que antes, es posible ajustar el resultado mediante análisis de la regresión, de manera que se ajusten de forma óptima las matrices que se hallan fuera de la diagonal. Esto implica disponer en forma de vector los elementos de las seis tablas situadas fuera de la diagonal, cada una de ellas con nueve elementos, como si fuera un vector de 54 elementos y, así, constituir los elementos de la variable *«y»* de la regresión. Tenemos que expresar estos elementos como en (19.2), es decir, como cocientes de contingencia menos 1. Formaremos las dos variables *«x»* (para la solución bidimensional) multiplicando las correspondientes coordenadas estándares. Hallaremos los valores óptimos de los factores de escala como anteriormente, por mínimos cuadrados ponderados (cap. 19), y así obtenemos un ajuste de $R^2 = 0{,}849$. Desafortunadamente en este caso no parece que exista atajo alguno, como ocurría para el ACM [ecuaciones (19.5) y (19.7)]. A partir de la regresión de mínimos cuadrados ponderada obtenemos los factores de escala 0,3570 y 0,1636, que hemos utilizado para obtener las coordenadas principales y el mapa de la imagen 21.4. Los cuadrados de estos factores de escala son las inercias principales, 0,1275 y 0,0268, que mostramos en los ejes. El porcentaje de inercia explicado por la solución bidimensional ajustada, R^2, es del 84,9% (como vimos anteriormente, no podemos calcular los porcentajes de los ejes individuales ya que la solución no es anidada).

Puntos adicionales en el AC de subgrupos

El procedimiento de representación de puntos adicionales depende de cómo hayamos dividido en grupos las filas o las columnas. Así, por ejemplo, en el caso de los datos de textos de diferentes autores, hemos dividido las letras en dos grupos, y analizamos la submatriz de vocales (imagen 21.2). En este caso, las filas (textos) que no hemos dividido en grupos están centradas. Sin embargo, las columnas, que sí se han dividido, no lo están. Si quisiéramos proyectar la letra *Y* sobre el mapa del AC del grupo de las vocales, para no tener que centrar el perfil de la *Y*, lo que haríamos es utilizar las coordenadas centradas en cero ϕ_{ik} (es decir, los vértices de las filas). Es decir, la media ponderada usual proporciona las coordenadas principales (véase el cap. 12) y la fórmula específica de transición (14.2) aplicable a este caso (para una solución bidimensional) es:

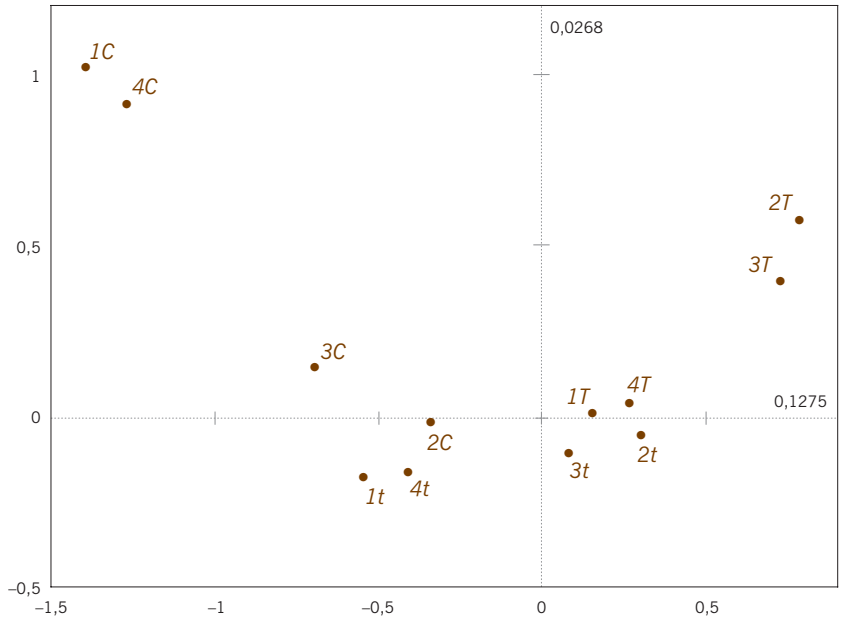


Imagen 21.4:
Mapa del AC de subgrupos de las respuestas categóricas sustantivas (excluidas las respuestas no sustantivas). Hemos ajustado la solución para hallar el mejor ajuste de las tablas de fuera de la diagonal, lo que lleva a una mejora considerable del ajuste total, explicándose el 84,9% de la inercia

$$\sum_i y_i \phi_{ik} \qquad k = 1,2 \tag{21.1}$$

donde y_i es el i-ésimo valor de perfil de Y. Por otra parte, si quisiéramos proyectar un nuevo texto con valores de perfil t_j en el subgrupo de interés (su suma es igual a la proporción de vocales en ese texto, no es 1), tendríamos que centrar los datos con relación a los valores originales c_j del centroide, antes de llevar a cabo el producto escalar con las coordenadas estándares de las columnas γ_{jk}:

$$\sum_j (t_j - c_j)\gamma_{jk} \qquad k = 1,2 \tag{21.2}$$

Fijémonos que para situar un punto adicional en el AC de subgrupos y en el AC habitual, se puede hacer siempre este tipo de centrado. Sin embargo, no es necesario cuando las coordenadas estándares cumplen que $\sum_i r_i \phi_{ik} = 0$ y $\sum_j c_j \gamma_{jk} = 0$, lo que ocurre cuando la suma se lleva a cabo con relación a todos los datos.

Puntos adicionales en el ACM de subgrupos

Los encuestados (filas) de la matriz binaria, así como cualquier agrupación de filas, por ejemplo, según el nivel de educación, género, etc., se pueden representar como puntos adicionales. Igual que en el ACM usual, representamos las categorías de las variables adicionales en los centroides de los puntos de los encuestados que se hallan en estos grupos. En el mapa de la imagen 21.5 mostramos las posiciones de varias categorías demográficas con relación a los mismos ejes principales de la imagen 12.4.

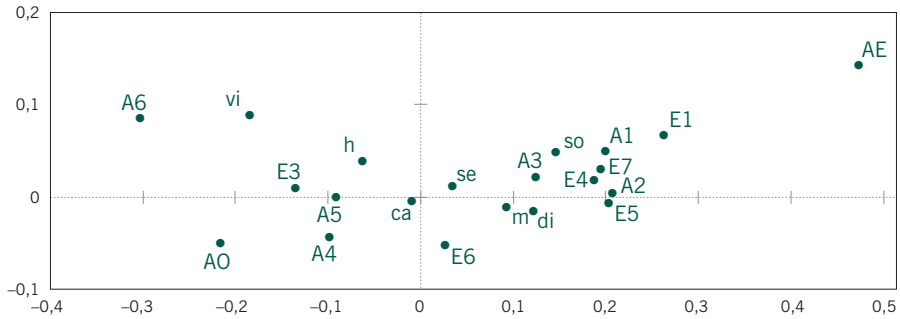


Imagen 21.5:
Posiciones de los puntos adicionales en el mapa de la imagen 21.4

Nota: El significado de las abreviaciones se puede consultar en el capítulo 17, página 161; AO y AE indican Alemania Occidental y del Este, respectivamente

RESUMEN: Análisis de correspondencias de subgrupos

1. La idea en el AC de *subgrupos* es visualizar una submatriz de filas o de columnas (o de ambas) en subespacios del espacio original que se obtiene con todos los datos. En este procedimiento, mantenemos el centroide original en el centro del mapa, también mantenemos las masas originales y los pesos de las distancias χ^2.
2. Dado que en el análisis de subgrupos se mantienen las propiedades del espacio original, podemos descomponer la inercia total original en partes que corresponden a las inercias de las distintas submatrices que componen la matriz original.
3. Podemos implementar fácilmente el AC de subgrupos dejando de calcular los valores marginales de las submatrices, de manera que en todos los cálculos habituales del AC utilizamos los valores marginales originales (masas).
4. Aplicar el ACM a determinadas submatrices de categorías nos proporciona una estrategia de análisis que puede ser muy útil en el análisis de datos procedentes de cuestionarios. Por ejemplo, podemos omitir los valores perdidos. Podemos centrarnos, para todas las respuestas, en un determinado tipo de categorías, y así visualizar las dimensiones de esta submatriz sin interferencias de las otras categorías.
5. Igual que en el ACM habitual, podemos aplicar el ACM de subgrupos a las matrices binaria y de Burt. A continuación podemos redimensionar los resultados para así optimizar el ajuste de las submatrices. Esto permite mejorar mucho los porcentajes de inercia explicados en el mapa.
6. Podemos añadir puntos adicionales en el mapa de una determinada submatriz. En el ACM de subgrupos esta posibilidad nos permite poder relacionar determinadas respuestas con categorías demográficas.

CAPÍTULO 9

Análisis de tablas cuadradas

En este capítulo vamos a considerar el caso particular de las tablas de frecuencias cuadradas. Es decir, cuando filas y columnas hacen referencia a los mismos objetos en dos circunstancias distintas. Encontramos este tipo de datos en muchas situaciones. Por ejemplo, en las tablas sobre la movilidad social, en las matrices de confusión utilizadas en psicología, en tablas de cambio de marcas de los consumidores, en investigación de mercados, en el estudio de referencias cruzadas de artículos científicos entre revistas, en matrices de transición entre comportamientos y en tablas de migración. A menudo este tipo de tablas se caracterizan por tener valores relativamente elevados en la diagonal. Estos valores indican una asociación muy fuerte que fácilmente enmascara las asociaciones más sutiles situadas fuera de la diagonal, que no quedan reflejadas en los ejes principales. Una aproximación para poder aplicar el AC a tablas cuadradas es dividir el análisis en dos partes: 1) un análisis de la parte *simétrica* de la tabla, que absorbe la mayor componente de la inercia, incluyendo la diagonal y 2) un análisis de la parte *antisimétrica* de la tabla, que contiene la información situada fuera de la diagonal. Es la visualización de esta última componente la que muestra la magnitud y el sentido del «flujo» de filas a columnas y viceversa.

Contenido

Conjunto de datos 12: movilidad social y profesiones de padres e hijos

Para entrar rápidamente en materia, consideremos unos datos clásicos sobre movilidad social. Son unos datos sobre profesiones de padres e hijos que fueron publicados por Karl Pearson hace más de 100 años. Los mostramos en la tabla de la imagen 22.1. Para crear esta tabla se hizo un recuento de las profesiones de padres e hijos. Dado que muchos hijos tenían la misma profesión que sus padres, los valores de la diagonal son elevados, algo que suele ser habitual en este tipo de tablas. Sin embargo, también suelen aparecer asimetrías destacables. Así, en el ejemplo que nos ocupa, podemos ver que la suma de los valores de la primera fila (padres militares) es 50, mientras que la suma de la primera columna (hijos militares) es 84. El flujo de hijos hacia el ejército se produce de padres propietarios rurales (fila 7) y comerciantes (fila 10). Por otra parte, los hijos de padres comerciantes experimentan un gran flujo de salida hacia otras profesiones. Así, observamos que hay 106 padres comerciantes, en cambio, los hijos que han elegido el comercio son 24. El flujo de salida se produce hacia artistas, teólogos, escritores y militares.

AC de tablas cuadradas

Podemos utilizar el AC para visualizar esta tabla de contingencia (imagen 22.2). La tabla tiene una inercia elevada (1,297), sin duda debido a las fuertes asociaciones existentes entre filas y columnas. Por tanto es adecuado utilizar el mapa asimétrico. Es decir, expresaremos los perfiles de los padres en coordenadas principales, y los de los hijos en coordenadas estándares. Si el perfil de una profesión de los padres tiene todo ceros excepto el valor de la diagonal, ello indica que esa profesión se halla en el vértice de esa profesión. Así, la segunda fila, correspondiente a artistas, es casi de este tipo, tienen el valor relativo más elevado (51 de

Imagen 22.1: *Tabla de contingencia correspondiente a las profesiones de padre e hijos. Vemos, por ejemplo, que las profesiones de los hijos de los 50 padres* militares *son: 28* militares, *4* maestros, *1* propietario rural, *3* abogados, *etc.*

	Profesión hijos														
Profesión padres	*MIL*	*ART*	*MAE*	*ARTE*	*TEO*	*AGR*	*PROP*	*ABO*	*ESC*	*COM*	*MED*	*MAR*	*PO*	*D+C*	*Sumas*
Militares	28	0	4	0	0	0	1	3	3	0	3	1	5	2	*50*
Artistas	2	51	1	1	2	0	0	1	2	0	0	0	1	1	*62*
Maestros	6	5	7	0	9	1	3	6	4	2	1	1	2	7	*54*
Artesanos	0	12	0	6	5	0	0	1	7	1	2	0	0	10	*44*
Teólogos	5	5	2	1	54	0	0	6	9	4	12	3	1	13	*115*
Agricultores	0	2	3	0	3	0	0	1	4	1	4	2	1	5	*26*
Propietarios rurales	17	1	4	0	14	0	6	11	4	1	3	3	17	7	*88*
Abogados	3	5	6	0	6	0	2	18	13	1	1	1	8	5	*69*
Escritores	0	1	1	0	4	0	0	1	4	0	2	1	1	4	*19*
Comerciantes	12	16	4	1	15	0	0	5	13	11	6	1	7	15	*106*
Médicos	0	4	2	0	1	0	0	0	3	0	20	0	5	6	*41*
Marinos	1	3	1	0	0	0	1	0	1	1	1	6	2	1	*18*
Políticos	5	0	2	0	3	0	1	8	1	2	2	3	23	1	*51*
Docentes y científicos	5	3	0	2	6	0	1	3	1	0	0	1	1	9	*32*
Sumas	*84*	*108*	*37*	*11*	*122*	*1*	*15*	*64*	*69*	*24*	*57*	*23*	*74*	*86*	*775*

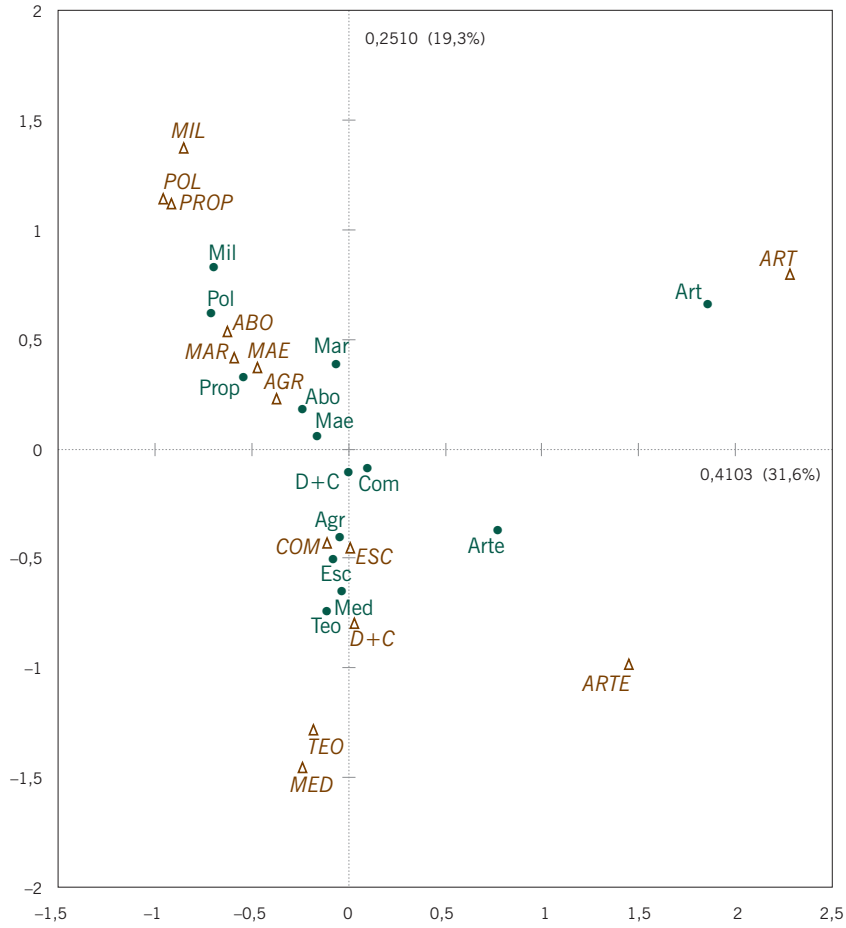


Imagen 22.2:
Mapa asimétrico del AC sobre los datos de movilidad de la imagen 22.1, las filas en coordenadas principales. Porcentaje de inercia explicado: 51,0%.

62, el 82%) de padres e hijos con la misma profesión. En el mapa de la imagen 22.2 vemos que esta situación se refleja por el hecho de que artistas (Art) quedan muy alejados. Los padres artistas (Art) casi alcanzan el vértice correspondiente a los hijos artistas *(ART)*. La fila de artesanos (Arte) se halla entre los vértices *ART* y *D+C* (docentes y científicos) ya que, en términos relativos, bastantes hijos de padres artesanos acaban en estas dos profesiones (fila 4 de la imagen 22.1)

La diagonal de la tabla domina el AC

El problema que encontramos al tratar de visualizar una matriz cuadrada de este tipo es la presencia de una poderosa diagonal que tiende a dominar el análisis. Dado que el AC trata de explicar tanta inercia como sea posible, no es de extrañar que el resultado del análisis se vea muy influido por la gran fuente de inercia que representa la diagonal, en detrimento del resto de la tabla que, sin embargo, contiene interesantes flujos entre profesiones de padres e hijos. Para apoyar esta

afirmación con algunos números, basta con ver que los 14 valores de la diagonal explican el 70,9% de la inercia total, mientras que los 182 valores que se hallan fuera de la diagonal contribuyen con sólo el 29,1% de la inercia; es decir, podemos descomponer la inercia total de la manera siguiente:

$$\begin{aligned}\text{inercia total} &= \text{inercia de la diagonal} + \text{inercia fuera de la diagonal}\\ 1{,}2974 &= 0{,}9100 + 0{,}3774\\ 100\% &= 70{,}9\% + 29{,}1\%\end{aligned} \tag{21.1}$$

El mapa de la imagen 22.2 explica una inercia de 0,6613, que es el 51,0% de la inercia total. Podemos repartir este valor entre los elementos de la diagonal y los que quedan fuera de ésta, de la siguiente manera:

$$\begin{aligned}\text{inercia explicada} &= \text{inercia explicada de la diagonal} + \text{inercia explicada fuera de la diagonal}\\ 0{,}6613\ (51{,}0\%) &= 0{,}5483\ (59{,}6\%) + 0{,}1130\ (29{,}9\%)\end{aligned}$$

Hemos expresado los porcentajes entre paréntesis con relación a las inercias de la expresión (22.1). Vemos que los elementos de fuera de la diagonal se explican pobremente en comparación con los de la diagonal.

Simetría y antisimetría en la tabla cuadrada

Podemos descomponer la tabla en dos partes: una que contenga la parte *simétrica* de la tabla, es decir, el flujo medio entre filas y columnas, y otra parte que incluya la componente *antisimétrica* que cuantifique el flujo diferencial. Podemos expresar la descomposición de la tabla original, simbolizada por **N**, de la siguiente manera:

$$\begin{aligned}\mathbf{N} &= \frac{1}{2}(\mathbf{N}+\mathbf{N}^{\mathsf{T}})+\frac{1}{2}(\mathbf{N}-\mathbf{N}^{\mathsf{T}})\\ &= \mathbf{S}+\mathbf{T}\end{aligned} \tag{22.2}$$

donde **S** es la parte simétrica, que contiene las medias de los valores situados en los lados opuestos de la diagonal, y **T** la componente antisimétrica, con la mitad de las diferencias:

$$s_{ij} = \frac{1}{2}(n_{ij}+n_{ji}) \qquad t_{ij} = \frac{1}{2}(n_{ij}-n_{ji}) \tag{22.3}$$

La siguiente expresión ilustra esta descomposición para los datos situados arriba a la izquierda en la tabla de la imagen 22.1:

$$\begin{bmatrix} 28 & 0 & 4 & 0 & \cdots \\ 2 & 51 & 1 & 1 & \cdots \\ 6 & 5 & 7 & 0 & \cdots \\ 0 & 12 & 0 & 6 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} 28 & 1 & 5 & 0 & \cdots \\ 1 & 51 & 3 & 6{,}5 & \cdots \\ 5 & 3 & 7 & 0 & \cdots \\ 0 & 6{,}5 & 0 & 6 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} + \begin{bmatrix} 0 & -1 & -1 & 0 & \cdots \\ 1 & 0 & -2 & -5{,}5 & \cdots \\ 1 & 2 & 0 & 0 & \cdots \\ 0 & 5{,}5 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

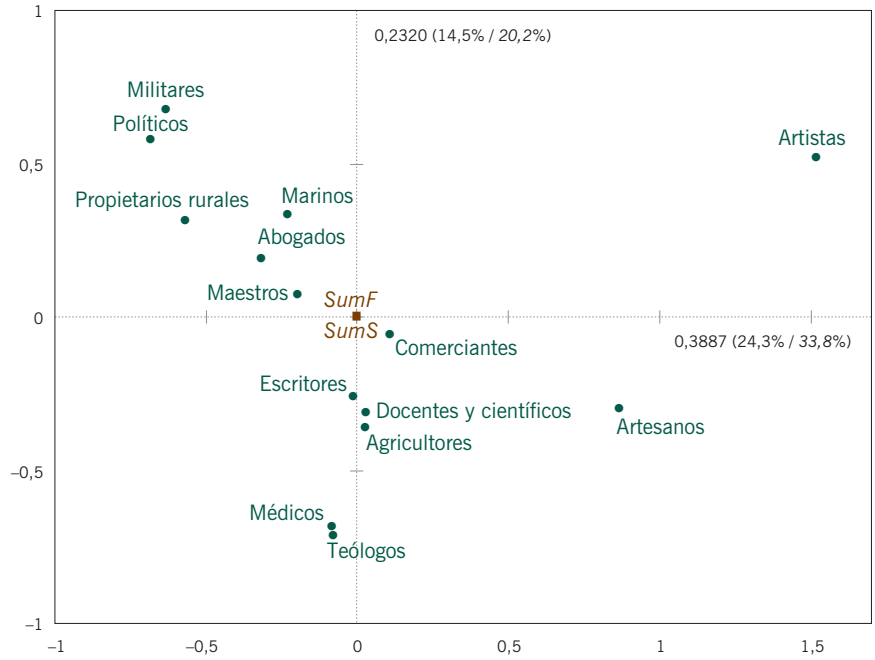


Imagen 22.3:
AC de la parte simétrica de la imagen 22.1. Los primeros porcentajes se han calculado con relación a la inercia total de 1,5991, mientras que los porcentajes en cursiva se han calculado con relación a la inercia de la parte simétrica de 1,1485

Por ejemplo, transformamos el valor 1 correspondiente a la segunda fila 1 (padre artista) y cuarta columna (hijo artesano) y el valor 12 correspondiente a la cuarta fila (padre artesano) y segunda columna (hijo artista), en su valor medio 6,5 que situados en las celdas correspondientes de **S**, y las desviaciones de la media ± 5,5 que situamos en las correspondientes celdas de **T**. La matriz simétrica tiene la misma diagonal que la tabla original y como hemos visto $s_{ij} = s_{ji}$. En cambio, la matriz antisimétrica tiene ceros en la diagonal. La antisimetría consiste en que los elementos situados en los lados opuestos de la diagonal tienen los mismos valores absolutos pero signos opuestos, es decir, $t_{ij} = -t_{ji}$.

AC de la parte simétrica

Aplicaremos el AC a la matriz simétrica y a la matriz antisimétrica, de forma separada. En el mapa de la imagen 22.3 mostramos el mapa de la matriz simétrica, con sólo unas coordenadas, ya que las coordenadas de filas y columnas son idénticas. Salvo por el hecho de que cada profesión tiene un solo punto, el mapa se parece mucho al de la imagen 22.2. El mapa muestra la asociación global entre profesiones de padres e hijos. En los ejes, el primer porcentaje, entre paréntesis, hace referencia a la inercia explicada con relación a la tabla asimétrica original, mientras que el porcentaje en cursiva, también entre paréntesis, hace referencia a la inercia explicada con relación a la inercia total de la parte simétrica **S** visualizada. Fijémonos en que los valores marginales de filas y de columnas de **S** son medias de los valores marginales de la matriz asimétrica **N**. De esta manera, si las masas

de filas y de columnas de esta última matriz son **r** y **c**, respectivamente, entonces, las masas de las filas y de las columnas de **S** son $\mathbf{w} = \frac{1}{2}(\mathbf{r} + \mathbf{c})$.

AC de la parte antisimétrica

Antes de que podamos aplicar el AC a la matriz antisimétrica **T**, tenemos que superar dos problemas. El primero es que **T** tiene valores positivos y valores negativos. En realidad, la suma de los elementos de la matriz es cero. Por tanto, no tiene sentido centrarla con relación a sus valores marginales, el primer paso del algoritmo del AC. Por tanto, tenemos que cambiar el algoritmo de manera que el AC analice los datos sin centrarlos, y así realizar directamente la normalización que conduce a las distancias χ^2. Sin embargo, ello nos conduce al segundo problema: la suma de filas y de columnas, como masas, no tiene sentido. La solución obvia es adoptar las mismas masas que en **S**, es decir, las masas **w** definidas anteriormente. Parece, pues, que necesitamos un algoritmo especial para analizar **T**. No obstante, existe un procedimiento que nos permite obtener la solución de forma simple recodificando los datos.

AC simultáneo de las partes simétrica y antisimétrica

El mencionado procedimiento de recodificación evita tener que implementar un algoritmo especial para hallar la solución del AC de la matriz antisimétrica. La idea es formar una matriz que sea cuatro veces la tabla original **N** con el siguiente formato (es fácil de obtener con R: véase la página 317, o con una hoja de cálculo):

$$\begin{bmatrix} \mathbf{N} & \mathbf{N}^\top \\ \mathbf{N}^\top & \mathbf{N} \end{bmatrix} \tag{22.4}$$

Es decir, formamos una nueva matriz compuesta en la que situamos la tabla original **N**, arriba a la izquierda y abajo a la derecha, y su transpuesta arriba a la derecha y abajo a la izquierda. A continuación llevamos a cabo el AC en esta nueva matriz compuesta. Si **N** es una matriz $I \times I$, entonces la nueva matriz compuesta es una matriz $2I \times 2I$, con $2I - 1$ dimensiones, $I - 1$ de las cuales corresponden exactamente a las dimensiones de la matriz simétrica **S** y el resto a las de la matriz antisimétrica **T**. Es fácil saber qué dimensiones corresponden a cada parte, ya que las dimensiones de la matriz antisimétrica siempre ocurren en pares de inercias principales iguales. Por ejemplo, en la tabla de la imagen 22.4, mostramos las 27 inercias principales (valores propios) correspondientes a los datos sobre la movilidad social, donde $I = 14$. Los siete pares de dimensiones con inercias principales iguales (que mostramos en negrita) corresponden al análisis antisimétrico y son: 3 y 4, 10 y 11, 14 y 15, 17 y 18, 19 y 20, 22 y 23, y finalmente 25 y 26. Las restantes 13 dimensiones corresponden al análisis simétrico. La inercia total de la matriz simétrica es la suma de sus respectivas 13 inercias principales: 0,3887 + 0,2320 + 0,1439 + ... = 1,1485, lo que corresponde al 71,8% de la inercia total de 1,5991. La inercia total de la matriz antisimétrica es la suma de 7 pares: 2 × 0,1584 + 2 × 0,0418 + ... = 0,4508, lo que corresponde al 28,2% del total (fijémonos en que la inercia total de 1,5991, es

Dim.	*Inercia principal*	*Dim.*	*Inercia principal*	*Dim.*	*Inercia principal*
1	0,38868	10	**0,04184**	19	**0,00309**
2	0,23204	11	**0,04184**	20	**0,00309**
3	**0,15836**	12	0,02287	21	0,00166
4	**0,15836**	13	0,02205	22	**0,00115**
5	0,14391	14	**0,01287**	23	**0,00115**
6	0,12376	15	**0,01287**	24	0,00062
7	0,08184	16	0,01036	25	**0,00038**
8	0,07074	17	**0,00759**	26	**0,00038**
9	0,04984	18	**0,00759**	27	0,00015

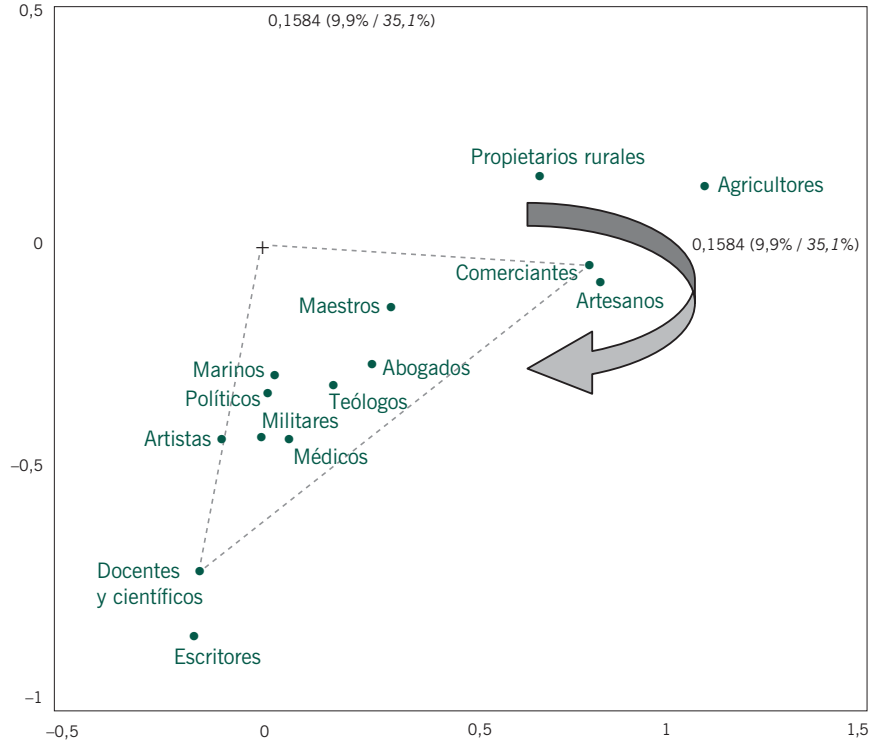
Imagen 22.4:
Inercias principales de las 27 dimensiones del análisis de la matriz compuesta de 28 × 28 (22.4) formada a partir de los datos sobre movilidad social. Las inercias principales que ocurren en pares corresponden a la parte antisimétrica

mayor que la inercia de la matriz original, de 1,2974 (22.1), ya que hemos centrado la tabla con relación a los valores marginales medios **w** de filas y columnas).

Visualización de las partes simétrica y antisimétrica

Como podemos ver en la tabla de la imagen 22.4, las dimensiones 1 y 2 son las que mejor visualizan la matriz simétrica. Explican una inercia de 0,6217 de un total de 1,1485, el 54,0%, El AC de la matriz compuesta proporciona los resultados de filas y de columnas repetidos; por tanto, utilizaremos sólo un conjunto de coordenadas principales para dibujar el mapa (para las dimensiones 1 y 2, hemos utilizado el primer conjunto de coordenadas de la imagen 22.6; véase también el apéndice de cálculo). Las dimensiones 3 y 4 son las mejores para visualizar la matriz antisimétrica. Explican 0,3167 de 0,4506, es decir, el 70,3% de la inercia de la parte antisimétrica. El mapa de la parte antisimétrica que mostramos en la imagen 22.5 tiene algunas propiedades particulares. La primera es que, debido a la igualdad de las inercias principales, las coordenadas pueden girar libremente en el mapa bidimensional. No se identifican con relación a los ejes principales, por tanto, no dibujamos estos ejes en el mapa. La segunda es que, debido a la antisimetría de la matriz, los resultados del AC de la matriz compuesta aparecen repetidos, pero con el signo cambiado. En el mapa dibujaremos un sólo conjunto de puntos —para dibujar el mapa de la imagen 22.5 hemos utilizado el segundo conjunto de coordenadas para las dimensiones 3 y 4 de la tabla de la imagen 22.6—. Para interpretar este tipo de mapas no utilizamos las distancias entre puntos, sino que interpretamos las áreas triangulares formadas por pares de puntos y el origen. Por ejemplo, comerciantes y docentes-científicos definen un gran triángulo con relación al origen, que interpretamos como un fuerte flujo diferencial entre estas dos profesiones. La flecha en el sentido de las agujas del reloj, que hemos dibujado en el mapa, indica el sentido del flujo de padres a hijos: los hijos de padres comerciantes son con relativa frecuencia docentes-científicos (en la tabla de la imagen 22.1 podemos ver que la frecuencia de padres comerciantes – hijos docente-científicos es 15, mientras que en sentido contrario la frecuencia es 0). Por tanto, generalizando, podemos decir que propietarios rurales, agricultores, comerciantes y artesanos experimentan flujos de salida hacia es-

Imagen 22.5:
AC de la parte antisimétrica de la tabla de la imagen 22.1. Hemos calculado los primeros porcentajes con relación a la inercia total de 1,5991, mientras que los porcentajes en cursiva se han calculado con relación a la inercia de la parte antisimétrica de 0,4506



critores y docentes-científicos. Algunos pares de profesiones definen, con relación al origen, áreas triangulares muy pequeñas como, por ejemplo, militares, políticos y marinos, lo que indica que no se produce un flujo entre estas profesiones. Sin embargo, podrían experimentar flujos de entrada de agricultores, artesanos, etc.

Imagen 22.6:
Coordenadas principales de algunas filas de la matriz compuesta de 28 × 28 (22.4) de los datos sobre movilidad social. Las coordenadas de las dimensiones simétricas (en este caso las dos primeras) son simples repeticiones, mientras que las de la parte asimétrica (las dimensiones 3 y 4) son iguales, pero de signo opuesto

Profesión	*Dim. 1*	*Dim. 2*	*Dim. 3*	*Dim. 4*	⋯
Militares	–0,632	0,671	–0,011	0,416	⋯
Artistas	1,521	0,520	0,089	0,423	⋯
Maestros	–0,195	0,073	–0,331	0,141	⋯
Artesanos	0,867	–0,298	–0,847	0,092	⋯
Teólogos	–0,077	–0,709	–0,189	0,305	⋯
⋮	⋮	⋮	⋮	⋮	
Militares	–0,632	0,671	0,011	–0,416	⋯
Artistas	1,521	0,520	–0,089	–0,423	⋯
Maestros	–0,195	0,073	0,331	–0,141	⋯
Artesanos	0,867	–0,298	0,847	–0,092	⋯
Teólogos	–0,077	–0,709	0,189	–0,305	⋯
⋮	⋮	⋮	⋮	⋮	⋯

RESUMEN:
Análisis de tablas cuadradas

1. Las tablas cuadradas con el mismo tipo de elementos en filas y en columnas son especiales debido a que los valores de su diagonal desempeñan un papel predominante en el análisis que, a menudo, enmascara las estructuras de la tabla que quedan fuera de la diagonal.
2. Una alternativa al AC habitual es dividir la tabla en dos partes: una parte *simétrica* y otra *antisimétrica*, de manera que esta última —en general, de menor inercia que la parte simétrica— capture las asimetrías de la tabla.
3. Analizamos la parte simétrica de la forma habitual, sin embargo, para analizar la parte antisimétrica utilizamos un algoritmo de cálculo del AC que elimina el centrado y la normalización de la tabla con respecto a sus valores marginales; para estos datos no tendría sentido utilizar los valores marginales como masas.
4. En ambos análisis, las masas que utilizamos para ponderar y para las distancias χ^2 son las medias de las masas de filas y de columnas de la tabla original.
5. Otra posibilidad de análisis es que apliquemos el AC a una matriz compuesta formada por la tabla original arriba a la izquierda y abajo a la derecha, y su transpuesta arriba a la derecha y abajo a la izquierda. Las dimensiones correspondientes a la parte simétrica tienen inercias principales únicas, mientras que las de la parte antisimétrica ocurren en pares iguales.
6. Interpretamos el mapa de la parte simétrica de la forma habitual, es decir, este mapa muestra la asociación global entre los elementos de la tabla.
7. Sin embargo, interpretamos el mapa de la parte antisimétrica, que posee una geometría especial, con relación a las áreas de triángulos formados por pares de puntos y el origen. Dichas áreas indican la intensidad de las asimetrías entre estos pares de elementos de la tabla. La dirección de la asimetría es la misma para todos los pares de puntos del mapa.

CAPÍTULO 23

Recodificación de datos

Hasta ahora, en los 22 capítulos anteriores hemos trabajado con tablas de frecuencias; con tablas simples (capítulos 1 a 16 y 22) o con tablas compuestas (capítulos 17 a 21). En este capítulo vamos a tratar con datos de naturaleza distinta. Veremos cómo recodificarlos, o transformarlos, para que podamos visualizarlos utilizando AC. Este procedimiento fue muy bien desarrollado por Benzécri que, antes de visualizar los datos utilizando AC, utilizaba distintos procedimientos para transformarlos. Los datos que veremos en este capítulo derivan de escalas de grados, de preferencias, de comparaciones por pares o de escalas continuas. En todos estos casos deberemos recordar el paradigma fundamental del AC: el AC analiza recuentos. Por tanto, si somos capaces de transformar los datos a algún tipo de recuento, entonces es probable que sea apropiado aplicar AC. Para comprobar que la transformación es apropiada, tendremos que confirmar que tienen sentido los conceptos de perfil, de masa y de distancia χ^2.

Contenido

Escalas de grados

En el capítulo 20 vimos una escala de grados típica, una escala de cinco puntos de acuerdo/desacuerdo que utilizamos en el ejemplo sobre ciencia y medio ambiente:

□	□	□	□	□
Muy de acuerdo	*Bastante de acuerdo*	*Ni de acuerdo ni en desacuerdo*	*Algo en desacuerdo*	*Muy en desacuerdo*

Analizamos estos datos como si se tratara de variables categóricas nominales. Con este fin creamos una variable binaria para cada categoría y para recodificar los datos. No aplicamos el AC a los datos expresados en la escala original de 1 a 5. En la escala original, el concepto de perfil no tendría sentido ya que el perfil de la respuesta [1 1 1 1] —muy de acuerdo con las cuatro afirmaciones— y el de la respuesta [5 5 5 5] —muy en desacuerdo con las cuatro afirmaciones— serían iguales. Otro tipo de escalas de grados que encontramos a menudo en ciencias sociales y en investigación de mercados son:

— Una escala de nueve puntos (añadimos una categoría extra entre los puntos de la escala de cinco puntos):

□	□	□	□	□	□	□	□	□
Muy de acuerdo		*Bastante de acuerdo*		*Ni de acuerdo ni en desacuerdo*		*Algo en desacuerdo*		*Muy en desacuerdo*

— Una escala de importancia de cuatro puntos:

□	□	□	□
Nada importante	*Bastante importante*	*Muy importante*	*Extremadamente importante*

— Una escala semántica diferencial de siete puntos en una encuesta sobre la satisfacción de los clientes:

Servicio antipático	□	□	□	□	□	□	□	*Servicio simpático*

— Una escala de grados continua (por ejemplo, una escala de 0 a 10):

Muy insatisfecho 0 ——————————————— 10 *Muy satisfecho*

En este último ejemplo el encuestado puede escoger cualquier valor entre 0 y 10, incluso, si quiere, con decimales. Sin embargo, seguimos considerando los datos como procedentes de una escala de grados. En consecuencia, llevaremos a cabo la recodificación de forma similar a la de los ejemplos anteriores. Fijémonos, sin embargo, en que cuando el número de puntos de las escalas es grande es poco manejable utilizar variables binarias para codificar el ACM.

Doblado de la escala de grados

El *doblado* es el procedimiento de recodificación que utilizamos habitualmente en AC para datos procedentes de escalas de grados. Lo que hacemos es redefinir las escalas de grados como un par de escalas complementarias; el polo «positivo» o «elevado», y el polo «negativo» o «bajo». Antes de llevar a cabo el doblado es recomendable que el extremo inferior de la escala de grados sea igual a cero. Así, por ejemplo,

Pregunta				*Pregunta A*		*Pregunta B*		*Pregunta C*		*Pregunta D*	
A	*B*	*C*	*D*	*A–*	*A+*	*B–*	*B+*	*C–*	*C+*	*D–*	*D+*
2	3	4	3	1	3	2	2	3	1	2	2
3	4	2	3	2	2	3	1	1	3	2	2
2	3	2	4	1	3	2	2	1	3	3	1
2	2	2	2	1	3	1	3	1	3	1	3
3	3	3	3	2	2	2	2	2	2	2	2
⋮	⋮	⋮	⋮		⋮		⋮		⋮		⋮
… y así para las 871 filas											

Imagen 23.1: *Datos originales correspondientes a las variables sobre la ciencia y el medio ambiente, y codificado doble, para los cinco primeros encuestados de $N = 871$ (muestra de Alemania Occidental)*

si trabajamos con las escalas de 1 a 5 y de 1 a 7, restaríamos 1 a los valores de las escalas para obtener las escalas de 0 a 4 y de 0 a 6, respectivamente. Estas nuevas escalas definen directamente el polo positivo. Damos por supuesto que los valores altos corresponden a valores sustantivamente altos de la escala (por ejemplo, mucha satisfacción, mucha importancia, muy de acuerdo). En las escalas de acuerdo/en desacuerdo que vimos anteriormente, los valores altos corresponden a un gran desacuerdo. Por tanto, en este caso, para evitar confusiones antes de proseguir el análisis tendríamos que haber invertido la escala. Obtenemos el polo negativo restando a M —el mayor valor del polo positivo (4, 6 u 8 para las escalas de grados, y 10 para la escala 0-10 que vimos anteriormente)— los valores de la escala del polo positivo. En la tabla de la imagen 23.1 hemos ilustrado este procedimiento para los datos sobre ciencia y medio ambiente que vimos en el capítulo 20. En la tabla de la imagen 23.1 mostramos las primeras 10 filas de datos y sus homólogos doblados. Por ejemplo, el primer valor del encuestado 1 es un 2, restando 1 obtenemos el valor 1, su valor doblado es 3. Por tanto, para la pregunta 1, aparecen los valores 1 y 3 en la primera columna de doblado que hemos etiquetado como *A*– y *A*+ para indicar que los valores calculados cuantifican el desacuerdo y el acuerdo, respectivamente, con relación a la primera pregunta. De manera similar, el valor original 3 de la segunda pregunta se convierte en un 2 y en un valor doblado de 2, es decir, valores iguales para los polos de desacuerdo y de acuerdo, *B*– y *B*+, y así sucesivamente.

Paradigma de recuento

Podemos considerar los valores doblados como si fueran recuentos. Efectivamente, los valores doblados 1 y 3 indican el número de puntos de la escala que quedan por debajo y por encima, respectivamente, del valor observado 1 (en la escala que empieza por cero). En la escala original, la respuesta 2 («bastante de acuerdo» tiene, en la escala, un punto por debajo [1] y tres por encima [3, 4 y 5]). De la misma forma, la respuesta 3, «ni de acuerdo ni en desacuerdo», se halla en el centro de la escala, ya que tiene dos puntos por encima y dos por debajo. Es decir, la tabla de datos doblados sustituye los datos originales midiendo la asociación entre cada encuestado y los polos de acuerdo y desacuerdo de la escala de grados. Es necesario medir esta asociación con ambos polos: si utilizáramos los valores de un

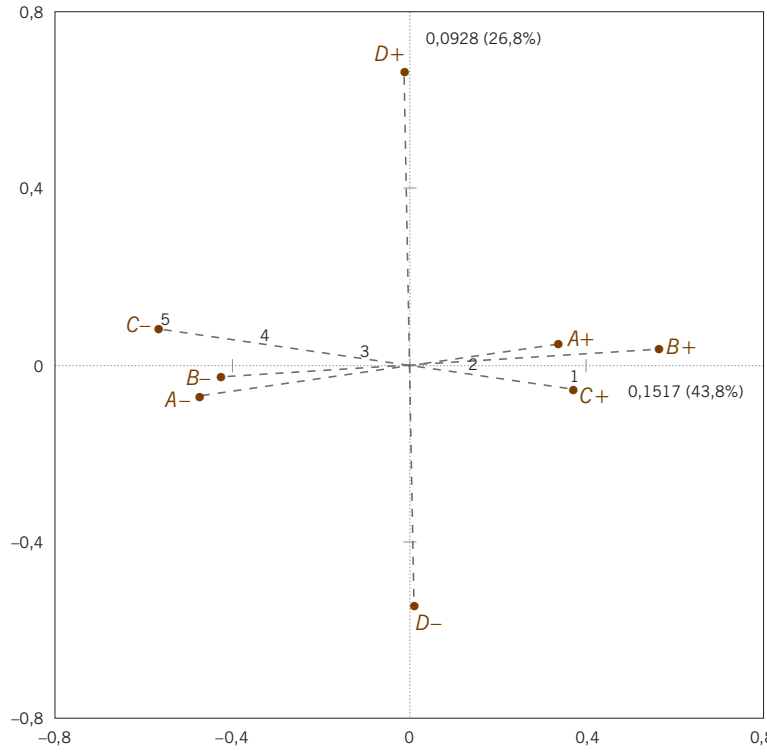


Imagen 23.2: *AC correspondiente al codificado doble de los datos sobre* ciencia *y* medio ambiente, *que muestra sólo los valores derivados del codificado doble. El porcentaje de inercia explicada es del 70,6%. En cada uno de los ejes, podemos imaginar la escala de rangos a intervalos iguales, conectando los polos (es decir, la escala de 1 a 5, de la pregunta C). La media de cada pregunta se halla exactamente en el origen*

solo polo, los perfiles del AC no tendrían sentido ya que, por ejemplo, estar muy de acuerdo o muy en desacuerdo con todas las preguntas tendrían los mismos perfiles y, por tanto, la misma posición en el mapa.

Mapa del AC con la escala de grados doble

Apliquemos el AC a la tabla de 8 columnas y 871 filas situada a la derecha de la imagen 23.1. Todas las filas dan la misma suma (16 en este ejemplo). Por tanto, las masas de los encuestados (filas) son iguales y, efectivamente, no tiene porqué haber diferencias en los pesos de los encuestados. Los cuatro pares de columnas dan la misma suma, por tanto, hay cuatro restricciones lineales en las columnas y no solamente una como en el AC habitual. En consecuencia, la dimensionalidad de la matriz de datos es 8 – 4 = 4. La inercia total y su descomposición en los cuatro ejes principales es la siguiente:

$$0{,}3462 = 0{,}1517\ (43{,}8\%) + 0{,}0928\ (26{,}8\%) + 0{,}0529\ (15{,}3\%) + 0{,}0488\ (14{,}1\%)$$

En el mapa de la imagen 23.2 hemos representado las columnas en coordenadas principales. Para cada pregunta tenemos dos puntos. Como muestran las líneas de trazo discontinuo, los polos positivos se hallan, con relación al origen, opuestos a sus homólogos negativos. Vemos claramente que la pregunta *D* se halla fuera de las alineaciones que muestran las otras tres preguntas. Lo mismo que vimos

en el capítulo 20. Quizás habríamos esperado que *D*– estuviera a la derecha y que *D*+ estuviera a la izquierda; en cualquier caso estas variables forman casi un ángulo recto con las restantes.

Los ejes de la escala de grados tienen el origen en la media

Los cuatro «ejes» de las escalas de grados pasan por el origen del mapa. Para recuperar la escala original, podemos subdividir las líneas discontinuas entre polos en cuatro intervalos iguales, y etiquetar los cinco puntos resultantes, como mostramos para la pregunta *C* utilizando las etiquetas de la escala original de 1 a 5 (1 corresponde a «muy de acuerdo»). Para todas las preguntas se cumple que el origen del mapa corresponde a la media de los valores de las respuestas en sus correspondientes escalas. Así, podemos ver en el mapa que las medias de las respuestas a las preguntas *A* y *C* se hallan más en el lado de acuerdo (+) de la escala de grados (para la pregunta *C* la media es 2,58), mientras que las medias de las preguntas *B* y *D* se hallan ligeramente en el lado de desacuerdo. Otra forma de verlo sería imaginar que los pesos de los puntos extremos de cada eje de escala de grados es proporcional a la media de los valores de su polo (de esta manera *C*+ se halla más próximo al origen que *C*–, porque es «más pesado»).

Correlaciones aproximadas por los cosenos de los ángulos

Los cosenos de los ángulos formados por los cuatro ejes del mapa de la imagen 23.2 son, aproximadamente, las correlaciones entre variables. Por tanto, las variables *A*, *B* y *C* estarán correlacionadas positivamente ente sí, mientras que no lo están con *D*. Los coeficientes de correlación de las cuatro variables son los siguientes:

Preguntas	*A*	*B*	*C*	*D*
A	1	0,378	0,357	0,036
B	0,378	1	0,436	0,016
C	0,357	0,436	1	–0,062
D	0,036	0,016	–0,062	1

Ello concuerda con nuestra deducción visual. Debido a que el mapa sólo explica el 70% de la inercia de los datos, no hemos recuperado de forma exacta los valores de las correlaciones. Por ejemplo, *B* y *C* deberían formar un ángulo más pequeño que *A* y *B*. Lo veríamos de forma más precisa en una representación tridimensional de los ejes de las escalas de grados.

Posiciones de filas y puntos adicionales

Igual que en el AC habitual, cada encuestado tiene un perfil y, por tanto, una posición en el mapa. No obstante, como ocurre en el ACM con datos procedentes de encuestas de grandes muestras, para nosotros tiene más interés representar como puntos adicionales grupos de individuos que las posiciones de cada individuo. Para ilustrarlo consideremos, para los datos anteriores, los datos clasificados en hombres y mujeres de seis grupos de edad, es decir 12 grupos. Podríamos calcular los valores medios de los mencionados grupos en la escala de grados. Lue-

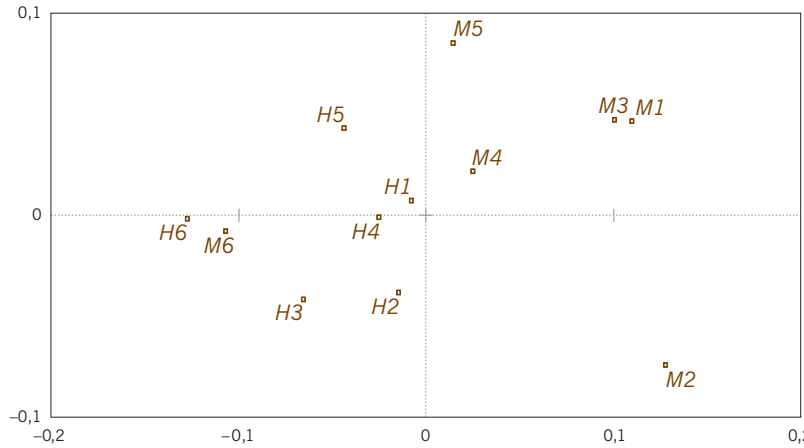


Imagen 23.3: *Puntos adicionales correspondientes a hombres y mujeres de cinco grupos de edad. Todas las mujeres se hallan en el lado derecho (acuerdo), mientras que los hombres —con excepción del grupo de mayor edad— se hallan en el lado de desacuerdo*

go los añadimos en el mapa como filas adicionales (dobladas). En el mapa de la imagen 23.3 mostramos sus posiciones. Vemos que todos los grupos de mujeres se sitúan a la derecha del mapa. Es decir, en el lado de acuerdo de las preguntas *A*, *B* y *C*. A excepción del grupo de hombres de mayor edad, los grupos de hombres se hallan en el lado de desacuerdo de estas preguntas, es decir, son menos críticos con el papel de la ciencia en el medio ambiente.

Datos de preferencias

Podríamos considerar los datos de preferencias como un caso especial de datos en escalas de grados y así visualizarlos en AC como estos últimos. En investigación de mercados es habitual pedir a los encuestados que ordenen productos de más a menos según su preferencia, o que ordenen atributos de los productos de más a menos importantes. Así, supongamos, por ejemplo, que tenemos seis productos, de *A* a *F*, y que un determinado individuo los ordena de la siguiente manera:

más preferidos : $B > E > A > C > F > D$: menos preferidos

De acuerdo con esta ordenación, a cada uno de los seis productos le corresponde los siguientes rangos:

A	*B*	*C*	*D*	*E*	*F*
3	1	4	6	2	5

Podemos considerar estos rangos como derivados de una escala de grados de seis puntos. La diferencia con la escala de rangos habitual es que los encuestados se han visto forzados a utilizar una sola vez cada uno de los valores de la escala. Sin embargo, los podemos doblar de la manera habitual. En las etiquetas asignadas a las columnas dobles, + indica mucha preferencia, y – muy poca preferencia:

A–	*A*+	*B*–	*B*+	*C*–	*C*+	*D*–	*D*+	*E*–	*E*+	*F*–	*F*+
2	3	0	5	3	2	5	0	1	4	4	1

Imagen 23.4: *Indicadores económicos de la Unión Europea y sus rangos del menor a mayor*

	Datos originales					*Rangos de los datos*				
Países	*TD*	*PIB*	*CI*	*CCI*	*CLUR*	*TD*	*PIB*	*CI*	*CCI*	*CLUR*
Bélgica	8,8	102	104,9	3,3	89,7	7	7	7	7,5	5,5
Dinamarca	7,6	134,4	117,1	1	92,4	5	12	11	1	8
Alemania	5,4	128,1	126	3	90	3	11	12	6	7
Grecia	8,5	37,7	40,5	2	105,6	6	2	2	2	12
España	16,5	67,1	68,7	4	86,2	12	4	4	11	3
Francia	9,1	112,4	110,1	2,8	89,7	8	9	9	4,5	5,5
Irlanda	16,2	64	60,1	4,5	81,9	11	3	3	12	2
Italia	10,6	105,8	106	3,8	97,4	10	8	8	10	10
Luxemburgo	1,7	119,5	110,7	2,8	95,9	1	10	10	4,5	9
Países Bajos	9,6	99,6	96,7	3,3	86,6	9	6	5	7,5	4
Portugal	5,2	32,6	34,8	3,5	78,3	2	1	1	9	1
Reino Unido	6,5	95,3	99,7	2,1	98,9	4	5	6	3	11

TD: tasa de desempleo (%); PIB: producto interior bruto per cápita (índice); CI: consumo individual (índice); CCI: cambio en el consumo individual (%); CLUR: costes laborales unitarios reales

De todas formas, es habitual que los encuestados puedan ordenar sólo los objetos más preferidos (por ejemplo, los tres más preferidos). En tal caso, asignamos el mismo rango a todos los objetos no seleccionados (la media de los rangos de las posiciones no seleccionadas). Por ejemplo, si ordenáramos sólo los tres mejores de seis productos, asignaríamos a los tres productos omitidos el rango 5, la media de 4, 5 y 6.

Comparaciones por pares

Las *comparaciones por pares* son un tipo de ordenación más libre que la ordenación por preferencias. Por ejemplo, supongamos que presentamos a los encuestados los 15 pares posibles de seis productos, de *A* a *F*, y le pedimos que elija los pares preferidos. Haríamos el doblado de las respuestas de los encuestados de la siguiente manera:

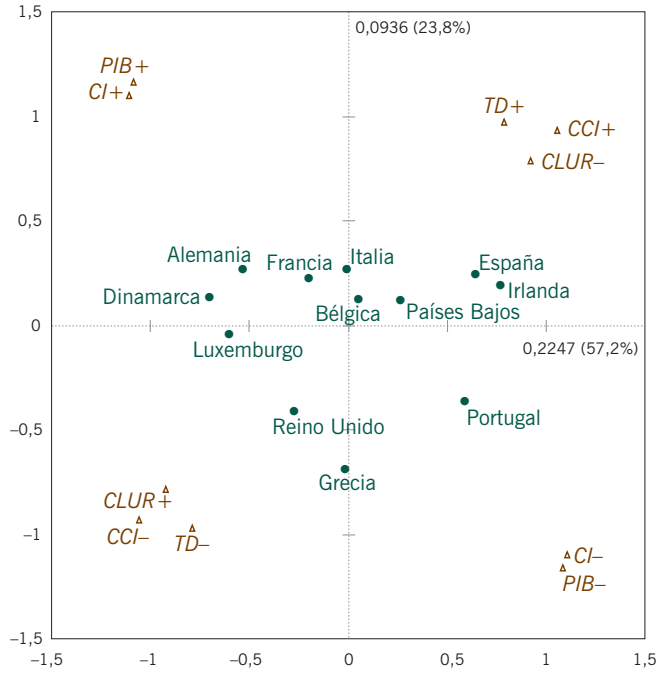
A+: número de veces que se ha preferido A a los restantes productos
A–: número de veces que se han preferido los restantes productos y no *A* (= 5 – *A*+),

y así sucesivamente con los demás productos. Luego, igual que antes, aplicamos el AC a los datos doblados.

Conjunto de datos 13: indicadores de la Unión Europea

Mediante AC, también podemos visualizar datos procedentes de escalas continuas, si previamente los recodificamos convenientemente. Con este fin, existen distintas posibilidades. Como ejemplo, consideremos los datos situados a la izquierda de la tabla de la imagen 23.4. Se trata de cinco indicadores económicos de 12 países de la Unión Europea a principios de los años noventa, que se han expresado en distintas escalas. Así, por ejemplo, la tasa de desempleo y el cambio en el consumo individual se expresan como porcentajes.

Imagen 23.5:
Mapa asimétrico del AC correspondiente a la recodificación, mediante rangos, de los indicadores de la Unión Europea. La inercia explicada es del 81,0%



Recodificación de datos continuos, ordenación y doblado

Para recodificar estos datos, primero expresaremos los valores de cada variable como rangos tal como se muestra en la tabla de la imagen 23.4. Así, por ejemplo, vemos que Luxemburgo tiene el desempleo más bajo y que, por tanto, le corresponde el rango 1, luego Portugal con el rango 2, etc. A los valores iguales les asignaremos también rangos iguales, o sea, la media de sus rangos. Por ejemplo, Francia y Luxemburgo tienen el mismo valor para el *CCI*, están empatados ocupando las posiciones 4 y 5. Les asignaremos como rango la media de 4 y 5, es decir 4,5. Una vez expresados los valores de cada variable como rangos, podemos llevar a cabo el doblado tal como hicimos anteriormente. Es decir, en primer lugar, para obtener el polo positivo restamos 1 a los valores de los rangos. Luego obtenemos el polo negativo como: 11 menos el valor del polo positivo. En la imagen 23.5, mostramos el mapa del AC de la matriz doblada. De nuevo, hemos unido los polos opuestos de cada variable. En este caso observamos que las distancias de los polos al origen son iguales. Ello se debe a que las medias de sus rangos son idénticas (por tanto, bastaría con dibujar el polo positivo). En el mapa podemos ver dos grupos de variables muy poco correlacionadas, pero muy correlacionadas dentro de cada grupo. Así, por ejemplo, fijémonos en que *CLUR* (costes laborales unitarios reales) están muy negativamente correlacionadas con *TD* (tasa de desempleo) y con *CCI* (cambio consumo individual); como estamos utilizando rangos, cuando hablamos de correlaciones, nos referimos a la *correlación por rangos de Spearman*. La posición de cada país depende de los valores de los rangos de cada variable y no de su valor concreto. Por tanto, dado que analizamos

rangos, y no valores originales, el análisis será robusto con relación a las observaciones atípicas. Así, pues, estamos llevando a cabo un AC *no paramétrico*.

Otras posibilidades de recodificación para datos continuos

La expresión de variables continuas como rangos conlleva la pérdida de algo de información. Sin embargo, nuestra experiencia nos muestra que esta pérdida es mínima por lo que respecta a la visualización de los datos. Por el contrario, en muchas situaciones, la robustez de los rangos es una ventaja. De todas formas, si necesitamos toda la información contenida en los datos, existen otras posibilidades. Por ejemplo, una transformación adecuada consiste en estandarizar todas las variables (puntuaciones z). Es decir, restamos a cada variable su media y esta diferencia la dividimos por su desviación típica. A continuación, a partir de z creamos dos versiones de cada variable utilizando la remodificación siguiente:

$$\text{valor positivo} = \frac{1+z}{2} \qquad \text{valor negativo} = \frac{1-z}{2} \tag{23.1}$$

A pesar de que obtenemos algunos valores negativos, los valores marginales de filas y columnas se mantienen positivos, e iguales para todas las filas y para todos los pares de columnas dobladas. Por tanto, la ponderación es igual para todos los casos y todas las variables. El AC de esta matriz doblada proporciona un mapa casi idéntico al mapa de la imagen 23.5. Como curiosidad, señalar que hasta donde llega nuestro conocimiento, se trata del único caso de matriz de datos con algunos valores negativos que podemos analizar de forma válida utilizando el AC.

RESUMEN: Recodificación de datos

1. Podemos recodificar datos procedentes de diferentes escalas de medida, de manera que sean adecuados para el AC.
2. Siempre que la matriz recodificada posea perfiles y sumas marginales que tengan sentido en el contexto de la aplicación, el AC proporcionará una visualización correcta de los datos.
3. Uno de los principales procedimientos de recodificación consiste en *doblar* las variables. Es decir, convertir cada variable en un par de variables para que la suma de los pares de variables sea constante.
4. Podemos llevar a cabo el doblado en el caso de escalas de grados, de preferencias y de comparaciones por pares de objetos. Obtenemos mapas en los que cada variable queda representada por dos puntos opuestos con relación al origen. En el caso particular de las variables expresadas en escalas de grados, el origen del mapa indica el valor medio de la variable que estemos considerando en relación con la escala delimitada por los dos polos extremos.
5. Podemos recodificar datos continuos como rangos dobles, lo que nos conduce a una forma no paramétrica del AC. Otra posibilidad es transformarlos en una par de variables continuas, a partir de sus valores estandarizados.

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.

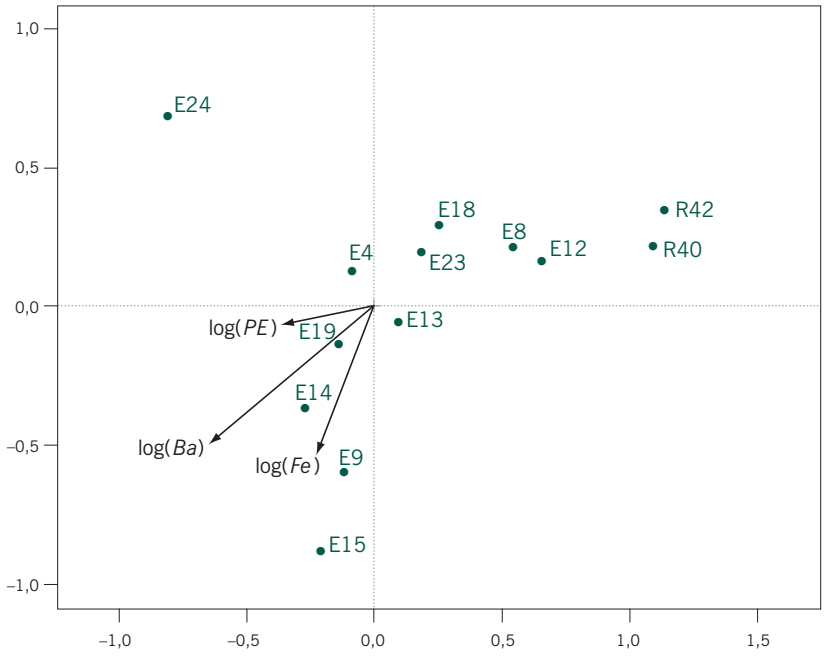


Imagen 24.2:
Mapa de las estaciones de la imagen 10.5, que muestra las posiciones de las tres variables ambientales externas como puntos adicionales, de acuerdo con sus correlaciones con los dos ejes principales

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).

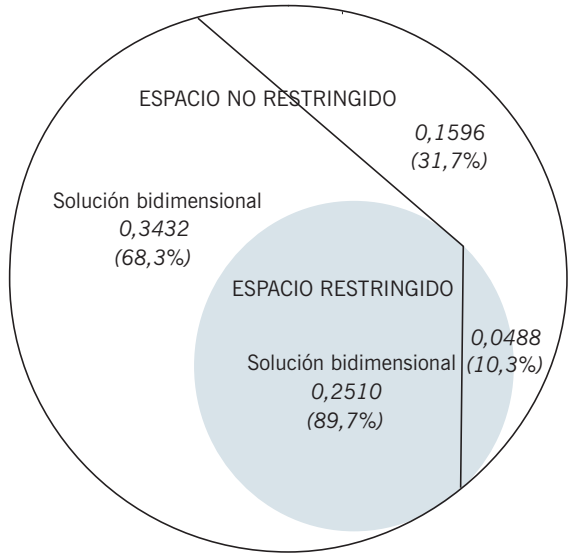
Espacios restringidos y no restringidos en ACC

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-

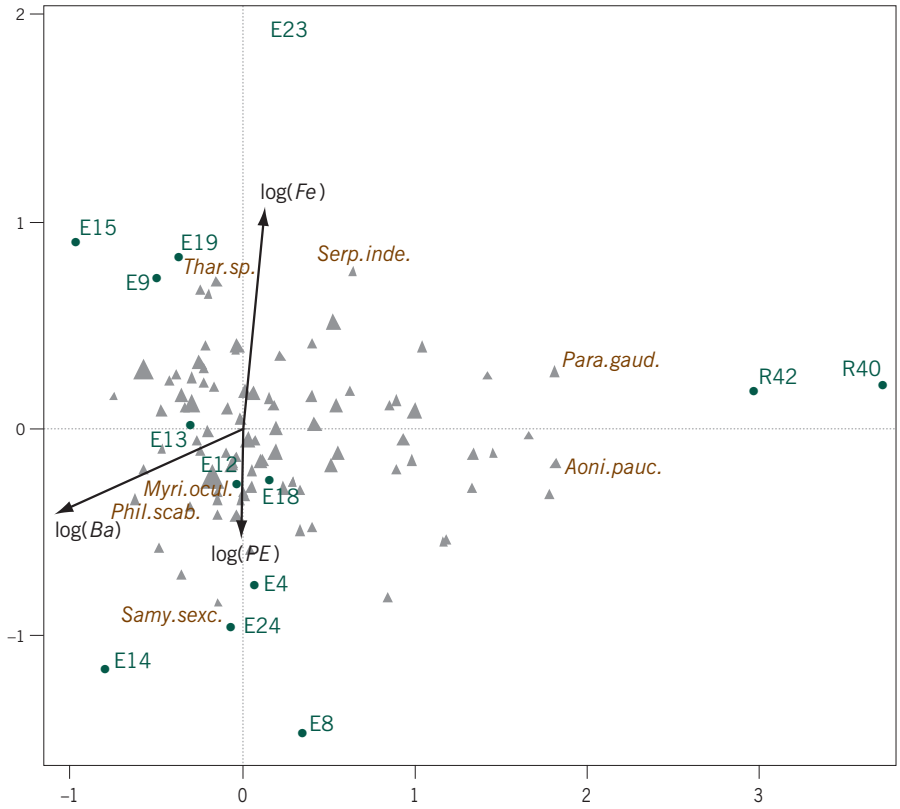


Imagen 24.6:
Triplot del ACC en el que hemos representado las especies (filas) y las localidades (columnas) en un mapa asimétrico de filas (es decir, las localidades en coordenadas estándares). Hemos situado las variables ambientales según los valores de sus coeficientes en las relaciones lineales con 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.

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

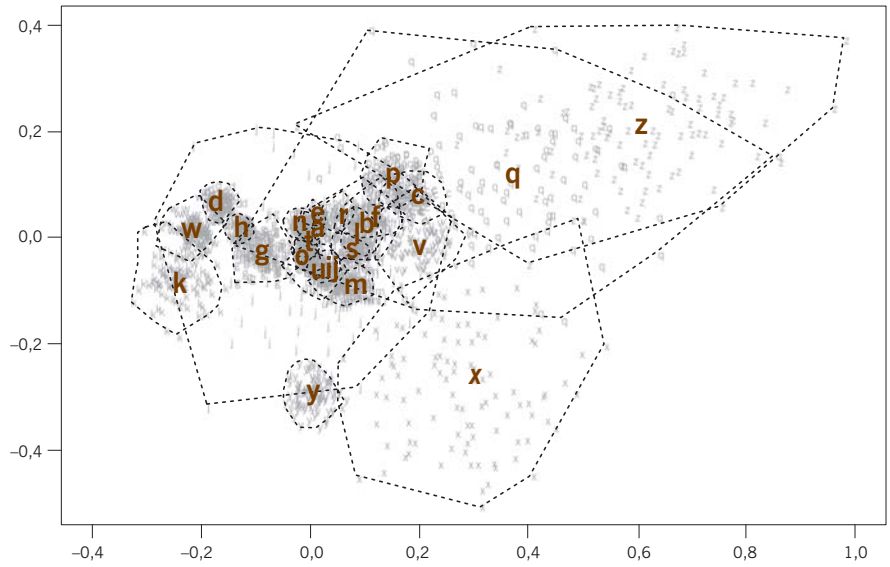
El *muestreo multinomial* es una manera equivalente de contemplar (y de llevar a cabo) el remuestreo. Se basa en que cada perfil fila define un conjunto de frecuencias que podemos considerar como las probabilidades de obtener, en cada texto, una `a`, una `b`, una `c`, etc. Por tanto, se trata de muestrear 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

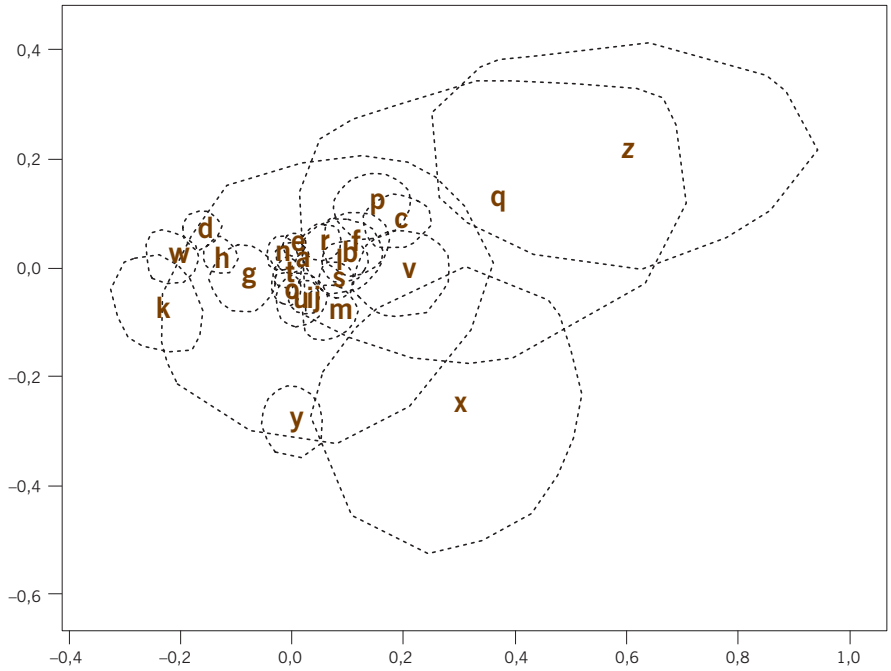


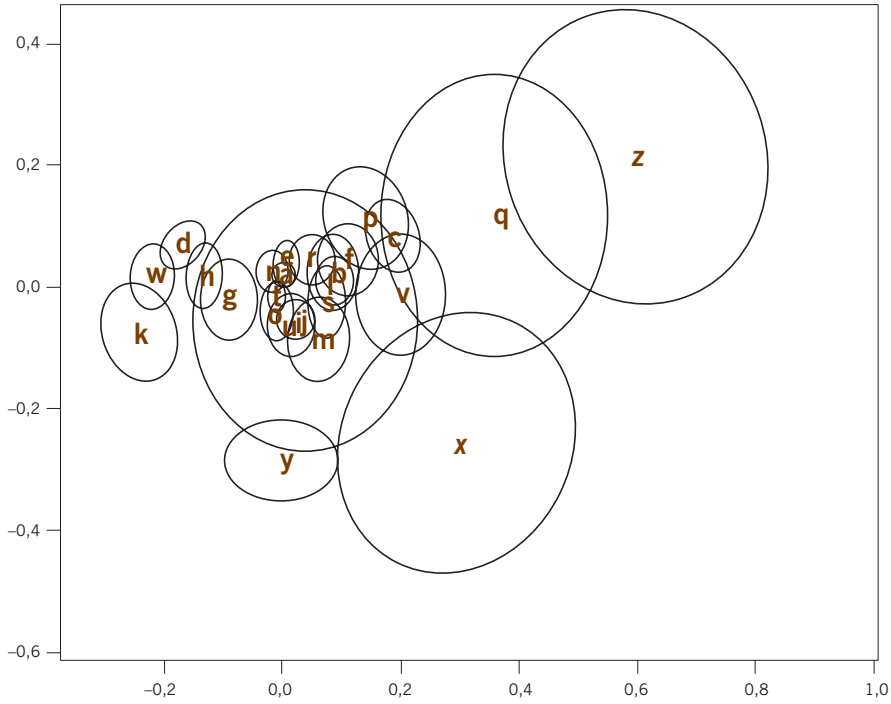
Imagen 25.2:
Recorte de perímetros convexos de puntos obtenidos de 1000 réplicas (10 veces más que en la imagen 25.1) que muestran, para sus distribuciones, regiones de confianza aproximadas al 95%

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

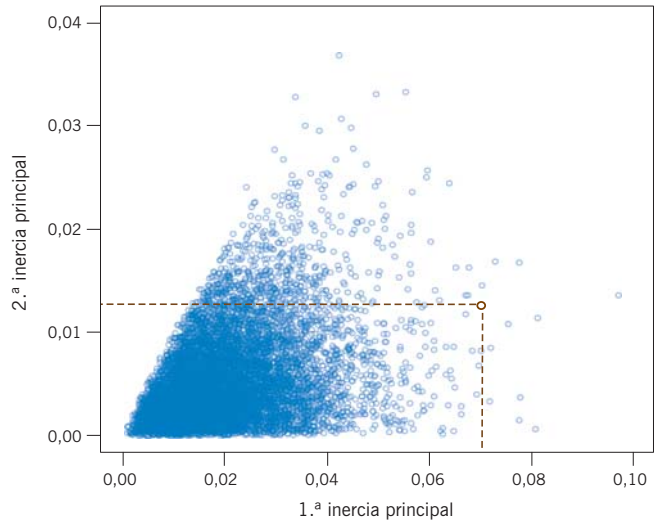


Imagen 25.5: *Diagrama de dispersión de las inercias principales del AC original y de las 9999 simulaciones de la tabla de contingencia de 5 × 3, bajo la hipótesis nula de que no existe asociación entre filas y columnas (en la imagen 25.4, se muestran dos de estas simulaciones). Las inercias principales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

APÉNDICE A

Teoría del análisis de correspondencias

El análisis de correspondencias se basa en resultados directos de la teoría de matrices. Utiliza especialmente la descomposición de una matriz en valores singulares (DVS), el fundamento de muchos métodos multivariantes, como el análisis de componentes principales, el análisis de correlaciones canónicas, todas las variantes de los biplots lineales, el análisis discriminante y el escalado métrico multidimensional. En este apéndice resumimos la teoría del análisis de correspondencias, así como la de los métodos relacionados que hemos visto en este libro. Hemos preferido utilizar la notación matricial porque es más concisa, y también porque se halla más próxima a la implementación del método en el lenguaje de cálculo R.

Contenido

Matriz de correspondencias y notación preliminar

Sea **N** una matriz de datos $I \times J$, con sumas positivas de filas y columnas (casi siempre **N** consta de valores no negativos, sin embargo, existen algunas excepciones como la que describimos al final del capítulo 23). Para simplificar la notación, en primer lugar, transformamos la matriz **N** en la *matriz de correspondencias* **P**, dividiendo **N** por la suma total de sus elementos $n = \sum_i \sum_j n_{ij} = \mathbf{1}^\top \mathbf{N} \mathbf{1}$ (utilizamos **1** para simbolizar un vector de unos, con una longitud adecuada a su uso. El primer **1** es $I \times 1$, el segundo es $J \times 1$ para coincidir con el número de filas y de columnas de **N**).

Matriz de correspondencias:

$$\mathbf{P} = \frac{1}{n}\mathbf{N} \tag{A.1}$$

Hemos utilizado la siguiente notación:

Masas de filas y columnas:

$$\begin{aligned} r_i = \textstyle\sum_{j=1}^{J} p_{ij} \qquad & c_j = \textstyle\sum_{i=1}^{I} p_{ij} \\ \text{es decir,}\quad \mathbf{r} = \mathbf{P}\mathbf{1} \qquad & \mathbf{c} = \mathbf{P}^\top \mathbf{1} \end{aligned} \tag{A.2}$$

Matrices diagonal de masas de filas y de columnas:

$$\mathbf{D}_r = \mathrm{diag}(\mathbf{r}) \qquad \text{y} \qquad \mathbf{D}_c = \mathrm{diag}(\mathbf{c}) \tag{A.3}$$

A partir de ahora, expresaremos todas las definiciones y todos los resultados en términos de estos valores relativos $\mathbf{P} = \{p_{ij}\}$, $\mathbf{r} = \{r_i\}$ y $\mathbf{c} = \{c_j\}$, cuyos elementos, en todos los casos, suman 1. Multiplicando por n recuperamos los elementos de la matriz original **N**: $np_{ij} = n_{ij}$, nr_i = suma de la i-ésima fila de **N**, nc_j = suma de la j-ésima columna de **N**.

Algoritmo básico de cálculo

El algoritmo de cálculo para obtener las coordenadas de los perfiles fila y de los perfiles columna en relación con los ejes principales, utilizando la descomposición en valores singulares (DVS), es el siguiente:

Paso 1 del AC. Cálculo de la matriz **S** *de residuos estandarizados:*

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)\mathbf{D}_c^{-\frac{1}{2}} \tag{A.4}$$

Paso 2 del AC. Cálculo de la DVS de **S***:*

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\top \quad \text{donde} \quad \mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I} \tag{A.5}$$

donde $\mathbf{D}_\alpha$ es la matriz diagonal de valores singulares (positivos) en orden descendente: $\alpha_1 \geq \alpha_2 \geq \ldots$

Paso 3 del AC. Coordenadas estándares de las filas Φ*:*

$$\Phi = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U} \tag{A.6}$$

Paso 4 del AC. Coordenadas estándares de las columnas Γ*:*

$$\Gamma = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V} \tag{A.7}$$

Paso 5 del AC. Coordenadas principales de las filas $\mathbf{F}$*:*

$$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\alpha = \Phi\mathbf{D}_\alpha \tag{A.8}$$

Paso 6 del AC. Coordenadas principales de las columnas $\mathbf{G}$*:*

$$\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\alpha = \Gamma\mathbf{D}_\alpha \tag{A.9}$$

Paso 7 del AC. Inercias principales λ_k*:*

$$\lambda_k = \alpha_k^2, \quad k = 1,2,\ldots,K \quad \text{donde} \quad K = \min\{I-1, J-1\} \tag{A.10}$$

Las filas de las matrices de coordenadas en (A.6)–(A.9) hacen referencia a las filas o a las columnas, según cada caso, de la tabla original. Mientras que las columnas de estas matrices hacen referencia a los ejes principales, o dimensiones, de las cuales hay $\min\{I-1, J-1\}$. Es decir, uno menos el número de filas o de columnas, el que sea menor. Fijémonos en el cálculo de las escalas de las coordenadas principales y estándares:

$$\mathbf{F}\mathbf{D}_r\mathbf{F}^\top = \mathbf{G}\mathbf{D}_c\mathbf{G}^\top = \mathbf{D}_\lambda \tag{A.11}$$

$$\Phi\mathbf{D}_r\Phi^\top = \Gamma\mathbf{D}_c\Gamma^\top = \mathbf{I} \tag{A.12}$$

Es decir, la suma de cuadrados ponderada de las coordenadas principales en la *k*-ésima dimensión (es decir, su inercia en la dirección de esta dimensión) es igual a la inercia principal (o valor propio) $\lambda_k = \alpha_k^2$, el cuadrado del *k*-ésimo valor singular. Mientras que la suma de cuadrados ponderada de las coordenadas estandarizadas es igual a 1. Todas las matrices de coordenadas tienen columnas ortogonales, en las que siempre utilizamos las masas para el cálculo de los productos escalares (ponderados).

Un apunte sobre la descomposición en valores singulares (DVS)

La DVS es un resultado matemático fundamental del AC, de la misma manera que lo es para otras técnicas de reducción de la dimensión, como el análisis de componentes principales, el análisis de correlaciones canónicas, y el análisis discriminante lineal. Esta descomposición expresa cualquier matriz rectangular como el producto de tres matrices de estructura simple, como vimos anteriormente en (A.5): $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\top$. Las columnas de las matrices $\mathbf{U}$ y $\mathbf{V}$ son, respectiva-

mente, los *vectores singulares izquierdo* y *derecho*. Los valores positivos α_k de la diagonal de $\mathbf{D}_\alpha$ son, en orden descendente, los *valores singulares*. Si $\mathbf{SS}^\mathsf{T}$ y $\mathbf{S}^\mathsf{T}\mathbf{S}$ son matrices cuadradas simétricas, la DVS está relacionada con la descomposición en *valores y vectores propios* de una matriz cuadrada simétrica de la siguiente manera $\mathbf{SS}^\mathsf{T} = \mathbf{UD}_\alpha^2\mathbf{U}^\mathsf{T}$ y $\mathbf{S}^\mathsf{T}\mathbf{S} = \mathbf{VD}_\alpha^2\mathbf{V}^\mathsf{T}$. Por tanto, los vectores singulares son también vectores propios de las respectivas matrices, y los valores singulares son las raíces cuadradas de sus valores propios. La utilidad práctica de la DVS es que podemos construir otra matriz $I \times J$ $\mathbf{S}_{(m)}$ a partir de las primeras m columnas $\mathbf{U}_{(m)}$ y $\mathbf{V}_{(m)}$, y los primeros m valores singulares $\mathbf{D}_{\alpha(m)}$ como: $\mathbf{S}_{(m)} = \mathbf{U}_{(m)}\mathbf{D}_{\alpha(m)}\mathbf{V}_{(m)}^\mathsf{T}$. Por lo tanto, $\mathbf{S}_{(m)}$ es la aproximación mínimos cuadrados de rango m a $\mathbf{S}$ (*teorema de Eckart-Young*). Dado que el objetivo de hallar los subespacios de pocas dimensiones que mejor se ajusten coincide con el objetivo de hallar matrices de rango pequeño por mínimos cuadrados, la DVS resuelve de forma satisfactoria y muy compacta nuestro problema. La única adaptación necesaria es incorporar en la DVS la ponderación de filas y de columnas con las masas de manera que obtengamos aproximaciones por mínimos cuadrados ponderados. Si definimos una forma generalizada de la DVS, con valores singulares normalizados y ponderados con las masas, entonces podremos obtener directamente el resultado del AC. Por ejemplo, la DVS generalizada de los *cocientes de contingencia* $p_{ij} / (r_i c_j)$, los elementos de la matriz $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}$, centrada en el valor constante 1, nos conduce directamente a las coordenadas estándares de filas y de columnas:

$$\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}^\mathsf{T} = \Phi\mathbf{D}_\alpha\Gamma^\mathsf{T} \quad \text{donde} \quad \Phi^\mathsf{T}\mathbf{D}_r\Phi = \Gamma^\mathsf{T}\mathbf{D}_c\Gamma = \mathbf{I} \tag{A.13}$$

El modelo bilineal del AC

Los pasos 1 a 4 del algoritmo básico nos permiten escribir los datos de $\mathbf{P}$ de la siguiente manera [véanse también (13.4), pág. 139, y (14.9), pág. 150]:

$$p_{ij} = r_i c_j \left(1 + \sum_{k=1}^{K} \sqrt{\lambda_k}\phi_{ik}\gamma_{jk}\right) \tag{A.14}$$

(*fórmula de reconstitución*). En notación matricial:

$$\mathbf{P} = \mathbf{D}_r(\mathbf{1}\mathbf{1}^\mathsf{T} + \Phi\mathbf{D}_\lambda^{1/2}\Gamma^\mathsf{T})\mathbf{D}_c \tag{A.15}$$

Dadas las relaciones simples (A.8) y (A.9), entre las coordenadas principales y las coordenadas estándares, podemos escribir este modelo bilineal de distintas maneras [véanse también (14.10) y (14.11) en la pág. 150].

Ecuaciones de transición entre filas y columnas

Los vectores singulares izquierdo y derecho están relacionados linealmente. Por ejemplo, multiplicando la DVS de la derecha por $\mathbf{V}$: $\mathbf{SV} = \mathbf{UD}_\alpha$. Las variaciones consistentes en expresar estas relaciones en términos de coordenadas principales y de coordenadas estándares proporcionan las ecuaciones de transición [véanse (14.1) y (14.2), así como (14.5) y (14.6) para las ecuaciones escalares equivalentes]:

Coordenadas principales en función de las coordenadas estándares (relaciones baricéntricas):

$$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{P}\boldsymbol{\Gamma} \qquad \mathbf{G} = \mathbf{D}_c^{-1}\mathbf{P}^\top\boldsymbol{\Phi} \tag{A.16}$$

Coordenadas principales en función de las coordenadas principales:

$$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{G}\mathbf{D}_\lambda^{-1/2} \qquad \mathbf{G} = \mathbf{D}_c^{-1}\mathbf{P}^\top\mathbf{F}\mathbf{D}_\lambda^{-1/2} \tag{A.17}$$

Las ecuaciones (A.16) son las que mencionamos en el capítulo 3. Expresan los perfiles como medias ponderadas de los vértices, en las que los pesos son los elementos del perfil. Son las ecuaciones que configuran los *mapas asimétricos*. Las ecuaciones (A.17) muestran que los dos conjuntos de coordenadas principales, que configuran los *mapas simétricos*, también están relacionados por una relación baricéntrica (media ponderada), pero con factores de escala diferentes en cada eje (las inversas de las raíces cuadradas de las inercias principales).

Puntos adicionales

Utilizamos las ecuaciones de transición para situar puntos adicionales en el mapa. Por ejemplo, dada una columna adicional con valores en $\mathbf{h}$ $(I \times 1)$, la dividimos por su total $\mathbf{1}^\top\mathbf{h}$ para obtener el perfil columna $\tilde{\mathbf{h}} = (1/\mathbf{1}^\top\mathbf{h})\mathbf{h}$. A continuación transponemos el perfil para obtener el vector fila en la segunda ecuación de (A.16). Por ejemplo, para calcular las coordenadas $\mathbf{g}$ de la columna adicional:

$$\mathbf{g} = \tilde{\mathbf{h}}^\top\boldsymbol{\Phi} \tag{A.18}$$

Inercia total y distancias χ^2

La inercia total de la matriz de datos es igual a la suma de los cuadrados de la matriz $\mathbf{S}$ de (A.4):

$$\text{inercia} = \text{traza}(\mathbf{S}\mathbf{S}^\top) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{A.19}$$

La inercia también es la suma de los cuadrados de los valores singulares, es decir, la suma de los valores propios:

$$\text{inercia} = \sum_{k=1}^{K}\alpha_k^2 = \sum_{k=1}^{K}\lambda_k \tag{A.20}$$

Las distancias χ^2 entre los perfiles fila y los perfiles columna son:

$$\text{Las distancias } \chi^2 \text{ entre las filas } i \text{ e } i' \text{:} \quad \sum_{j=1}^{J}\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right)^2 / c_j \tag{A.21}$$

$$\text{Las distancias } \chi^2 \text{ entre las columnas } j \text{ y } j' \text{:} \quad \sum_{i=1}^{I}\left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}}\right)^2 / r_i \tag{A.22}$$

Escribir el conjunto completo de distancias χ^2 en forma de matriz cuadrada simétrica exige un poco más de trabajo. En primer lugar, calculamos la matriz **A** de «productos escalares χ^2» entre los perfiles fila, por ejemplo como:

$$\text{Productos escalares } \chi^2 \text{ entre filas:} \quad \mathbf{A} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}\mathbf{P}^\top\mathbf{D}_r^{-1} \tag{A.23}$$

A continuación, definimos el vector **a** como los elementos de la diagonal de esta matriz (es decir, los productos escalares de los perfiles fila por ellos mismos):

$$\mathbf{a} = \text{diag}(\mathbf{A}) \tag{A.24}$$

Entonces, la matriz, $I \times I$, de distancias χ^2 al cuadrado es:

$$\text{Matriz de distancias } \chi^2 \text{ al cuadrado entre filas:} \quad \mathbf{a}\mathbf{1}^\top + \mathbf{1}\mathbf{a}^\top - 2\mathbf{A} \tag{A.25}$$

Para calcular la matriz $J \times J$ de distancias χ^2 al cuadrado entre perfiles columna, intercambiamos las filas por las columnas en (A.23). Definimos **A** como $\mathbf{D}_c^{-1}\mathbf{P}^\top\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}$ y a continuación seguimos con (A.24) y (A.25).

Contribuciones de filas y columnas a las inercias principales

Las contribuciones de filas y de columnas a la inercia de la k-ésima dimensión son los componentes de la inercia:

$$\text{Para la fila } i\text{:} \quad \frac{r_i f_{ik}^2}{\lambda_k} = r_i \phi_{ik}^2 \qquad \text{para la columna } j\text{:} \quad \frac{c_j g_{jk}^2}{\lambda_k} = c_j \gamma_{jk}^2 \tag{A.26}$$

recordando la relación entre coordenadas principales y coordenadas estándares dadas en (A.8) y (A.9): $f_{ik} = \sqrt{\lambda_k}\phi_{ik}$, $g_{jk} = \sqrt{\lambda_k}\gamma_{jk}$. [Fijémonos en que las raíces cuadradas de los valores de (A.26) son exactamente las coordenadas propuestas para el biplot estándar de AC del capítulo 13, que muestra que las raíces cuadradas de estas coordenadas son las contribuciones a los ejes principales.]

Contribuciones de los ejes principales a las inercias de los puntos (correlaciones al cuadrado)

Las contribuciones de las dimensiones de la inercia a la i-ésima fila y a la j-ésima columna (es decir, los cosenos al cuadrado o las correlaciones al cuadrado) son:

$$\text{Para la fila } i\text{:} \quad \frac{f_{ik}^2}{\sum_k f_{ik}^2} \qquad \text{para la columna } j\text{:} \quad \frac{g_{jk}^2}{\sum_k g_{jk}^2} \tag{A.27}$$

Como vimos en el capítulo 11, los denominadores de (A.27) son los cuadrados de las distancias χ^2 entre los correspondientes perfiles y el perfil medio.

Agrupación de Ward de perfiles fila o de perfiles columna

Aquí describimos la agrupación del capítulo 15 en términos de filas. Podemos aplicar exactamente lo mismo a la agrupación de columnas. En cada paso del algoritmo, agrupamos las filas para minimizar la disminución del valor del estadístico χ^2 (de forma equivalente, para minimizar la disminución del valor de la

inercia, dado que la inercia = χ^2/n, donde n es el total de la tabla). Este criterio de agrupación es equivalente a la agrupación de Ward, en la que ponderamos cada grupo con la masa total de sus miembros. Podemos demostrar que la medida de la diferencia entre filas es igual a la forma ponderada de la distancia ji-cuadrado entre perfiles. Supongamos que $\mathbf{a}_i$ y r_i, con $i = 1, ..., I$, simbolizan, respectivamente, los perfiles de la fila I de la matriz de datos y sus masas. A continuación identificamos el par que da la menor disminución de inercia, lo que es equivalente a buscar el par de filas (i, i') que minimiza la siguiente expresión:

$$\frac{r_i r_{i'}}{r_i + r_{i'}} \|\mathbf{a}_i - \mathbf{a}_{i'}\|_c^2 \qquad \text{(A.28)}$$

A continuación reunimos las dos filas sumando sus frecuencias, y recalculamos su perfil y su masa. En cada etapa de agrupación de los perfiles fila calculamos la misma medida de diferencia que vimos en (A.28) [véase (15.2) en la pág. 149 para una fórmula equivalente basada en perfiles de grupos], y reunimos los dos perfiles con la mínima diferencia. Por tanto, (A.28) es el nivel de agrupación en términos de disminución de inercia, o si multiplicamos por n, de disminución de χ^2. En el caso de tablas de contingencia, podemos contrastar la significación del nivel de agrupación utilizando la tabla del final de este apéndice.

Tablas concatenadas

Supongamos que concatenamos horizontal o verticalmente las tablas $\mathbf{N}_{qs}$, con $q = 1, ..., Q$, $s = 1, ..., S$, para formar una matriz compuesta $\mathbf{N}$. Si las frecuencias marginales son las mismas en todas las filas y en todas las columnas (caso que sucede cuando cruzamos de forma independiente los mismos individuos en varias tablas), entonces, la inercia de $\mathbf{N}$ es la media de las inercias de las tablas $\mathbf{N}_{qs}$:

$$\text{inercia}(\mathbf{N}) = \frac{1}{QS}\sum_{q=1}^{Q}\sum_{s=1}^{S}\text{inercia}(\mathbf{N}_{qs}) \qquad \text{(A.29)}$$

AC múltiple

Supongamos que la matriz original de datos categóricos es $N \times Q$, es decir, N casos y Q variables. El AC múltiple clásico (ACM) tiene dos versiones. En la primera transformamos los casos clasificados por variables en una matriz binaria $\mathbf{Z}$ en la que recodificamos los datos categóricos como variables binarias. Si la variable q-ésima tiene J_q categorías, esta matriz binaria tendrá $J = \sum_q J_q$ columnas (mostramos un ejemplo en la tabla de la imagen 18.1). La versión binaria del ACM no es más que la aplicación del algoritmo clásico de AC a la matriz $\mathbf{Z}$. Así obtenemos las coordenadas de los N casos y de las J categorías. La segunda versión del ACM, consiste en calcular la matriz de Burt $\mathbf{B} = \mathbf{Z}^\top\mathbf{Z}$ de todos los cruces de las Q variables (mostramos un ejemplo en la tabla de la imagen 18.4). La versión Burt del ACM consiste en aplicar el algoritmo básico del AC a la matriz $\mathbf{B}$. Obtenemos así las coordenadas de las J categorías ($\mathbf{B}$ es una matriz simétrica). Las

coordenadas estándares de las categorías son idénticas en los dos versiones del ACM, y las inercias principales de la versión Burt son los cuadrados de los de la versión binaria.

AC conjunto

En el AC conjunto (ACCo) ajustamos las tablas situadas fuera de la diagonal de la matriz de Burt, ignorando las matrices situadas en la diagonal. El algoritmo que utilizamos es un procedimiento iterativo de mínimos cuadrados alternados. Aplicamos el AC sobre la matriz de Burt, que vamos modificando sucesivamente sustituyendo las matrices de la diagonal por valores estimados en el AC de la iteración previa. En el apéndice de cálculo explicamos más detalladamente este algoritmo. Cuando en el algoritmo del ACCo llegamos a la convergencia, llevamos a cabo el AC en la última matriz de Burt modificada, $\tilde{\mathbf{B}}$, que tiene en su diagonal matrices perfectamente ajustadas por construcción. Así, por ejemplo, si suponemos que la solución que buscamos es bidimensional, entonces las matrices de la diagonal modificadas cumplen exactamente (A.14), en la que utilizamos sólo dos términos del modelo del AC bilineal (o fórmula de reconstitución).

Porcentaje de inercia explicado por el ACCo

De la misma forma que la inercia total de $\tilde{\mathbf{B}}$ incluye la inercia Δ atribuible a las matrices de la diagonal, las dos inercias principales, $\tilde{\lambda}_1$ y $\tilde{\lambda}_2$, también la incluyen. Para hallar el porcentaje de inercia explicado por la solución bidimensional tenemos que descontar la inercia Δ tanto de la inercia total como de las dos inercias principales. Podemos obtener Δ a partir de la diferencia entre la inercia de la matriz de Burt original $\mathbf{B}$ (de la que conocemos las inercias de la diagonal) y la de la matriz de Burt modificada $\tilde{\mathbf{B}}$ de la siguiente manera [aquí utilizamos el resultado (A.29) que podemos aplicar a las matrices $\mathbf{B}_{qs}$ de $\mathbf{B}$, y a las de $\tilde{\mathbf{B}}$, que tienen las mismas matrices fuera de la diagonal]:

$$\begin{aligned}
\text{inercia}(\mathbf{B}) &= \frac{1}{Q^2}\left(\sum\sum_{q\neq s}\text{inercia}(\mathbf{B}_{qs}) + \sum_q \text{inercia}(\mathbf{B}_{qq})\right) \\
&= \frac{1}{Q^2}\left(\sum\sum_{q\neq s}\text{inercia}(\mathbf{B}_{qs}) + (J-Q)\right) \\
\text{inercia}(\tilde{\mathbf{B}}) &= \frac{1}{Q^2}\left(\sum\sum_{q\neq s}\text{inercia}(\mathbf{B}_{qs})\right) + \Delta
\end{aligned}$$

Restando lo anterior nos lleva a:

$$\text{inercia}(\mathbf{B}) - \text{inercia}(\tilde{\mathbf{B}}) = \frac{J-Q}{Q^2} - \Delta \qquad \text{(A.30)}$$

Lo que nos da el valor Δ:

$$\Delta = \frac{J-Q}{Q^2} - \left(\text{inercia}(\mathbf{B}) - \text{inercia}(\tilde{\mathbf{B}})\right) \qquad \text{(A.31)}$$

Restando este valor del total y de, la suma de las inercias principales (suponiendo una solución bidimensional) nos proporciona el porcentaje de inercia explicado por la solución del ACCo:

$$100 \times \frac{\tilde{\lambda}_1 + \tilde{\lambda}_2 - \Delta}{\text{inercia}(\tilde{\mathbf{B}}) - \Delta} \tag{A.32}$$

En la sección anterior vimos cómo eliminar la inercia extra de las matrices de la diagonal resultante del ACC de la matriz de Burt. A nivel de cada punto (fila o columna), se nos presenta la misma situación. Es decir, cada categoría *j* tiene un valor adicional de inercia, δ_j, atribuible a las matrices modificadas de la diagonal. Para la matriz de Burt original **B** sabemos cómo calcular este valor, atribuible a las matrices de la diagonal: para el *j*-ésimo punto es $(1 - Qc_j)/Q^2$, donde c_j es la *j*-ésima masa (sumando estos valores para $j = 1,..., J$, obtenemos $(J - Q)/Q^2$ que es la inercia extra total atribuible a las matrices diagonales de **B**). Tal como vimos en el apartado anterior, podemos obtener las contribuciones de la solución bidimensional de las inercias de los puntos de la manera siguiente:

$$\text{inercia}(j\text{-ésima categoría de } \mathbf{B}) = \text{componentes de fuera de la diagonal} + \frac{1 - Qc_j}{Q^2}$$

$$\text{inercia}(j\text{-ésima categoría de } \tilde{\mathbf{B}}) = \text{componentes de fuera de la diagonal} + \delta_j$$

Restando lo anterior (los «componentes fuera de la diagonal» son los mismos) nos lleva a:

$$\text{inercia}(j\text{-ésima categoría de } \mathbf{B}) - \text{inercia } (j\text{-ésima categoría de } \tilde{\mathbf{B}}) = \frac{1 - Qc_j}{Q^2} - \delta_j$$

lo que nos permite calcular δ_j:

$$\delta_j = \frac{1 - Qc_j}{Q^2} - \left(\text{inercia}(j\text{-ésima categoría de } \mathbf{B}) - \text{inercia } (j\text{-ésima categoría de } \tilde{\mathbf{B}})\right) \tag{A.33}$$

Descontando este valor de inercia de la *j*-ésima categoría y, de forma similar, de la suma de los componentes de la inercia de las dos dimensiones, obtenemos las contribuciones relativas (calidades) en relación al resultado bidimensional del ACCo:

$$\frac{c_j \tilde{g}_{j1}^2 + c_j \tilde{g}_{j2}^2 - \delta_j}{\left(\sum_k c_j \tilde{g}_{jk}^2\right) - \delta_j} \tag{A.34}$$

donde $\tilde{g}_{jk}$ es la coordenada principal de la *j*-ésima categoría del eje *k* del AC de $\tilde{\mathbf{B}}$ (solución del ACCo). La suma en el denominador es para todas las dimensiones. Fijémonos en que $\sum_j \delta_j = \Delta$; es decir, sumando (A.33) obtenemos (A.31).

Inercias ajustadas en ACM

Podemos ajustar la solución del ACM para optimizar el ajuste de las matrices situadas fuera de la diagonal (lo podríamos denominar ACC *condicionado al* resultado del ACM). Según vimos en el capítulo 19, podemos obtener los ajustes óptimos por regresión de mínimos cuadrados ponderada. Sin embargo, debido a que la solución no está anidada, preferimos introducir una ligera modificación —fácil de calcular a partir del resultado del ACM de la matriz de Burt— que sí nos permite obtener una solución anidada aunque sea subóptima. El mencionado ajuste lo realizamos de la siguiente manera (podemos consultar las págs. 198-200 del cap. 19):

Inercia total ajustada de la matriz de Burt:

$$\text{inercia total ajustada} = \frac{Q}{Q-1} \times \left(\text{inercia de } \mathbf{B} - \frac{J-Q}{Q^2} \right) \tag{A.35}$$

Las *inercias principales ajustadas* (*valores propios*) de la *matriz de Burt*:

$$\lambda_k^{\text{adj}} \left(\frac{Q}{Q-1} \right)^2 \times \left(\sqrt{\lambda_k} - \frac{1}{Q} \right)^2, \quad k = 1,2,\ldots \tag{A.36}$$

Aquí λ_k indica la *k*-ésima inercia principal de la matriz de Burt. Por tanto $\sqrt{\lambda_k}$ es la *k*-ésima inercia principal de la matriz binaria. Hacemos los ajustes sólo en las dimensiones para las cuales $\sqrt{\lambda_k} > \frac{1}{Q}$, y no utilizamos más dimensiones; por tanto, los porcentajes de inercia no suman el 100%. Podríamos demostrar que estos porcentajes son estimaciones que están por debajo de los porcentajes que obtenemos en el ACCo, sin embargo, en la práctica son valores muy similares.

AC y ACM de subgrupos

El AC de subgrupos es simplemente una aplicación del algoritmo del AC a una parte seleccionada de la matriz estandarizada residual **S** que vimos en (A.4) (*no* es una parte de la matriz original). Sin embargo, para todos los cálculos utilizamos las masas de la matriz completa. Supongamos que trabajamos en una parte de la matriz original que contiene todas las filas y sólo una parte de las columnas. En consecuencia, las filas estarán centradas (es decir, sus medias ponderadas se hallan en el origen del mapa), mientras que no ocurrirá lo mismo con las columnas. Llevamos a cabo el ACM de subgrupos realizando el AC de un subgrupo de la matriz binaria o de la matriz de Burt. En el caso de la matriz de Burt, trabajar con una selección de categorías implica que tenemos que especificar el subgrupo, tanto para filas como para columnas.

Análisis de tablas asimétricas cuadradas

Si la matriz de datos **N** es una matriz cuadrada asimétrica, en la que tanto las filas como las columnas hacen referencia a los mismos objetos, podemos expresar **N** como la suma de la parte simétrica y la parte asimétrica:

$$\begin{aligned} \mathbf{N} &= \frac{1}{2}(\mathbf{N} + \mathbf{N}^\top) + \frac{1}{2}(\mathbf{N} - \mathbf{N}^\top) \\ &= \text{simétrica} + \text{asimétrica} \end{aligned} \tag{A.37}$$

Aplicamos el AC a cada una de estas matrices, con una ligera variación para la parte asimétrica. En la parte simétrica $\frac{1}{2}(\mathbf{N}+\mathbf{N}^{\mathsf{T}})$ procedemos con el AC usual, y así obtenemos un conjunto de coordenadas. Las masas son las medias de las masas de las filas y de las columnas correspondientes al mismo objeto: $w_i = \frac{1}{2}(r_i + c_i)$. El análisis de la parte asimétrica $\frac{1}{2}(\mathbf{N}-\mathbf{N}^{\mathsf{T}})$ consiste en la aplicación del algoritmo de AC, *sin centrar* y utilizando las mismas masas que las del análisis simétrico. Es decir, la matriz de «residuos estandarizados» de (A.4) es la matriz de «diferencias estandarizadas»:

$$\mathbf{S} = \mathbf{D}_w^{-\frac{1}{2}}\left[\frac{1}{2}(\mathbf{P}-\mathbf{P}^{\mathsf{T}})\right]\mathbf{D}_w^{-\frac{1}{2}} \tag{A.38}$$

donde **P** es la matriz de correspondencias y $\mathbf{D}_w$ es la matriz diagonal de las masas w_i. Tal como vimos en el capítulo 22, podemos sustituir estos dos análisis por un AC habitual de la matriz compuesta:

$$\begin{bmatrix} \mathbf{N} & \mathbf{N}^{\mathsf{T}} \\ \mathbf{N}^{\mathsf{T}} & \mathbf{N} \end{bmatrix} \tag{A.39}$$

Al desarrollar el AC de la matriz compuesta —obtenida a partir de la matriz **N** de $I \times I$—, observamos que las inercias principales de las dimensiones simétricas son únicas, mientras que las inercias principales de las dimensiones asimétricas ocurren por pares. Por tanto, es fácil atribuir las $2I-1$ dimensiones obtenidas del AC de la matriz compuesta. Por otra parte, observamos que en las coordenadas de las dimensiones simétricas aparecen repeticiones, mientras que para las dimensiones asimétricas aparecen repetidas pero con el signo cambiado (podemos ver un ejemplo en el capítulo 22).

AC canónico

En el AC canónico (ACC) disponemos de una matriz adicional de variables explicativas **X** (independientes). En el análisis exigimos que las dimensiones del AC estén relacionadas linealmente con **X**. Dividimos la inercia total en la inercia del *espacio restringido* directamente relacionada con las variables explicativas y la inercia correspondiente al *espacio no restringido*. Dado que **X** son filas o columnas adicionales, el ACC es un análisis «asimétrico». En ACC es habitual que las filas sean las unidades de muestreo y que **X** sea un conjunto adicional de M columnas, es decir, que tengamos una matriz adicional de $I \times M$. En el ACC llevamos a cabo una regresión en la que calculamos la matriz restringida $I \times J$ cuyas columnas están linealmente relacionadas con **X**. La diferencia entre **P** y la matriz restringida es la matriz no restringida cuyas columnas no están linealmente relacionadas con **X**. Por tanto el ACC consiste en aplicar el AC a una matriz restringida y (opcionalmente) a una matriz no restringida. En todos los cálculos, mantenemos las masas originales de filas y de columnas. Los resultados obtenidos —como coordenadas, inercias principales, contribuciones, fórmulas de reconstitución, etc.— son los

mismos que los del AC habitual. En el cálculo de medias y varianzas, damos por supuesto que las columnas de **X** están estandarizadas y que utilizamos las masas de las filas como pesos. Si tenemos varias variables categóricas independientes, las codificamos como variables binarias de forma similar a como lo haríamos en un análisis de regresión que estandarizamos de la forma habitual.

Los pasos del ACC son los siguientes:

Paso 1 del ACC – Cálculo de la matriz $\mathbf{S}$ *de residuos estandarizados como en AC:*

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)\mathbf{D}_c^{-\frac{1}{2}} \tag{A.40}$$

Paso 2 del ACC – Cálculo de la matriz de proyección $I \times I$*, de rango* M*, que proyecta sobre el espacio restringido:*

$$\mathbf{Q} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^\top\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}_r^{\frac{1}{2}} \tag{A.41}$$

Paso 3 del ACC – Proyección de los residuos estandarizados para obtener la matriz restringida:

$$\mathbf{S}^* = \mathbf{Q}\mathbf{S} \tag{A.42}$$

Paso 4 del ACC – Aplicación a $\mathbf{S}^*$ *de los pasos 1-6 del AC (página 252):*

Paso 5 del ACC – Inercias principales λ_k^* *en el espacio restringido:*

$$\lambda_k^* = \alpha_k^2, \quad k = 1,2,\ldots,K \quad \text{donde} \quad K = \min\{I-1, J-1, M\} \tag{A.43}$$

Paso 6 del ACC (opcional) – Proyección de los residuos estandarizados en el espacio no restringido:

$$\mathbf{S}^\perp = (\mathbf{I} - \mathbf{Q})\mathbf{S} = \mathbf{S} - \mathbf{S}^* \tag{A.44}$$

Paso 7 del ACC (opcional) – Aplicación de los pasos 1-6 del AC a $\mathbf{S}^\perp$.

Tal como describimos en el capítulo 24, podemos expresar las inercias principales de (A.43) como porcentajes de la inercia total, o como porcentajes de la inercia restringida, que es la suma de los cuadrados de los elementos de $\mathbf{S}^*$, igual a $\sum_k \lambda_k^*$.

Tablas para contrastar agrupaciones o dimensiones significativas

En el caso de tablas de contingencia obtenidas de muestras aleatorias, podemos contrastar la significación estadística de la primera inercia principal. Se trata de la misma prueba que utilizamos para la agrupación de Ward en el capítulo 15. En ese último caso podemos obtener el valor crítico del estadístico χ^2 en la tabla de la imagen A.1, de acuerdo con el tamaño de la tabla (en la página 161, podemos consultar el ejemplo sobre las tiendas de comida. Se trata de una tabla de 5 × 4,

	J								
I	3	4	5	6	7	8	9	10	11
3	8,59								
4	10,74	13,11							
5	12,68	15,24	17,52						
6	14,49	17,21	19,63	21,85					
7	16,21	19,09	21,62	23,95	26,14				
8	17,88	20,88	23,53	25,96	28,23	30,40			
9	19,49	22,62	25,37	27,88	30,24	32,48	34,63		
10	21,06	24,31	27,15	29,75	32,18	34,50	36,70	38,84	
11	22,61	25,96	28,90	31,57	34,08	36,45	38,72	40,91	43,04
12	24,12	27,58	30,60	33,35	35,93	38,36	40,69	42,93	45,10
13	25,61	29,17	32,27	35,09	37,73	40,22	42,60	44,90	47,12

Fuente: Pearson, E.S. y Hartley, H.O. (1972). *Biometrika Tables for Statisticians*, Volumen 2: Tabla 51. Cambridge University Press, Gran Bretaña.

Imagen A.1:
Valores críticos para la prueba de comparaciones múltiples en una tabla de contingencia de $I \times J$ (o $J \times I$). Podemos utilizar los mismos valores críticos para contrastar la significación de una inercia principal. El nivel de significación es del 5%

para la que el valor crítico de la tabla de la imagen A.1 es 15,24). Estos valores críticos son los mismos que utilizamos para contrastar la significación de la primera inercia principal. Así, para el mismo ejemplo sobre las tiendas de comida, que mostramos en la imagen 15.3, el valor de la primera inercia principal era de 0,02635, que expresada como χ^2, es de $0{,}02635 \times 700 = 18{,}45$. Dado que 18,45 es mayor que el valor crítico 15,24, la primera inercia principal es estadísticamente significativa (a un nivel del 5%).

Apéndice B

Cálculo del análisis de correspondencias

En este apéndice veremos el cálculo del AC utilizando el lenguaje de programación R, un software de alto nivel que podemos bajar libremente de la página web:

http://www.r-project.org

Supondremos que el lector ya tiene algunos conocimientos básicos sobre este lenguaje, que se ha convertido de facto en el software estándar para el cálculo estadístico. En caso contrario, en el sitio de Internet mencionado podemos encontrar muchos recursos para aprenderlo. Los programas que veremos en este apéndice se hallan también en el web de la red CARME (siglas en inglés de *Correspondence Analysis and Related MEthods*):

http://www.carme-n.org

Al final de este apéndice veremos también algunos programas comerciales y describiremos diferentes opciones para la creación de mapas.

Contenido

El programa R

El programa R proporciona todas las herramientas necesarias para obtener mapas de AC. La más importante es la descomposición en valores singulares (DVS). Estas herramientas son las *funciones* R. Algunas funciones y material relacionado los encontramos en forma de *paquetes* R. Así, el paquete **ca** nos permite llevar a cabo todas las modalidades del AC que hemos descrito en este libro. Lo iremos

viendo en este apéndice. También veremos el paquete **rgl** que nos permite crear mapas en tres dimensiones. De todas formas, empezaremos paso a paso haciendo algunos cálculos sencillos utilizando R. Con la letra tipo Courier indicaremos las instrucciones y las salidas en R. Por ejemplo, vamos a crear la matriz (13.2) de la página 137. Calcularemos su DVS y la guardamos en un objeto R tipo «svd». Luego visualizaremos la parte del objeto etiquetada como «d» (los valores propios):

```
table.T <- matrix(c(8,5,-2,2,4,2,0,-3,3,6,2,3,3,-3,-6,
                    -6,-4,1,-1,-2),nrow=5)
table.SVD <- svd(table.T)
table.SVD$d
[1] 1.412505e+01 9.822577e+00 1.376116e-15 7.435554e-32
```

(Las instrucciones las expresamos en color marrón, y los datos y resultados en color verde.)

Entrada de datos en R

La entrada de datos en R tiene sus peculiaridades. Sin embargo, una vez dominadas éstas, ¡el resto es muy fácil! La funcion `read.table()` es muy útil para introducir matrices de datos. Las fuentes de datos más fácilmente manejables son los archivos de texto o los archivos Excel. Por ejemplo, supongamos que queremos introducir la tabla de datos de 5 × 3 sobre los tipos de lectura que mostramos en la tabla de la imagen 3.1. Veamos tres opciones para leer estos datos.

1. Supongamos que los datos se hallan en un archivo texto como el siguiente:

```
      C1   C2   C3
E1     5    7    2
E2    18   46   20
E3    19   29   39
E4    12   40   49
E5     3    7   16
```

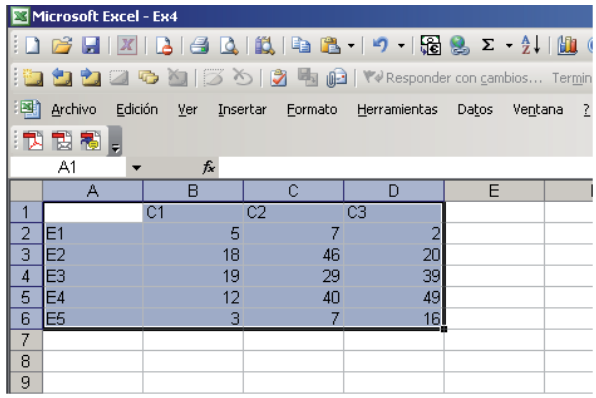
   que llamamos `reader.txt` y que hemos guardado en el directorio de trabajo R. Para leer los datos ejecutaremos la instrucción R siguiente:

```
read.table("reader.txt")
```

2. Otra posibilidad es seleccionar la tabla con el procesador de textos o con Word y luego copiarlo en el portapapeles utilizando la opción copiar del menú de Edit o clicando el botón de la derecha del ratón (suponiendo una plataforma Windows). Para leer directamente la tabla contenida en el portapapeles ejecutaremos la instrucción siguiente:

```
read.table("clipboard")
```

3. De manera similar, podemos leer los datos de un archivo Excel,* seleccionando los datos como mostramos a continuación:



y copiarlos en el portapapeles. A continuación ejecutaremos la siguiente instrucción:

```
table <- read.table("clipboard")
```

Con esta opción la tabla queda guardada como un *data frame* de R llamado `table`. Para indicar en la función `read.table()` que la primera línea contiene las etiquetas de las columnas y que la primera columna de las líneas posteriores contiene las etiquetas de las filas, tenemos que dejar un espacio vacío en la primera línea de la tabla copiada —así, podemos ver que hemos dejado una celda vacía en la esquina de arriba a la izquierda de la tabla Excel. Haríamos lo mismo si se tratara de un archivo de texto. Podemos ver el contenido de `table` ejecutando:

```
table
```

```
     C1   C2   C3
E1    5    7    2
E2   18   46   20
E3   19   29   39
E4   12   40   49
E5    3    7   16
```

El objeto incluye las etiquetas de filas y columnas. Las podemos ver escribiendo `rownames(table)` y `colnames(table)`, por ejemplo:

* Utilizando el paquete **foreign** de R (que se distribuye con el programa) es posible leer otros formatos, como por ejemplo Stata, Minitab, SPSS, SAS, Systat y DBF.

```
rownames(table)
[1] "E1" "E2" "E3" "E4" "E5"
```

Textos R para cada capítulo

A continuación describiremos de forma sistemática cómo realizar con R los cálculos de cada uno de los capítulos del libro. Empezaremos por el capítulo 2 viendo las funciones más básicas de R y el paquete **rgl** de representación gráfica tridimensional. Dejaremos para más adelante las explicaciones del paquete **ca**, que realiza los cálculos del AC de una manera mucho más compacta.

Capítulo 2: Perfiles y espacio de perfiles

En el capítulo 2 mostramos algunos diagramas triangulares correspondientes a los datos de mis viajes. Supongamos que hemos introducido los datos de los perfiles de la tabla de la imagen 2.1, como hemos descrito anteriormente, y que los guardamos en el *data frame* `profiles` de R:

```
profiles <- read.table("clipboard")
profiles
```

	Dias_de_fiesta	Medias_jornadas	Jornadas_completas
Noruega	0.333	0.056	0.611
Canada	0.067	0.200	0.733
Grecia	0.138	0.862	0.000
Francia/Alemania	0.083	0.083	0.833

(Fijémonos en que no hay espacios en blanco en las etiquetas, si los hubiera, los datos no se habrían leído correctamente.) Podemos generar una imagen en tres dimensiones de los perfiles utilizando el paquete **rgl**[*] de la manera siguiente (suponemos que hemos instalado y cargado **rgl**):

Ejemplo de figura tridimensional utilizando el paquete **rgl**

```
rgl.lines(c(0,1.2), c(0,0),   c(0,0))
rgl.lines(c(0,0),   c(0,1.2), c(0,0))
rgl.lines(c(0,0),   c(0,0),   c(0,1.2))
rgl.lines(c(0,0),   c(0,1),   c(1,0), size = 2)
rgl.lines(c(0,1),   c(1,0),   c(0,0), size = 2)
rgl.lines(c(0,1),   c(0,0),   c(1,0), size = 2)
rgl.points(profiles[,3], profiles[,1], profiles[,2], size = 4)
rgl.texts(profiles[,3], profiles[,1], profiles[,2],
    text = row.names(profiles))
```

En la figura de la imagen B.1 mostramos el diagrama de dispersión tridimensional desde un determinado punto de visión. Presionando el botón de la izquierda del ratón podemos hacer girar la figura para dar una mejor sensación tridimensional. En la figura de la imagen B.2 mostramos una de estas rotaciones en la que

[*] El paquete **rgl** no es uno de los paquetes proporcionados como estándar con R, hay que instalarlo bajándolo de la página web de R o de www.carme-n.org.

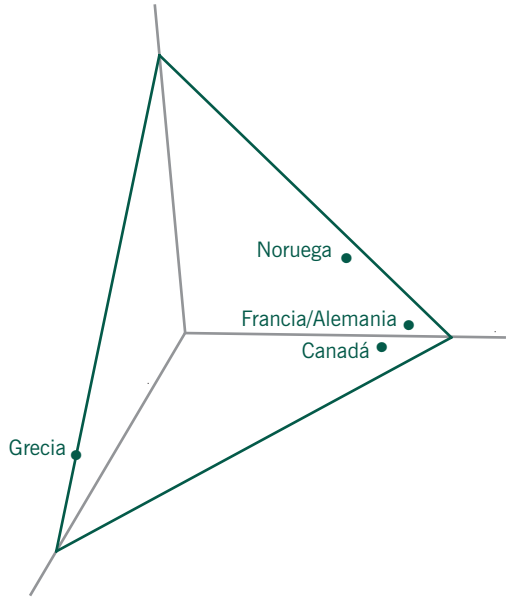


Imagen B.1:
Vista tridimensional de los perfiles fila de los países con los datos sobre los viajes, utilizando el paquete **rgl** *en* R

el punto de vista es plano con relación al triángulo que contiene los puntos correspondientes a los perfiles. La rueda del ratón permite acercar la imagen.

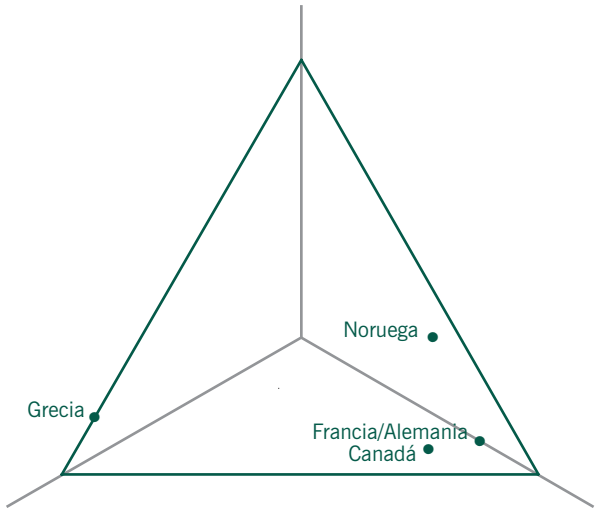
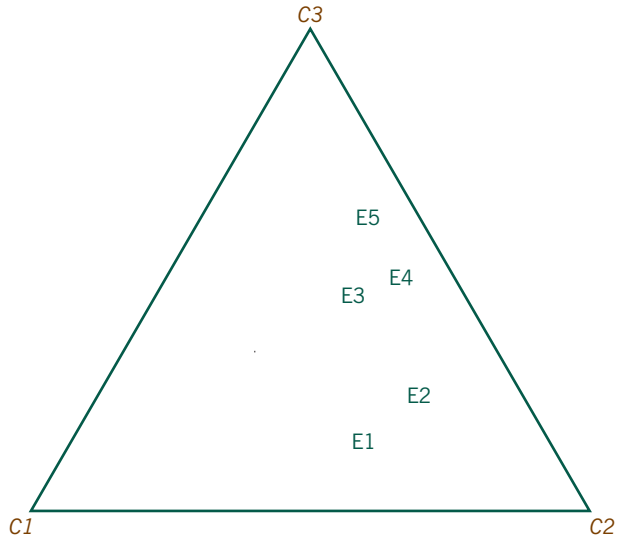


Imagen B.2:
Rotación del espacio tridimensional para mostrar dónde se hallan los puntos correspondientes a los perfiles

Como ilustración de las figuras del capítulo 3, vamos a ver cómo dibujar utilizando R, el diagrama de coordenadas triangular de la imagen 3.2. Para calcular las posiciones (x, y) de los puntos necesitamos un poco de trigonometría.

Capítulo 3:
Masas y centroide

Imagen B.3:
Diagrama de los perfiles de cinco niveles educativos en el espacio de coordenadas triangular



La función `apply()`

Supongamos que hemos leído la tabla como vimos al final de la página 267 y que tenemos los datos almacenados en el *data frame* `table`. Las siguientes instrucciones en R permiten obtener la figura de la imagen B.3. En la primera instrucción calculamos los perfiles de las filas mediante la función `apply()` y los guardamos en `table.pro`. Podemos aplicar la función `apply()` tanto a filas como a columnas. En nuestro caso el parámetro «`1`» indica filas, y `sum` la suma de sus elementos. En este caso dividimos los elementos de cada una de las columnas por la suma de los elementos de las correspondientes filas. En las dos líneas siguientes calculamos las coordenadas *x* e *y* de los cinco perfiles en un triángulo equilátero de lado 1. Para ello utilizamos los valores del primer y del tercer perfil (para situar los puntos sólo necesitamos dos de las tres coordenadas).

Ejemplo de figura bidimensional

```
table.pro <- table/apply(table, 1, sum)
table.x   <- 1 - table.pro[,1] - table.pro[,3]/2
table.y   <- table.pro[,3] * sqrt(3)/2
plot.new()
lines(c(0,1,0.5,0), c(0,0,sqrt(3)/2,0), col = "gray")
text(c(0,1,0.5), c(0,0,sqrt(3)/2), labels = colnames(table))
text(table.x, table.y, labels = rownames(table))
```

Capítulo 4: Distancias ji-cuadrado e inercia

En el capítulo 4 calculamos el estadístico χ^2, la inercia y las distancias χ^2. Vamos a ver cómo llevar a cabo cada uno de estos cálculos. Los realizaremos con los datos sobre los tipos de lectura que anteriormente guardamos en el *data frame* `table`.

— *Estadístico χ^2 e inercia total*

```
table.rowsum <- apply(table, 1, sum)
table.colsum <- apply(table, 2, sum)
table.sum    <- sum(table)
table.exp    <- tabla.rowsum %o% table.colsum / table.sum
chi2         <-sum((table - table.exp)^2 / table.exp)
chi2
[1] 25.97724
chi2 / table.sum
[1] 0.08326039
```

Producto externo, el operador `%o%`

Fijémonos en la utilización del operador `%o%` de *producto externo*, en la cuarta línea del programa anterior. Multiplica cada elemento del vector situado a su izquierda por cada elemento del vector situado a su derecha.

— *Distancias χ^2 de los perfiles fila al centroide*

Vamos a ver cómo calcular el cuadrado de la distancia χ^2 para la quinta fila de la tabla de perfiles, como vimos en (4.4). Para calcular la suma de los tres términos utilizamos la iteración `for` de R:

Ejemplo de iteración en R *utilizando* `for`

```
chidist <- 0
for(j in 1:3)
{chidist<-chidist + (table.pro[5,j] –
table.colmass[j]^2/table.colmass[j]}
chidist
       C1
0.1859165
```

R da la etiqueta `C1` al valor de `chidist`, probablemente porque es la primera columna de la iteración. Otra posibilidad es calcular las cinco distancias de una vez. Para ello, tenemos que restar a cada fila de la matriz de perfiles el correspondiente valor de la columna que contiene las masas de las filas, elevar estas diferencias al cuadrado, y luego dividir otra vez cada fila por las masas para, finalmente, sumar las filas. Debido a que en R, las matrices se guardan como columnas de vectores, las operaciones con filas son ligeramente más complicadas. Una posible solución es: primero transponer la matriz de perfiles, utilizando la función de transposición `t()` para a continuación sumar las columnas del objeto transpuesto (anteriormente filas) utilizando la función `apply()` con los parámetros que indican las suma de columnas `2,sum`:

Función de transposición `t()`

```
apply((t(table.pro)-table.colmass)^2/table.colmass,2,sum)
        E1            E2            E3            E4            E5
0.35335967    0.11702343    0.02739229    0.03943842    0.18591649
```

Función de distancia `dist()`

Adición de filas utilizando `rbind()`

La versátil función `sweep()`

Podemos calcular todas las distancias χ^2 entre perfiles y, en particular, entre éstos y su perfil media, mediante la función `dist()` que, por defecto, calcula la distancia euclídea matricial entre las filas de una matriz. Vamos a añadir la fila que contiene las masas de las columnas (perfil fila medio) a la matriz de perfiles utilizando la función `rbind()` (adición de filas) para formar la matriz de perfiles de 6×3, `tablec.pro`. A continuación dividiremos cada elemento del perfil por la correspondiente raíz cuadrada de la media. Una alternativa a la utilización de la operación de transposición es utilizar la versátil función `sweep()`, similar a `apply()` pero con más opciones. En la tercera línea del texto en R que mostramos a continuación las opciones de la función `sweep()` son `2` (operar en columnas), `sqrt(table.colmass)`(el vector utilizado para la operación) y `"/"` (división):

```
tablec.pro <- rbind(tablec.pro, table.colmass)
rownames(tablec.pro)[6] <- "ave"
dist(sweep(tablec.pro, 2, sqrt(table.colmass), FUN="/"))
```

```
            E1        E2        E3        E4        E5
E2   0.3737004
E3   0.6352512 0.4696153
E4   0.7919425 0.5065568 0.2591401
E5   1.0008054 0.7703644 0.3703568 0.2845283
ave  0.5944406 0.3420869 0.1655062 0.1985911 0.4311803
```

El resultado de la función `dist()` es un objeto que contiene una matriz triangular con todas las distancias entre los cinco perfiles. La última línea de la salida del programa —que hemos etiquetado como "`ave`" (en la segunda línea del programa y que corresponde al perfil fila medio añadido)— muestra las distancias χ^2 de los perfiles a su media.

Capítulo 5: Representación de distancias ji-cuadrado

En el capítulo 5 visualizamos las distancias χ^2 comprimiendo los ejes de coordenadas mediante factores inversamente proporcionales a las raíces cuadradas de sus correspondientes masas. Para hacer la representación gráfica seguiremos una secuencia de codificación similar a la que vimos anteriormente para el diagrama tridimensional del capítulo 2, con la excepción de que dividiremos cada coordenada por `sqrt(table.colmass)`. Un detalle importante para reproducir la figura de la imagen 5.2 es decidir qué elementos del perfil van en cada dimensión. Lo dejamos como ejercicio para el lector (de todas formas, en la página web que mencionamos podemos encontrar el texto del programa).

Capítulo 6: Reducción de la dimensionalidad

En el capítulo 6 consideramos el AC como un método que permite reducir dimensiones. Vamos a llevar a cabo nuestra primera descomposición en valores singulares (DVS). Para ello, en primer lugar, introducimos los datos sobre la autopercepción de la salud en `health`. A continuación, seguimos los pasos que vimos en la página 266 del apéndice A. Las pasos preliminares (A.1-A.3) son los siguientes:

```
health.P    <-health/sum(health)
health.r    <-apply(health.P, 1, sum)
health.c    <-apply(health.P, 2, sum)
health.Dr   <-diag(health.r)
health.Dc   <-diag(health.c)
health.Drmh <-diag(1/sqrt(health.r))
health.Dcmh <-diag(1/sqrt(health.c))
```

La función `diag()` *para obtener matrices diagonal*

Las dos últimas instrucciones anteriores crean $\mathbf{D}_r^{-\frac{1}{2}}$ y $\mathbf{D}_c^{-\frac{1}{2}}$, respectivamente. Posteriormente necesitaremos estas matrices de forma repetida (el nombre de objeto `mh` viene de "minus half"). Para poder llevar a cabo el producto de matrices (A.4), tenemos que transformar el *data frame* `health.P` en una matriz del entorno R. Efectuaremos el producto de matrices utilizando el operador `%*%`. Realizamos la DVS indicada en (A.5) con la función `svd()`.

Operador para la multiplicación de matrices `%*%`

```
health.P   <- as.matrix(health.P)
health.S   <- health.Drmh %*% (health.P-health.r %o% health.c)
                          %*% health.Dcmh
health.svd <- svd(health.S)
```

Ejemplo de DVS, la función `svd()`

Calculamos las coordenadas principales y estándares (`pc` y `sc`) como vimos en (A.6-A.9):

```
health.rsc <-health.Drmh %*% health.svd$u
health.csc <-health.Dcmh %*% health.svd$v
health.rpc <-health.rsc %*% diag(health.svd$d)
health.cpc <-health.csc %*% diag(health.svd$d)
```

¡Y esto es todo! Las 14 instrucciones en R anteriores constituyen el algoritmo de cálculo del AC; para calcular coordenadas del mapa de AC simplemente tenemos que sustituir `health` por cualquier otro objeto.

Por ejemplo, para ver los valores de las coordenadas principales de las filas en el primer eje escribiremos:

```
health.rpc[,1]
[1] -0.37107411 -0.32988430 -0.19895401  0.07091332  0.39551813 ...
```

(Fijémonos en que los signos están cambiados con relación al mapa de la imagen 6.3. Ello ocurre a menudo cuando utilizamos distintos softwares. Podemos cambiar sin problemas los signos de todas las coordenadas.)

En el capítulo 7 vimos las propiedades del escalado óptimo del AC. Por tanto, en este capítulo no realizamos cálculos complicados. Simplemente ilustramos el cálculo de la transformación de la escala óptima (7.5) utilizando las funciones R

Capítulo 7: Escalado óptimo

para el cálculo de valores máximos y mínimos. (Debido al cambio de signos de las coordenadas del primer eje, hemos invertido la escala. Es decir, la escala transformada va de 0 = muy buena a 100 = muy mala. Restando 100 a los valores resultantes obtenemos los valores de la tabla de la imagen 7.2.):

```
health.range <- max(health.csc[,1] – min(health.csc[,1])
health.scale <- (health.csc[,1] – min(health.csc[,1]))
    *100/health.range
health.scale
[1] 0.00000 18.86467 72.42164 98.97005 100.00000
```

Capítulo 8: Simetría entre el análisis de filas y el de columnas

En el capítulo 8 vimos más propiedades de los resultados del AC. La figura de la imagen 8.5 no la creamos utilizando R, la creamos directamente en LaTeX (véanse las descripciones sobre la composición gráfica al final de este apéndice). A continuación veremos cómo utilizando algunas instrucciones R, podemos ilustrar las propiedades de máxima correlación del AC. Por ejemplo, la ecuación (8.2) de la página 92. Así, podemos calcular la correlación entre los valores de edad y de salud en la primera dimensión, $\phi^{\mathsf{T}}\mathbf{P}\gamma$, donde ϕ y γ son las coordenadas estándares en la primera dimensión, y $\mathbf{P}$ es la matriz de correspondencias.

```
health.cor <- t(health.rsc[,1] %*% health.P %*% health.csc[,1]
health.cor^2

            [,1]
[,1]  0.1366031
```

El cuadrado de esta correlación es la primera inercia principal (el resultado anterior aparece como una matriz de 1 × 1, ya que resulta de una multiplicación de matrices). El cálculo anterior de las correlaciones queda justificado por el hecho de que las varianzas son 1. Vamos a ver la estandarización (A.12), por ejemplo, para las filas.

```
t(health.rsc[,1]) %*% health.Dr %*% health.rsc[,1]

     [,1]
[,1]    1
```

Capítulo 9: Representaciones bidimensionales

Contacto inicial con el paquete **ca**

En el capítulo 9 vimos la geometría de los mapas bidimensionales. Comparamos los mapas asimétricos con los simétricos. Vamos a aprovechar que el paquete **ca** contiene los datos sobre los fumadores para iniciarnos en su utilización. Una vez instalado y cargado el paquete **ca**, podemos obtener los mencionados datos ejecutando la instrucción:

```
data(smoke)
```

que nos proporciona el *data frame* `smoke`:

```
smoke
```

```
     none  light  medium  heavy
SM      4      2       3      2
JM      4      3       7      4
SE     25     10      12      4
JE     18     24      33     13
SC     10      6       7      2
```

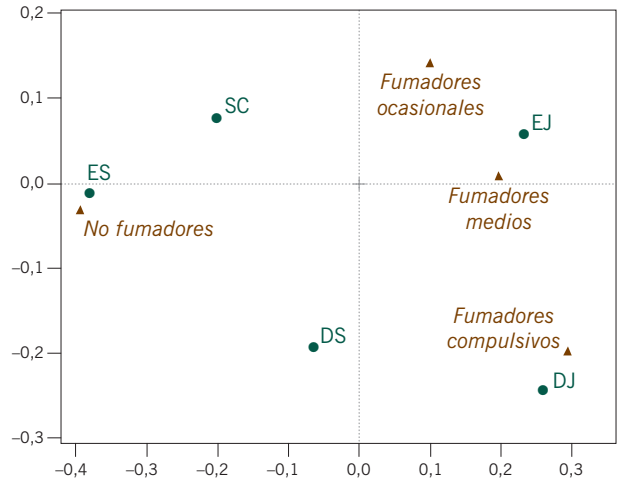
La función `ca()` —una de las funciones del paquete **ca**— nos permite llevar a cabo un AC simple. Así, podemos obtener fácilmente el AC de los datos sobre los fumadores escribiendo `ca(smoke)`:

```
ca(smoke)
```

```
Principal inertias (eigenvalues):

             1          2          3
Value        0.074759   0.010017   0.000414
Percentage   87.76%     11.76%     0.49%

Rows:

                SM         JM         SE         JE         SC
Mass      0.056995   0.093264   0.264249   0.455959   0.129534
ChiDist   0.216559   0.356921   0.380779   0.240025   0.216169
Inertia   0.002673   0.011881   0.038314   0.026269   0.006053
Dim. 1   -0.240539   0.947105  -1.391973   0.851989  -0.735456
Dim. 2   -1.935708  -2.430958  -0.106508   0.576944   0.788435

Columns:

              none      light     medium      heavy
Mass      0.316062   0.233161   0.321244   0.129534
ChiDist   0.394490   0.173996   0.198127   0.355109
Inertia   0.049186   0.007059   0.012610   0.016335
Dim. 1   -1.438471   0.363746   0.718017   1.074445
Dim. 2   -0.304659   1.409433   0.073528  -1.975960
```

De los resultados anteriores nos tendrían que resultar familiares las inercias principales y sus porcentajes. Y para filas y columnas, las masas, las distancias χ^2 al centroide, las inercias y las coordenadas estándares en las dos primeras dimensiones. Más adelante describiremos mucho más detalladamente las características de este paquete. Por el momento sólo vamos a mostrar lo fácil que resulta hacer una representación gráfica. Para obtener el mapa simétrico del AC de la imagen B.4 basta con escribir y ejecutar la función `plot()` con `ca(smoke)`:

Imagen B.4:
Mapa simétrico de los datos sobre los fumadores, utilizando el paquete **ca**




```
plot(ca(smoke))
```

Fijémonos en que, con relación al mapa de la imagen 9.5, aparecen invertidos los dos ejes principales. Para obtener mapas asimétricos, añadiremos la opción `map="rowprincipal"` o `map="colprincipal"` a la función `plot()`. Por ejemplo, podemos obtener el mapa de la imagen 9.2 de la siguiente manera:

```
plot(ca(smoke), map = "rowprincipal")
```

Capítulo 10: Tres ejemplos más

Lo que hemos visto sobre el paquete **ca** nos basta para poder llevar a cabo los análisis del capítulo 10. Los datos que utilizamos están disponibles en la página web www.carme-n.org en formatos texto y Excel. Los datos sobre los autores se hallan también en el paquete **ca**, de manera que los podemos obtener, igual que hicimos con los datos sobre los fumadores, con la instrucción R `data(author)`. Para visualizar los datos sobre los autores en un mapa de AC tridimensional, podemos probar lo siguiente (suponemos que hemos cargado el paquete **ca**):

```
data(author)
plot3d.ca(ca(author), labels = c(2,1), sf = 0.000001)
```

Capítulo 11: Contribuciones a la inercia

En el capítulo 11 introducimos algunos cálculos nuevos. Todos ellos implementados en el paquete **ca**. Sin embargo, igual que antes, primero hagamos los cálculos «a mano». Podemos leer los datos sobre la financiación de la investigación científica como describimos anteriormente —supongamos que hemos llamado `fund` al *data frame* que contiene estos datos. Calculamos la matriz de residuos estanda-

rizados de esta tabla como hicimos en el capítulo 4. Las inercias de la tabla de la imagen 11.1 son las sumas de cuadrados de las filas y de las columnas de la matriz de residuos.

```
fund.P    <- as.matrix(fund/sum(fund))
fund.r    <- apply(fund.P, 1, sum)
fund.c    <- apply(fund.P, 2, sum)
fund.Drmh <- diag(1/sqrt(fund.r))
fund.Dcmh <- diag(1/sqrt(fund.c))
fund.res  <- fund.Drmh %*% (fund.P - fund.r %o% fund.c) %*% fund.Dcmh
round(apply(fund.res^2, 1, sum), 5)
[1] 0.01135 0.00990 0.00172 0.01909 0.01621 0.01256 0.00083
[8] 0.00552 0.00102 0.00466
round(apply(fund.res^2, 2, sum),5)
[1] 0.01551 0.00911 0.00778 0.02877 0.02171
```

Contribuciones de cada celda de la tabla a la inercia total

Las contribuciones, expresadas en tantos por mil, de la tabla de la imagen 11.2 son los cuadrados de los residuos estandarizados con relación al total:

```
round (1000*fund.res^2/sum(fund.res^2), 0)
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0	32	16	0	89
[2,]	0	23	4	44	48
[3,]	3	12	1	0	5
[4,]	9	15	11	189	8
[5,]	106	11	2	74	3
[6,]	1	11	38	1	102
[7,]	2	0	0	3	5
[8,]	51	4	0	10	2
[9,]	10	0	0	2	0
[10,]	5	3	22	26	0

(Después de las multiplicaciones de matrices, hemos perdido las etiquetas de filas y de columnas. Las podemos recuperar utilizando las funciones `rownames()` y `colnames()`.)

Las inercias principales de la tabla de la imagen 11.3 son los cuadrados de los valores singulares resultantes de hacer la DVS (`"svd"`) de la matriz de residuos:

```
fund.svd <- svd(fund.res)
fund.svd$d^2
[1] 3.911652e-02 3.038081e-02 1.086924e-02 2.512214e-03 4.252722e-33
```

(obtenemos cinco valores, sin embargo, el último es teóricamente igual a cero).

Para calcular los componentes individuales de la inercia de las filas en los cuatro ejes, en primer lugar, necesitamos calcular las coordenadas principales f_{ik} [véase (A.8)] y luego los valores $r_i f_{ik}^2$:

```
fund.F    <- fund.Drmh %*% fund.svd$u %*% diag(fund.svd$d)
fund.rowi <- diag(fund.r) %*% fund.F^2
fund.rowi[,1:4]
```

```
                [,1]          [,2]          [,3]          [,4]
 [1,]  6.233139e-04  9.775878e-03  8.222230e-04  1.301601e-04
 [2,]  1.178980e-03  7.542243e-03  8.385857e-04  3.423076e-04
 [3,]  2.314352e-04  8.787604e-04  2.931994e-04  3.211261e-04
 [4,]  1.615600e-02  1.577160e-03  6.274587e-04  7.271264e-04
 [5,]  1.426048e-02  1.043783e-04  1.691831e-03  1.562740e-04
 [6,]  1.526183e-03  9.407586e-03  1.273528e-03  3.573707e-04
 [7,]  7.575664e-06  5.589276e-04  7.980532e-05  1.868385e-04
 [8,]  3.449918e-03  1.601539e-04  1.799425e-03  1.091335e-04
 [9,]  5.659639e-04  7.306881e-06  4.185906e-04  3.022249e-05
[10,]  1.116674e-03  3.684113e-04  3.024590e-03  1.516545e-04
```

lo que concuerda con los valores de la tabla de la imagen 11.5. Fijémonos en que, en la última instrucción anterior, sólo hemos considerado las primeras cuatro columnas (`fund.rowi[,1:4]`). Dado que el quinto valor singular es cero, los valores de la quinta columna teóricamente son cero. Finalmente, expresamos estos componentes con relación a la inercia de un punto (suma de filas) o a la inercia de un eje (sumas de columnas, es decir, las inercias principales) [véanse (A.27) y (A.26), respectivamente], y al mismo tiempo los expresamos en tantos por mil, de la siguiente manera:

Cálculo de las contribuciones relativas (cosenos al cuadrado o correlaciones)

```
round(1000*(fund.rowi/apply(fund.rowi, 1, sum)) [,1:4], 0)
```

```
      [,1] [,2] [,3] [,4]
 [1,]   55  861   72   11
 [2,]  119  762   85   35
 [3,]  134  510  170  186
 [4,]  846   83   33   38
 [5,]  880    6  104   10
 [6,]  121  749  101   28
 [7,]    9  671   96  224
 [8,]  625   29  326   20
 [9,]  554    7  410   30
[10,]  240   79  649   33
```

lo que concuerda con los datos de la tabla de la imagen 11.6 (para obtener las calidades que aparecen en la tabla de la imagen 11.8 sumamos las primeras dos columnas de la tabla anterior). Respecto a las sumas de columnas, es decir, las inercias principales:

Cálculo de las contribuciones a cada eje principal

```
round(1000*t(t(fund.rowi)/fund.svd$d^2) [,1:4], 0)
```

```
      [,1] [,2] [,3] [,4]
 [1,]   16  322   76   52
 [2,]   30  248   77  136
 [3,]    6   29   27  128
 [4,]  413   52   58  289
 [5,]  365    3  156   62
 [6,]   39  310  117  142
 [7,]    0   18    7   74
 [8,]   88    5  166   43
 [9,]   14    0   39   12
[10,]   29   12  278   60
```

que muestra cómo se ha construido cada eje. Por ejemplo, las filas 4 y 5 (Física y Zoología) son las que más contribuyen al primer eje.

Anticipándonos un poco a la descripción completa del paquete **ca**, podemos ver que si aplicamos la función `summary()` a `ca(fund)` obtenemos los resultados completos del análisis anterior:

```
summary(ca(fund))
```

```
Principal inertias (eigenvalues):

      dim      value      %      cum%  scree plot
[1,]  1        0.039117   47.2   47.2  ************************
[2,]  2        0.030381   36.7   83.9  ******************
[3,]  3        0.010869   13.1   97.0  ******
[4,]  4        0.002512    3.0  100.0
[5,]           --------  -----
[6,]  Total:   0.082879  100.0

Rows:

     name   mass  qlt  inr     k=1  cor  ctr     k=2  cor  ctr
 1 |  Gel |  107  916  137 |    76   55   16 |   303  861  322 |
 2 |  Bic |   36  881  119 |   180  119   30 |  -455  762  248 |
 3 |  Chm |  163  644   21 |    38  134    6 |    73  510   29 |
 4 |  Zol |  151  929  230 |  -327  846  413 |   102   83   52 |
 5 |  Phy |  143  886  196 |   316  880  365 |    27    6    3 |
 6 |  Eng |  111  870  152 |  -117  121   39 |  -292  749  310 |
 7 |  Mcr |   46  680   10 |    13    9    0 |  -110  671   18 |
 8 |  Bot |  108  654   67 |  -179  625   88 |   -39   29    5 |
 9 |  Stt |   36  561   12 |   125  554   14 |    14    7    0 |
10 |  Mth |   98  319   56 |   107  240   29 |   -61   79   12 |
```

```
Columns:
     name   mass  qlt  inr     k=1  cor  ctr     k=2  cor  ctr
1  |    A |   39  587  187 |   478  574  228 |    72   13    7 |
2  |    B |  161  816  110 |   127  286   67 |   173  531  159 |
3  |    C |  389  465   94 |    83  341   68 |    50  124   32 |
4  |    D |  162  968  347 |  -390  859  632 |   139  109  103 |
5  |    E |  249  990  262 |   -32   12    6 |  -292  978  699 |
```

Capítulo 12: Puntos adicionales

En el capítulo 12 vimos cómo añadir puntos a un mapa utilizando la relación baricéntrica existente entre las coordenadas estándares de las columnas y las coordenadas principales de las filas. Es decir, los perfiles se hallan situados a medias ponderadas de los vértices. El ejemplo que vimos en la página 130 muestra cómo situar el punto adicional *Museos* [4 12 11 19 7], cuya suma total es 53. Si el vector **m** contiene el perfil de *Museos,* los productos escalares de éste con las coordenadas estándares de las columnas, $\mathbf{m}^{\mathsf{T}}\Gamma$, proporciona las coordenadas buscadas:

```
fund.m     <- c(4,12,11,19,7)/53
fund.Gamma <- fund.Dcmh %*% fund.svd$v
t(fund.m) %*% fund.Gamma[,1:2]
```

```
             [,1]       [,2]
[1,]  -0.3143203  0.3809511
```

(En comparación con el mapa de la imagen 12.2, en esta solución, el signo del segundo eje aparece cambiado. Si llevamos a cabo la misma operación con los vectores unitarios de la tabla de la imagen 12.4 como puntos adicionales, y luego los multiplicamos por las coordenadas estándares de las columnas, veríamos que sus posiciones coinciden con estas últimas.)

Capítulo 13: Biplots de análisis de correspondencias

En el capítulo 13 vimos las diferentes escalas de los biplots del AC. En el mapa correspondiente al biplot estándar del AC de la imagen 13.3, las filas están en coordenadas principales y las columnas en coordenadas estándares multiplicadas por la raíz cuadrada de las masas respectivas de las columnas. Dadas las coordenadas estándares calculadas anteriormente en `fund.Gamma`, el cálculo de las coordenadas en este biplot estándar, en las dos primeras dimensiones, es el siguiente:

```
diag(sqrt(fund.c)) %*% fund.Gamma[,1:2]
```

```
              [,1]         [,2]
[1,]   0.47707276   0.08183444
[2,]   0.25800640   0.39890356
[3,]   0.26032157   0.17838093
[4,]  -0.79472740   0.32170520
[5,]  -0.08046934  -0.83598151
```

Perfiles estimados a partir del biplot

En las siguientes instrucciones, en notación matricial, primero guardamos los productos escalares del lado derecho de (13.7), para $K^* = 2$, en `fund.est` y luego calculamos los perfiles estimados multiplicando por las raíces cuadradas $\sqrt{c_j}$ y sumando c_j:

```
fund.est <- fund.F[,1:2] %*% t(diag(sqrt(fund.c)) %*%
    fund.Gamma[,1:2])
oner <- rep(1, dim(fund)[1])
round(fund.est %*% diag(sqrt(fund.c)) + oner %o% fund.c, 3)
```

```
           A     B     C     D     E
 [1,]  0.051 0.217 0.436 0.177 0.120
 [2,]  0.049 0.107 0.368 0.046 0.431
 [3,]  0.044 0.176 0.404 0.160 0.217
 [4,]  0.010 0.143 0.348 0.280 0.219
 [5,]  0.069 0.198 0.444 0.065 0.225
 [6,]  0.023 0.102 0.338 0.162 0.375
 [7,]  0.038 0.145 0.379 0.144 0.294
 [8,]  0.021 0.136 0.356 0.214 0.272
 [9,]  0.051 0.176 0.411 0.124 0.238
[10,]  0.048 0.162 0.400 0.120 0.270
```

resultado que podemos comparar con los verdaderos valores de los perfiles:

```
round(fund.P/fund.r, 3)
```

```
           A     B     C     D     E
Geol   0.035 0.224 0.459 0.165 0.118
Bioc   0.034 0.069 0.448 0.034 0.414
Chem   0.046 0.192 0.377 0.162 0.223
Zool   0.025 0.125 0.342 0.292 0.217
Phys   0.088 0.193 0.412 0.079 0.228
Engi   0.034 0.125 0.284 0.170 0.386
Micr   0.027 0.162 0.378 0.135 0.297
Bota   0.000 0.140 0.395 0.198 0.267
Stat   0.069 0.172 0.379 0.138 0.241
Math   0.026 0.141 0.474 0.103 0.256
```

Calculando las diferencias entre los valores verdaderos y los valores estimados de los perfiles obtenemos una aproximación a los errores individuales. La suma de los cuadrados de estas diferencias, convenientemente ponderadas, nos da un error general del AC bidimensional. Tenemos que ponderar cada fila de diferencias al cuadrado con la correspondiente masa de la fila r_i y cada columna con la inversa del valor esperado $1/c_j$. El cálculo es el siguiente (se trata de una instrucción empaquetada en dos líneas, ¡un ejemplo de programación R concentrada!):

```
sum(diag(fund.r) %*% (fund.est %*% diag(sqrt(fund.c))+
    oner %o% fund.c - fund.P/fund.r)^2 %*% diag(1/fund.c))
[1] 0.01338145
```

Para ver que el resultado es correcto, tenemos que sumar las inercias principales pero *no* las de los dos primeros ejes:

```
sum(fund.svd$d[3:4]^2)
[1] 0.01338145
```

lo que confirma los cálculos anteriores (es el 16% de la inercia no explicada que aparece en la página 143).

Calibración de los ejes del biplot

El cálculo de las calibraciones de los biplots es bastante complicado ya que implica mucha trigonometría. En vez de dar un listado de todo el procedimiento, recomendamos a los lectores interesados que consulten la página web en la que detallamos la programación de la función `biplot.ca()` que calcula las coordenadas de los puntos inicial y final, así como todas las marcas de los ejes del biplot para las columnas.

Capítulo 14: Relaciones de transición y de regresión

En el capítulo 14 vimos varias relaciones lineales entre las coordenadas de filas, de columnas y de datos. Aquí ilustraremos algunas de estas relaciones utilizando la función de modelización lineal de R, `lm()`, que permite especificar pesos en la regresión de mínimos cuadrados. Por ejemplo, vamos a llevar a cabo la regresión de mínimos cuadrados de las coordenadas estándares de las filas (eje *y* de la figura de la imagen 14.2) con relación a las coordenadas estándares (eje *x*). Las variables de la regresión tienen 10 × 5 valores que podemos vectorizar, partiendo de la matriz original expresada en columnas. De esta manera, la variable *x* es el vector (que llamaremos `fund.vecc`) en el que las coordenadas de la primera columna en la primera dimensión se repiten 10 veces, luego la segunda coordenada 10 veces, y así sucesivamente. Mientras que la variable *y* (`fund.vecr`) tiene repetidas las coordenadas de la primera dimensión cinco veces en una columna (calculamos las coordenadas estándares de las filas como `fund.Phi`). Cuando llevemos a cabo los cálculos podemos comprobar los valores de `fund.vecc` y de `fund.vecr`. Los pesos de las regresiones serán las frecuencias de la tabla original `fund`; para vectorizarlos, tenemos que convertir el *data frame* primero en una matriz y luego en un vector utilizando `as.vector()`:

Conversión de objetos de datos utilizando `as.matrix()` *y* `as.vector()`

```
fund.vec  <- as.vector(as.matrix(fund))
fund.Phi  <- fund.Drmh %*% fund.svd$u
fund.vecr <- rep(fund.Phi[,1], 5)
fund.vecc <- as.vector(oner %*% t(fund.Gamma[,1]))
```

Llevamos a cabo la regresión de mínimos cuadrados ponderada de la manera siguiente:

Ejemplo de `lm()`*, función para regresiones lineales utilizando la opción* `weights`

```
lm(fund.vecr~fund.vecc, weights = fund.vec)

Call:
lm(formula = fund.vecr ~ fund.vecc, weights = fund.vec)

Coefficients:
(Intercept)     fund.vecc
 -2.015e-16     1.978e-01
```

lo que muestra que la constante es cero y que el coeficiente es 0,1978, la raíz cuadrada de la primera inercia principal.

Para llevar a cabo la regresión descrita en la página 151 entre los cocientes de contingencia de Geología con relación a las coordenadas estándares en las primeras dos dimensiones, llevamos a cabo la regresión de la respuesta `fund.y` sobre las dos los primeras columnas de la matriz de coordenadas Γ en `fund.Gamma`, con los pesos $\mathbf{c}$ en `fund.c`, de la manera siguiente (para obtener más resultados, aplicamos la función `summary()` a la instrucción `lm()`):

```
fund.y <- (fund.P[1,]/fund.r[1]/fund.c
summary (lm(fund.y ~ fund.Gamma[,1] + fund.Gamma[,2],
    weights = fund.c))

Call:
lm(formula = fund.y ~ fund.Gamma[, 1]+fund.Gamma[, 2], weights = fund.c)

Residuals:
       A          B          C          D          E
-0.079708   0.016013   0.037308  -0.030048  -0.003764

Coefficients:
                  Estimate  Std. Error    t value    Pr(>|t|)
(Intercept)        1.00000     0.06678     14.975     0.00443 **
fund.Gamma[, 1]    0.07640     0.06678      1.144     0.37105
fund.Gamma[, 2]    0.30257     0.06678      4.531     0.04542 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06678 on 2 degrees of freedom
Multiple R-Squared: 0.9161, Adjusted R-squared: 0.8322
F-statistic: 10.92 on 2 and 2 DF, p-value: 0.0839
```

resultado que confirma los coeficientes que vimos al final de la página 151 (otra vez, el segundo coeficiente tiene el signo opuesto debido a que las coordenadas de la segunda dimensión tienen también signos opuestos) y R^2 es 0,916.

La función `lm()` no proporciona los coeficientes de regresión estandarizados. Sin embargo los podemos obtener utilizando la función de covarianza ponderada `cov.wt()` con la opción `cor=TRUE` para el cálculo de correlaciones ponderadas.

Ejemplo de la función `cov.wt()` *para calcular la correlación ponderada*

```
cov.wt(cbind(fund.y,fund.Gamma[,1:2], wt = fund.c, cor = TRUE)$cor $cor

            [,1]          [,2]          [,3]
[1,]  1.0000000  2.343286e-01  9.280040e-01
[2,]  0.2343286  1.000000e+00  2.359224e-16
[3,]  0.9280040  2.359224e-16  1.000000e+00
```

lo que concuerda, excepto por algunos cambios de signo, con la matriz de correlaciones que vimos al principio de la página 152.

Capítulo 15: Agrupación de filas o de columnas

En el capítulo 15 vimos la agrupación de Ward para realizar la agrupación ponderada de filas o de columnas con sus masas. La función R `hclust()` con la que llevamos a cabo la agrupación jerárquica no permite ponderar [véase (15.2)]. Tampoco lo permite la función `agnes()` del paquete **cluster**. Sin embargo, el paquete estadístico comercial XLSTAT, que describiremos más adelante, sí presenta esta posibilidad que también incluye los programas R de Fionn Murtagh (pág. 327).

Capítulo 16: Tablas de múltiples entradas

En el capítulo 16, describimos la codificación interactiva de variables. Para llevarla a cabo partimos de los datos originales o bien de una tabla de múltiples entradas derivada de los mismos. Por ejemplo, en el caso de los datos sobre la salud que vimos en el capítulo 16, los datos originales tenían el siguiente aspecto (mostramos las primeras cuatro filas de un total de 6371):

```
. . .   health  age  gender    . . .
. . .      4     5      2      . . .
. . .      2     3      1      . . .
. . .      2     4      1      . . .
. . .      3     5      1      . . .
. . .      .     .      .      . . .
. . .      .     .      .      . . .
```

Para obtener la tabla de la imagen 16.2, tenemos que combinar las siete categorías de edad y las dos categorías de género, para formar una nueva variable `age_gender` con 14 categorías. Lo conseguimos con la siguiente transformación:

Codificación interactiva

```
age_gender <- 7*(gender – 1) + age
```

que numerará los grupos de edad de hombres (`gender=1`), de 1 a 7, y los de las mujeres (`gender=2`), de 8 a 14. A partir de ahí, crearemos la tabla de contingen-

cia cruzando las variables `age_gender` y `health`. En R, creamos las tablas de contingencia con la función `table()`. Por ejemplo,

Tablas de contingencia con `table()`

```
table(age-gender, health)
```

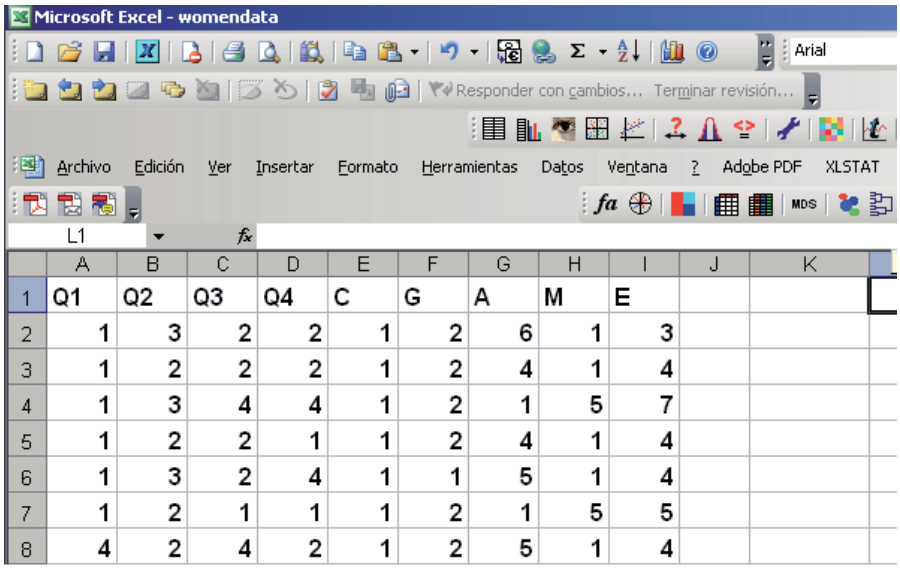
proporcionaría la tabla de contingencia de la imagen 16.2.

Supongamos ahora que los datos originales sobre el trabajo de las mujeres se halla en un archivo Excel como el que mostramos más adelante: cuatro preguntas, de `Q1` a `Q4`, país (`C`, de *country*), género (`G`, de *gender*), edad (`A`, de *age*), estado civil (`M`, de *marital status*) y educación (`E`, de *education*). Para introducir los datos en R, copiaremos, como vimos anteriormente las columnas en el portapapeles, y utilizaremos la función `read.table()`. Sin embargo, ahora las filas de la tabla no tienen nombres. Además, no hay un espacio en blanco en la celda de arriba a la izquierda. Por tanto, tenemos que especificar la opción `header=T` (`T` es la abreviatura de `TRUE`):

```
women <- read.table("clipboard"), header = T)
```

Podemos asignar nombres de las columnas del *data frame* `women` utilizando la función `colnames()`:

```
colnames(women)
[1] "Q1" "Q2" "Q3" "Q4" "C" "G" "A" "M" "E"
```



	A	B	C	D	E	F	G	H	I	J	K
1	Q1	Q2	Q3	Q4	C	G	A	M	E		
2	1	3	2	2	1	2	6	1	3		
3	1	2	2	2	1	2	4	1	4		
4	1	3	4	4	1	2	1	5	7		
5	1	2	2	1	1	2	4	1	4		
6	1	3	2	4	1	1	5	1	4		
7	1	2	1	1	1	2	1	5	5		
8	4	2	4	2	1	2	5	1	4		

Ejemplo de la función `attach()`

Para obtener la tabla de la imagen 16.4 podemos utilizar la función `attach()`, que permite disponer de las etiquetas de las columnas como si fueran un objeto de R (para que no sean disponibles podemos invertir la operación utilizando `detach()`):

```
attach(women)
table(C, Q3)
```

```
   Q3
C        1      2     3     4
   1   256   1156   176   191
   2   101   1394   581   248
   3   278    691    62    66
   4   161    646    70   107
   .     .      .     .     .
   .     .      .     .     .
  21   243    448   484    25
  22   468    664    92    63
  23   203    671   313   120
  24   738   1012   514   230
```

(compárese con la imagen 16.4).

Para obtener la tabla de la imagen 16.6 original con la variable fila codificada interactivamente:

```
CG <- 2*(C-1) + G
table(CG, Q3)
```

```
   Q3
CG       1     2     3     4
   1   117   596   114    82
   2   138   559    60   109
   3    43   675   357   123
   4    58   719   224   125
   .     .     .     .     .
   .     .     .     .     .
  47   348   445   294   112
  48   390   566   218   118
  51     1     2     0     0
  55     1     1     2     1
```

Fijémonos en que las dos últimas filas de la tabla corresponden a unos pocos valores perdidos de género que codificamos como `9`. Para visualizar los recuentos de frecuencias de las columnas, ejecutaremos la instrucción `lapply(women, table)`. De todas formas, primero deberíamos eliminar todos los valores perdidos —en la página 308 mostramos cómo eliminar las filas con valores perdidos—. También podríamos asignar el código NA de R a los valores perdidos. Así, para los valores perdidos de género de la columna 6:

```
women[,6] [G==9] <- NA
attach(women)
CG <- 2*(C – 1) + G
```

(fijémonos en que tenemos que aplicar de nuevo la función `attach()` al *data frame* `women` y recalcular `CG`). Suponiendo que hemos recodificado todos los valores perdidos (o eliminado las correspondientes filas), para construir una variable con 228 categorías que codifique interactivamente país, género, y edad (para edad no hay valores perdidos), codificamos las combinaciones de `CG` y `A` de la manera siguiente:

```
CGA <- 6*(CG – 1) + A
```

Capítulo 17: Tablas concatenadas

En el capítulo 17 vimos el AC de varias tablas de contingencia concatenadas. Las funciones `rbin()` y `cbin()` permiten agregar filas y columnas. Por ejemplo, supongamos que disponemos de la matriz de datos `women` sobre la que hemos aplicado la función `attached()` como vimos anteriormente. Podemos obtener la matriz compuesta de cinco tablas de contingencia correspondientes a la pregunta 3, que esquematizamos en la imagen 17.1 utilizando la iteración `for` de la siguiente manera:

```
women.stack <- table(C, Q3)
for (j in 6:9) {women.stack <- rbind(women.stack,
    table(women[,j], Q3))}
```

Podemos acceder a las columnas de `women` por su etiqueta o por el número de columna. Si miramos el contenido de `women.stack` veremos que para todas las variables demográficas, excepto país y grupo de edad, existen varias filas con valores perdidos. Antes de llevar a cabo el AC tenemos que omitir estos valores. Lo podemos hacer de tres maneras distintas: 1) excluyendo estas filas de la matriz; por ejemplo, podemos eliminar las filas 38, 39, 47 y 48 de la siguiente manera:

```
women.stack <- women.stack[-c(38,39,47,48),]
```

(el signo negativo antes de los números de las filas indica exclusión); 2) cambiando los códigos de los valores perdidos a NA, como describimos en la página anterior; o 3) declarando que las filas con valores perdidos se hallan fuera del subgrupo de interés en el AC de subgrupos que vimos en el capítulo 21 (dado que mantiene el tamaño de la muestra de todas las tablas, es la mejor opción).

El estadístico χ^2 utilizando la función de la prueba χ^2, `chisq.test()`

Para comprobar las inercias de la tabla de la página 168, podemos utilizar la función de la prueba χ^2 de R, `chisq.test()`. Uno de sus resultados es el estadístico χ^2, que podemos especificar mediante `$statistic`. Vamos a hacer los cálculos para la tabla de contingencia que cruza la variable edad con la pregunta 3, lo

que corresponde a las filas de la 27 a la 32 de la matriz compuesta (después de las 24 filas de país y las 2 de género). Obtenemos la inercia dividiendo el estadístico por el tamaño de la muestra, el total de la tabla.

```
chisq.test(women.stack[27:32,])$statistic/sum(women.stack[27:32,])
```

```
x-squared
0.0421549
```

Lo que concuerda con el valor de edad de la tabla de la página 168.

Para unir horizontalmente, con relación a las cuatro preguntas, las cuatro tablas (compuestas asimismo de cinco tablas unidas verticalmente) esquematizadas en la imagen 17.3 utilizaremos la función `cbind()` que permite unir columnas.

El paquete **ca**

Para llevar a cabo el ACM y métodos relacionados dejaremos los cálculos «a mano» que hemos utilizado hasta ahora, para empezar a utilizar las funciones del paquete **ca**. El paquete contiene funciones que permiten realizar el AC simple, múltiple y conjunto, así como funciones que facilitan el análisis de subgrupos y la inclusión de variables adicionales. También ofrece funciones para la representación gráfica de los resultados en dos y en tres dimensiones. El paquete comprende los siguientes componentes:

- AC simple
  - Cálculo: `ca()`
  - Salidas y resúmenes: `print.ca()` y `summary.ca()` (y `print.summary.ca()`)
  - Diagramas: `plot.ca()` y `plot3d.ca()`
- ACM y ACCo
  - Cálculo: `mjca()`
  - Salidas y resúmenes: `print.mjca()` y `summary.mjca()` (y `print.summary.mjca()`)
  - Diagramas: `plot.mjca()` y `plot3d.mjca()`
- Conjuntos de datos
  - `smoke`, `autor` y `wg93`

El paquete contiene más funciones, como `iterate.mjca()` para la actualización de la matriz de Burt en ACCo.

Función `ca()`

La función `ca()` calcula el CA simple, por ejemplo:

```
library(ca)    # carga el paquete ca, en caso de que no se haya hecho
                 antes utilizando el menú de R
```

```
data(smoke)
ca(smoke)
```

lleva a cabo un AC simple con los datos de `smoke` (véanse páginas 288-290). Con la función `names()` podemos obtener una lista de todos los componentes de `ca()`:

```
names(ca(smoke))
 [1]  "sv"          "nd"          "rownames"   "rowmass"    "rowdist"
 [6]  "rowinertia"  "rowcoord"    "rowsup"     "colnames"   "colmass"
[11]  "coldist"     "colinertia"  "colcoord"   "colsup"     "call"
```

Los resultados de `ca()` están estructurados como una lista de objetos. Por ejemplo, obtenemos las coordenadas estándares de las filas con:

```
ca(smoke)$rowcoord
```

La función `ca()` incluye una opción para fijar el número de dimensiones de la solución (`nd`), también incluye una opción para indicar las filas y/o columnas que queremos tratar como puntos adicionales (`suprow` y `supcol`, respectivamente) y opciones para indicar las filas y/o columnas que queremos seleccionar para llevar a cabo el AC de subgrupos (`subsetrow` y `subsetcol`, respectivamente). La función `summary()` permite obtener una salida más detallada:

```
summary(ca(smoke))
```

proporciona el siguiente resumen del AC:

```
Principal inertias (eigenvalues):

      dim      value      %      cum%  scree plot
[1,]  1        0.074759   87.8   87.8  ************************
[2,]  2        0.010017   11.8   99.5  ***
[3,]  3        0.000414    0.5  100.0
[4,]           --------  -----
[5,]  Total:   0.085190  100.0

Rows:

     name   mass   qlt  inr      k=1  cor  ctr      k=2  cor  ctr
1  |   SM |   57   893   31 |   -66   92    3 |  -194  800  214 |
2  |   JM |   93   991  139 |   259  526   84 |  -243  465  551 |
3  |   SE |  264  1000  450 |  -381  999  512 |   -11    1    3 |
4  |   JE |  456  1000  308 |   233  942  331 |    58   58  152 |
5  |   SC |  130   999   71 |  -201  865   70 |    79  133   81 |
```

```
Columns:
     name   mass   qlt  inr      k=1  cor  ctr      k=2  cor  ctr
1  |  non |  316  1000  577 |  -393  994  654 |   -30    6   29 |
2  |  lgh |  233   984   83 |    99  327   31 |   141  657  463 |
3  |  mdm |  321   983  148 |   196  982  166 |     7    1    2 |
4  |  hvy |  130   995  192 |   294  684  150 |  -198  310  506 |
```

Vemos que proporciona los valores propios y los porcentajes de inercia explicada de cada dimensión. También proporciona la inercia explicada en forma de porcentajes acumulados y el diagrama de descomposición (*scree plot* en inglés). En `Rows` y `Columns` encontramos las coordenadas principales de las dos primeras dimensiones (`k=1` y `k=2`). Junto con las coordenadas de los puntos hallamos las correlaciones al cuadrado (`cor`) y las contribuciones (`ctr`). Los valores de estas tablas están multiplicados por 1000. Por tanto, `cor` y `ctr` están expresadas en tantos por mil (‰). También proporciona la calidad (`qlt`) del resultado del AC solicitado. Así, en este ejemplo, la calidad es la suma de los cuadrados de las correlaciones de las dos primeras dimensiones. En el caso de haber variables adicionales, éstas se señalan con un asterisco junto a los nombres de las variables. Por ejemplo, el `summary` del AC de los datos `smoke`, en la que hemos considerado la categoría `none` (la primera columna) como una variable adicional, obtenemos:

```
summary(ca(smoke, supcol=1))
```

y en la correspondiente sección de la salida aparece lo siguiente:

```
...
Columns:
       name    mass qlt   inr     k=1 cor   ctr     k=2 cor   ctr
1 |  (*)non |  <NA>  55  <NA> |  292  39  <NA> |  -187  16  <NA> |
...
```

mostrando que las masas, las inercias y las contribuciones son *"not applicable"*.

Representaciones gráficas con el paquete **ca**

Por defecto, la función `plot()` del paquete **ca** visualiza los resultados del AC y del ACM en forma de mapas *simétricos* (`map="symmetric"`). Las restantes opciones son:

— `"symmetric"` Filas y columnas en coordenadas principales (por defecto), es decir, se realiza el calibrado de manera que la inercia es igual a la inercia principal (valor propio o cuadrado del valor singular).

— `"rowprincipal"` Filas en coordenadas principales y columnas en coordenadas estándares.

— `"colprincipal"` Columnas en coordenadas principales y filas en coordenadas estándares.

— `"symbiplot"` Las coordenadas de filas y columnas se calibran para que tengan inercias iguales a los valores singulares.

— `"rowgab"` Filas en coordenadas principales y columnas en coordenadas estándares multiplicadas por la masa (de acuerdo con la propuesta de Gabriel).

— `"colgab"` Columnas en coordenadas principales y filas en coordenadas estándares multiplicadas por la masa.

— `"rowgreen"` Filas en coordenadas principales y columnas en coordenadas estándares multiplicadas por la raíz cuadrada de la masa (de acuerdo con la propuesta de Greenacre [véase cap. 13]).

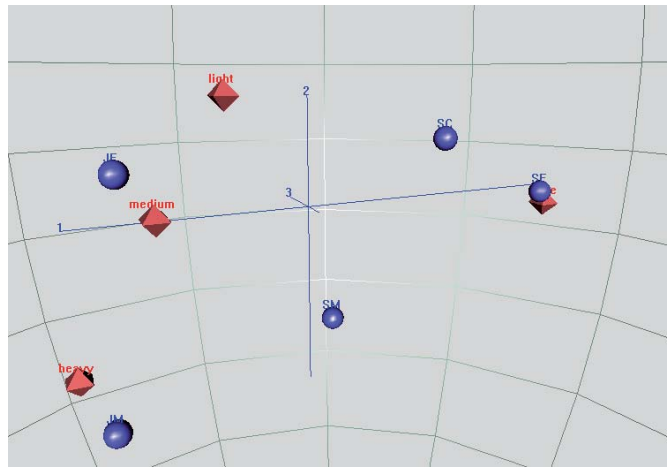
— `"colgreen"` Columnas en coordenadas principales y filas en coordenadas estándares multiplicadas por la raíz cuadrada de la masa.

Por defecto, las variables adicionales aparecen en el mapa con un símbolo distinto. Podemos definir los símbolos con la opción `pch` de `plot.ca()`. Esta opción toma cuatro valores en el orden siguiente: tipo de punto o símbolo para: 1) filas activas, 2) filas adicionales, 3) columnas activas y 4) columnas adicionales. Como regla general, las opciones que incluyen especificaciones para filas y para columnas contienen primero las de las filas y luego las de las columnas. Por ejemplo, especificamos el color de los símbolos con la opción de `col`. Por defecto, `col=c("#000000", "#FF0000")` (negro para las filas y rojo para las columnas). Además de estos códigos hexadecimales existe una lista reducida de nombres: `"black"`, `"red"`, `"blue"`, `"green"`, `"gray"`, etc.

Con la opción `what` podemos especificar las filas y las columnas que queremos visualizar en el mapa. Así podemos indicar `"all"` (todas), `"active"` (activas), `"passive"` (adicionales) o `"none"` (ninguna). Así, por ejemplo, con `what=c("active", "active")` creamos un diagrama con sólo puntos activos (es decir, sin puntos adicionales).

Además de las opciones de escalado de `map`, existen varias opciones que permiten añadir determinados atributos gráficos en los mapas. La opción `mass` hace que el tamaño de los puntos sea proporcional a su masa. De forma similar, utilizando la opción `contrib` podemos indicar mediante la intensidad de color del diagrama las contribuciones relativas o absolutas de los puntos.

Imagen B.5:
Mapa tridimensional de AC simple (lo podemos comparar con el mapa bidimensional de la imagen B.4)



La opción `dim` selecciona las dimensiones del mapa. Por defecto, `dim=c(1,2)` —es decir, se representan las primeras dos dimensiones—. Especificando `dim=c(2,3)`, obtendríamos un mapa con la segunda y la tercera dimensión. Para obtener un mapa en tres dimensiones podemos utilizar las funciones `plot3d.ca()` y `plot3d.mjca()`. Estas dos funciones necesitan del paquete **rgl** de R. Su estructura es similar a la de sus homólogos para dos dimensiones. Por ejemplo,

```
plot3d(ca(smoke, nd=3))
```

crea un mapa tridimensional de AC, como el que mostramos en la imagen B.5, que podemos hacer girar, aumentar o disminuir utilizando el ratón.

La función `mjca()` *del paquete* **ca**

Para llevar a cabo el ACM y el ACCo utilizamos la función `mjca()`. La estructura de esta función es similar a la del AC simple. Las dos diferencias más destacables son el formato de los datos de entrada y la limitación al análisis de columnas (sólo se proporcionan resultados para las columnas). Además, los puntos adicionales se limitan a columnas. Para ejecutar la función `mjca()` es necesario proporcionar los datos en forma de matriz; según el tipo de análisis que realicemos, la función transforma la matriz de datos en una matriz binaria o en una matriz de Burt. Podemos especificar el tipo de análisis a realizar con la opción `lambda` de la función `mjca()`:

— `lambda="indicator"`: análisis basado en un AC simple sobre la matriz binaria;

— `lambda="Burt"`: análisis basado en la descomposición en valores propios de la matriz de Burt;

— `lambda="adjusted"`: análisis basado en la matriz de Burt con inercias ajustadas (por defecto);

— `lambda="JCA"`: análisis de correspondencias conjunto.

Por defecto, la función `mjca()` lleva a cabo un análisis ajustado, es decir, `lambda="adjusted"`. En el ACC (`lambda="JCA"`), la matriz de Burt se actualiza iterativamente por mínimos cuadrados, mediante la función interna `iterate.mjca()`. Esta función de actualización tiene dos criterios de convergencia, `epsilon` y `maxit`. La opción `epsilon` compara la diferencia máxima absoluta de la matriz de Burt de cada iteración con la de la anterior. En la opción `maxit` especificamos el número máximo de iteraciones a realizar. El programa va iterando hasta que se satisface una de las dos condiciones anteriores. Podemos ignorar uno de los dos criterios indicando `NA`. Por ejemplo, podemos llevar a cabo exactamente 50 iteraciones, e ignorar el criterio de convergencia indicando `maxit=50` y `epsilon=NA`.

Igual que en el AC simple, mediante la opción `nd` podemos limitar la solución a dos dimensiones. Sin embargo, para las versiones «binaria» y «Burt» del ACM el programa proporciona los valores propios de $(J - Q)$ dimensiones. En el caso de un análisis ajustado o de un ACC, el programa proporciona sólo los valores propios de las k dimensiones, para las que los valores singulares de la matriz de Burt λ_k (es decir, las inercias principales de la matriz binaria) satisfacen la condición $\lambda_k > 1/Q$.

Capítulo 18: Análisis de correspondencias múltiples

En el capítulo 18, analizamos los datos sobre el trabajo de las mujeres, para las muestras de Alemania Occidental y del Este, utilizando la versión binaria y de Burt del ACM. Supongamos que previamente hemos leído el *data frame* `women` (con 33590 filas) y que hemos aplicado la función `attached()`. Los códigos de las dos muestras alemanas son 2 y 3, respectivamente. Podemos acceder a la parte de `women` correspondiente a estas dos muestras utilizando un vector *lógico* que llamaremos `germany`:

Ejemplo de operación lógica

```
germany <- C==2 / C==3
womenG <- women[germany,]
```

La primera instrucción crea un vector de longitud 33590 con valores `TRUE` para las filas de las muestras alemanas. En caso contrario, dichos valores son `FALSE`. La segunda instrucción crea un *data frame* llamado `womenG` sólo con las 3421 filas con valores `TRUE`. Sin embargo, la matriz que analizamos en el capítulo 18 tenía sólo 3418 porque eliminamos tres casos a los que les faltaban algunos datos demográficos. Supongamos que codificamos los valores perdidos de las variables género y estado civil con el código `9`, y los de educación con los códigos `98` y `99`.

Para eliminar las filas a las que les faltan valores seguimos los mismos pasos que vimos anteriormente, es decir, primero identificamos las filas con valores perdidos y luego las eliminamos:

Eliminación de valores perdidos fila a fila

```
missing <- G==9 / M ==9 / E ==98 / E ==99
womenG  <- (womenG[!missing,])
```

(si codificáramos los valores perdidos mediante el código `NA` de R, como vimos en la página 300, entonces para identificar y luego eliminar las correspondientes filas utilizaríamos este código).

Obtendremos la versión binaria del ACM correspondiente a las cuatro primeras columnas (las cuatro preguntas sobre el trabajo de las mujeres) de la siguiente manera:

```
mjca(womenG[,1:4], lambda = "indicator")

Eigenvalues:

            1        2        3        4        5        6
Value       0.693361 0.513203 0.364697 0.307406 0.21761 0.181521
Percentage  23.11%   17.11%   12.16%   10.25%   7.25%   6.05%
            7        8        9        10       11       12
Value       0.164774 0.142999 0.136322 0.113656 0.100483 0.063969
Percentage  5.49%    4.77%    4.54%    3.79%    3.35%    2.13%

Columns:

             Q1.1      Q1.2      Q1.3      Q1.4      Q2.1      Q2.2
Mass     0.182929  0.034816  0.005778  0.026477  0.013239  0.095012 ...
ChiDist  0.605519  2.486096  6.501217  2.905510  4.228945  1.277206 ...
Inertia  0.067071  0.215184  0.244222  0.223523  0.236761  0.154988 ...
Dim. 1  -0.355941 -0.244454 -0.279167  2.841498 -0.696550 -0.428535 ...
Dim. 2  -0.402501  1.565682  3.971577 -0.144653 -2.116572 -0.800930 ...
```

y la versión de Burt del ACM:

```
mjca(womenG[,1:4], lambda = "Burt")

Eigenvalues:

            1        2        3        4        5        6
Value       0.480749 0.263377 0.133004 0.094498 0.047354 0.03295
Percentage  41.98%   23%      11.61%   8.25%    4.13%    2.88%
            7        8        9        10       11       12
Value       0.027151 0.020449 0.018584 0.012918 0.010097 0.004092
Percentage  2.37%    1.79%    1.62%    1.13%    0.88%    0.36%
```

```
Columns:
             Q1.1      Q1.2      Q1.3      Q1.4      Q2.1      Q2.2
Mass     0.182929  0.034816  0.005778  0.026477  0.013239  0.095012 ...
ChiDist  0.374189  1.356308  3.632489  2.051660  2.354042  0.721971 ...
Inertia  0.025613  0.064046  0.076244  0.111452  0.073363  0.049524 ...
Dim. 1   0.355941  0.244454  0.279167 -2.841498  0.696550  0.428535 ...
Dim. 2  -0.402501  1.565682  3.971577 -0.144653 -2.116572 -0.800930 ...
```

En ambos casos, igual que en el AC simple, podemos calcular la inercia total como la suma de los cuadrados de los valores singulares:

```
sum(mjca(womenG[,1:4], lambda = "indicator")$sv^2)
[1] 3
sum(mjca(womenG[,1:4], lambda = "Burt")$sv^2)
[1] 1.145222
```

Mediante el componente `subinertia` de la función `mjca()` podemos obtener las contribuciones de cada una de las tablas de la matriz de Burt a la inercia total. A partir de su suma obtenemos la inercia total:

```
sum(mjca(womenG[,1:4], lambda = "Burt")$subinertia)
[1] 1.145222
```

Dado que la inercia total es la media de las inercias de las 16 tablas, la inercia de las tablas individuales es 16 veces los valores de `$subinertia`:

```
16*mjca(womenG[,1:4], lambda = "Burt")$subinertia
          [,1]      [,2]      [,3]      [,4]
[1,] 3.0000000 0.3657367 0.4261892 0.6457493
[2,] 0.3657367 3.0000000 0.8941517 0.3476508
[3,] 0.4261892 0.8941517 3.0000000 0.4822995
[4,] 0.6457493 0.3476508 0.4822995 3.0000000
```

Para hallar las posiciones de las variables adicionales:

```
summary(mjca(womenG, lambda = "Burt", supcol = 5:9))

Principal inertias (eigenvalues):

dim      value      %      cum%  scree plot
1        0.480749   42.0   42.0  *************************
2        0.263377   23.0   65.0  **************
3        0.133004   11.6   76.6  *******
4        0.094498    8.3   84.8  *****
5        0.047354    4.1   89.0  **
```

```
6         0.032950     2.9    91.9   **
7         0.027151     2.4    94.2   *
8         0.020449     1.8    96.0   *
9         0.018584     1.6    97.6   *
10        0.012918     1.1    98.8
11        0.010097     0.9    99.6
12        0.004092     0.4   100.0
          --------   -----
Total:    1.145222   100.0
```

```
Columns:
```

```
       name   mass  qlt  inr      k=1  cor  ctr      k=2  cor  ctr
1   |   Q1.1 |  183  740    6 |    247  435   23 |   -207  305   30 |
2   |   Q1.2 |   35  367   14 |    169   16    2 |    804  351   85 |
3   |   Q1.3 |    6  318   16 |    194    3    0 |   2038  315   91 |
4   |   Q1.4 |   26  923   24 |  -1970  922  214 |    -74    1    1 |
5   |   Q2.1 |   13  255   16 |    483   42    6 |  -1086  213   59 |
6   |   Q2.2 |   95  494   11 |    297  169   17 |   -411  324   61 |
.          .     .    .    .       .    .    .       .    .    .
.          .     .    .    .       .    .    .       .    .    .
17  | (*)C.2 | <NA>  283 <NA> |    -89   48 <NA> |    195  234 <NA> |
18  | (*)C.3 | <NA>  474 <NA> |    188   81 <NA> |   -413  393 <NA> |
19  | (*)G.1 | <NA>   26 <NA> |    -33    5 <NA> |     67   21 <NA> |
20  | (*)G.2 | <NA>   24 <NA> |     34    5 <NA> |    -68   19 <NA> |
21  | (*)A.1 | <NA>   41 <NA> |   -108   12 <NA> |   -170   29 <NA> |
22  | (*)A.2 | <NA>   52 <NA> |    -14    0 <NA> |   -172   52 <NA> |
.          .     .    .    .       .    .    .       .    .    .
.          .     .    .    .       .    .    .       .    .    .
```

Las categorías adicionales se han señalado con un `*`, no tienen ni masa (`mass`), ni valores de inercia (`inr`), ni contribuciones a los ejes principales (`ctr`).

Capítulo 19: Análisis de correspondencias conjunto

Para obtener el mapa del ACCo de la imagen 19.3, simplemente tenemos que cambiar la opción `lambda` por `"JCA"`. Dado que los ejes no están anidados, no se dan los porcentajes de inercia de los diferentes ejes, solamente se dan para el resultado global de todo el espacio.

```
summary(mjca(womenG[,1:4], lambda = “JCA”))
```

```
Principal inertias (eigenvalues):
```

```
1         0.353452
2         0.128616
3         0.015652
4         0.003935
          --------
Total:    0.520617
```

```
Diagonal inertia discounted from eigenvalues: 0.125395
Percentage explained by JCA in 2 dimensions: 90.2%
(Eigenvalues are not nested)
[Iterations in JCA: 31 , epsilon = 9.33e-05]

Columns:

      name   mass  qlt  inr      k=1  cor  ctr     k=2  cor ctr
1   |  Q1.1 |  183  969   21 |    204  693   22 |  -129  276  24 |
2   |  Q1.2 |   35  803   23 |    144   61    2 |   503  742  69 |
3   |  Q1.3 |    6  557   32 |    163    9    0 |  1260  548  71 |
4   |  Q1.4 |   26  992  137 |  -1637  991  201 |   -45    1   0 |
5   |  Q2.1 |   13  597   31 |    394  125    6 |  -764  471  60 |
6   |  Q2.2 |   95  956   26 |    250  431   17 |  -276  525  56 |
.       .     .    .    .        .    .    .       .    .   .
.       .     .    .    .        .    .    .       .    .   .
```

En el ACCo, las correlaciones al cuadrado, y en consecuencia también las calidades, son todas mucho mayores.

En el resultado del ACCo la inercia «total» es la inercia de la matriz de Burt modificada, que incluye una parte debida a las matrices modificadas de la diagonal. Para obtener la inercia de las matrices situadas fuera de la diagonal, tenemos que restar de la inercia total la "`Diagonal inertia discounted from eigenvalues: 0.125395`". Dado que la solución buscada es bidimensional y que por construcción ajusta los valores de las matrices de la diagonal, los primeros dos valores propios también contienen esta parte adicional, que tenemos que descontar. La proporción de inercia (de fuera de la diagonal) explicada es, por tanto:

$$\frac{0,3534+0,1286-0,1254}{0,5206-0,1254}=0,9024$$

es decir, el porcentaje del 90,2% indicado anteriormente [véase el apéndice teórico (A.32)]. El valor del denominador de la expresión anterior, el total ajustado 0,5206 – 0,1254 = 0,3952, también lo podemos obtener como:

$$\text{inercia de } \mathbf{B}-\frac{J-Q}{Q}=1,1452-\frac{12}{16}=0,3952$$

Para obtener la solución del ACM ajustado, es decir, las mismas coordenadas estándares del ACM y (casi) los mismos factores de escala («casi» óptimos ya que mantenemos el anidamiento, lo que no ocurre con el ajuste), escribiremos lo siguiente (no es necesario especificar la opción `lambda "adjusted"` ya que es la opción por defecto):

```
summary(mjca(womenG[,1:4]))

Principal inertias (eigenvalues):

dim      value      %     cum%   scree plot
1        0.349456   66.3  66.3   ************************
2        0.123157   23.4  89.7   *********
3        0.023387    4.4  94.1   *
4        0.005859    1.1  95.2

Adjusted total inertia: 0.526963

Columns:

      name   mass   qlt  inr      k=1   cor  ctr     k=2  cor ctr
1   |  Q1.1 |  183   996   22 |    210   687   23 |  -141  309  30 |
2   |  Q1.2 |   35   822   26 |    145    53    2 |   549  769  85 |
3   |  Q1.3 |    6   562   38 |    165     8    0 |  1394  554  91 |
4   |  Q1.4 |   26  1009  141 |  -1680  1008  214 |   -51    1   1 |
5   |  Q2.1 |   13   505   36 |    412   119    6 |  -743  387  59 |
6   |  Q2.2 |   95   947   27 |    253   424   17 |  -281  522  61 |
.      .      .     .    .        .     .    .       .    .   .
.      .      .     .    .        .     .    .       .    .   .
```

Calculamos la inercia total ajustada, utilizada en los porcentajes anteriores, a partir de la expresión (19.5) de la página 200. Y calculamos las dos primeras inercias principales ajustadas (valores propios) a partir de la expresión (19.6) [véanse también (A.35) y (A.36)].

Capítulo 20: Propiedades del escalado óptimo del ACM

En el capítulo 20 generalizamos las ideas que vimos en los capítulos 7 y 8 al caso multivariante. En este capítulo utilizaremos los datos sobre ciencia y medio ambiente que se hallan disponibles en nuestro paquete **ca**. Para cargar estos datos basta con que ejecutemos la instrucción:

```
data(wg93)
```

El *data frame* resultante `wg93` contiene los resultados de las cuatro preguntas descritas en la página 206, así como los de tres variables demográficas: género, edad y educación (las dos últimas con seis categorías cada una de ellas). Después de salvar los resultados del ACM en el objeto `wg93.mca`, podemos obtener el mapa del ACM de la imagen 20.1 de la siguiente manera:

```
wg93.mca <- mjca(wg93[,1:4], lambda = "indicator")
plot(wg93.mca, what = c("none", "all"))
```

El mapa resultante tiene el primer y el segundo eje invertidos, pero —como dijimos anteriormente— esto no tiene consecuencia alguna.

Obtuvimos la tabla de la imagen 20.2 calculando las contribuciones al eje 1 de una matriz de 5 × 4 (primero calculamos las coordenadas principales `wg93.F`, luego calculamos las contribuciones de las filas `wg93.coli`):

```
wg93.F    <- wg93.mca$colcoord %*% diag(sqrt(wg93.mca$sv))
wg93.coli <- diag(wg93.mca$colmass) %*% wg93.F^2
matrix(round(1000*wg93.coli[,1]/wg93.mca$sv[1]^2, 0), nrow = 5)
```

```
     [,1]  [,2]  [,3]  [,4]
[1,]  115   174   203    25
[2,]   28    21     6     3
[3,]   12     7    22     9
[4,]   69    41    80     3
[5,]   55    74    32    22
```

En las siguientes instrucciones calculamos las primeras coordenadas estándares como las puntuaciones de los cuatro ítems de cada uno de los 871 encuestados, así como la puntuación media:

```
Ascal <- wg93.mca$colcoord[1:5,1]
Bscal <- wg93.mca$colcoord[6:10,1]
Cscal <- wg93.mca$colcoord[11:15,1]
Dscal <- wg93.mca$colcoord[16:20,1]
As    <- Ascal[wg93[,1]]
Bs    <- Bscal[wg93[,2]]
Cs    <- Cscal[wg93[,3]]
Ds    <- Dscal[wg93[,4]]
AVEs  <- (AS+Bs+Cs+Ds)/4
```

Situando en una misma matriz las puntuaciones anteriores, podemos calcular sus correlaciones al cuadrado mediante la función `cor()`:

La función de correlación `cor()`

```
cor(cbind(As,Bs,Cs,Ds,AVEs))^2
```

```
              As           Bs          Cs           Ds        AVEs
As    1.000000000  0.139602528  0.12695057  0.005908244  0.5100255
Bs    0.139602528  1.000000000  0.18681032  0.004365286  0.5793057
Cs    0.126950572  0.186810319  1.00000000  0.047979010  0.6273273
Ds    0.005908244  0.004365286  0.04797901  1.000000000  0.1128582
AVEs  0.510025458  0.579305679  0.62732732  0.112858161  1.0000000
```

En la última fila (o columna) aparecen las correlaciones al cuadrado (en análisis de homogeneidad diríamos valores de discriminación) de la página 210. Su media proporciona la inercia principal de la matriz binaria.

```
sum(cor(cbind(As,Bs,Cs,Ds,AVEs))[1:4,5]^2)/4
[1] 0.4573792

wg93.mca$sv[1]^2
[1] 0.4573792
```

La función de covarianza `cov()`

Otro resultado, no mencionado en el capítulo 20, es que el ACM también maximiza la covarianza media entre las puntuaciones de los cuatro ítems. Para verlo, primero, calculamos la matriz de covarianzas de 4×4 de las puntuaciones (dado que la función `cov()` calcula las covarianzas habituales «no sesgadas», dividiendo por $N - 1$, multiplicando por $(N - 1)/N$ obtenemos las covarianzas «sesgadas»). Luego calculamos el valor medio de los 16 valores utilizando la función `mean()`:

```
cov(cbind(As,Bs,Cs,Ds,AVEs)) * 870/871

             As          Bs         Cs          Ds
As   1.11510429  0.44403796  0.4406401  0.04031951
Bs   0.44403796  1.26657648  0.5696722  0.03693604
Cs   0.44064007  0.56967224  1.3715695  0.12742741
Ds   0.04031951  0.03693604  0.1274274  0.24674968

mean(cov(cbind(As,Bs,Cs,Ds)) * 870/871)
[1] 0.4573792
```

Fijémonos en que la suma de las varianzas de las puntuaciones de los cuatro ítems es igual a 4:

```
sum(diag(cov(cbind(As,Bs,Cs,Ds)) * 870/871))
[1] 4
```

En (20.2) calculamos las varianzas individuales y su media sobre toda la muestra:

```
VARs <- ((As-AVEs)^2 + (Bs-AVEs)^2 + (Cs-AVEs)^2 + (Ds-AVEs)^2)/4
mean(VARs)
[1] 0.5426208
```

que es la pérdida de homogeneidad: 1 menos la primera inercia principal.

Suprimiendo las etiquetas de las filas, podemos obtener el mapa de la imagen 20.3 como un mapa `"rowprincipal"` (ejecutando `help(plot.ca)` podemos visualizar las opciones para hacer diagramas):

```
plot(wg93.mca, map = "rowprincipal", labels = c(0,2))
```

Actualmente, el análisis de subgrupos del capítulo 21 solamente se halla en la función `ca()`. Sin embargo, dado que con la función `mjca()` podemos obtener la matriz de Burt, es fácil hacer el ACM de subgrupos con la matriz de Burt. Empecemos con un análisis AC de subgrupos de vocales y consonantes con los datos de los autores contenidos en el paquete **ca**.

```
data(author)
vowels <- c(1,5,9,15,21)
consonants <- c(1:26)[-vowels]
summary(ca(author, subsetcol = consonants))

Principal inertias (eigenvalues):

       dim     value       %     cum%  scree plot
 [1,]  1       0.007607   46.5  46.5  ************************
 [2,]  2       0.003253   19.9  66.3  ***********
 [3,]  3       0.001499    9.2  75.5  *****
 [4,]  4       0.001234    7.5  83.0  ****
  .    .         .           .     .
  .    .         .           .     .
[12,]          --------  -----
[13,]  Total:  0.016371  100.0

Rows:

      name   mass  qlt  inr      k=1  cor  ctr     k=2  cor ctr
1   |  td( |   85   59   29 |     7    8    1 |  -17   50    7 |
2   |  d() |   80  360   37 |   -39  196   16 |  -35  164   31 |
3   |  lw( |   85  641   81 |  -100  637  111 |    8    4    2 |
4   |  ew( |   89  328   61 |    17   27    4 |   58  300   92 |
.      .       .    .    .       .    .    .      .    .   .
.      .       .    .    .       .    .    .      .    .   .

Columns:

      name   mass  qlt  inr      k=1  cor  ctr     k=2  cor ctr
1   |    b |   16  342   21 |   -86  341   15 |   -6    2    0 |
2   |    c |   23  888   69 |  -186  699  104 |  -97  189   66 |
3   |    d |   46  892  101 |   168  783  171 |  -63  110   56 |
4   |    f |   19  558   33 |  -113  467   33 |  -50   91   15 |
.      .       .    .    .       .    .    .      .    .   .
.      .       .    .    .       .    .    .      .    .   .

summary(ca(author, subsetcol = vowels))

Principal inertias (eigenvalues):

      dim     value       %     cum%  scree plot
[1,]  1       0.001450   63.7  63.7  ************************
[2,]  2       0.000422   18.6  82.3  ******
```

```
[3,]  3       3e-04000   13.2   95.5  ****
[4,]  4       0.000103    4.5  100.0
[5,]          --------  -----
[6,]  Total:  0.002276  100.0

Rows:

     name   mass  qlt  inr    k=1  cor  ctr    k=2  cor  ctr
1  |  td( |   85  832  147 |   58  816  195 |    8   15   13 |
2  |  d() |   80  197   44 |  -12  118    9 |  -10   79   20 |
3  |  lw( |   85  235   33 |   14  226   12 |   -3    9    2 |
4  |  ew( |   89  964  109 |   31  337   60 |   42  627  382 |
.     .      .    .    .      .    .    .      .    .    .
.     .      .    .    .      .    .    .      .    .    .

Columns:

     name   mass  qlt  inr    k=1  cor  ctr    k=2  cor  ctr
1  |    a |   80  571   79 |    9   34    4 |  -35  537  238 |
2  |    e |  127  898  269 |   67  895  393 |    4    3    5 |
3  |    i |   70  800  221 |  -59  468  169 |   50  332  410 |
4  |    o |   77  812  251 |  -79  803  329 |   -8    9   12 |
5  |    u |   30  694  179 |  -71  359  105 |  -69  334  335 |
```

Ahora vamos a realizar el ACM de subgrupos versión Burt que vimos en las páginas 220-221 con los datos sobre el trabajo de las mujeres que hemos guardado en `womenG` después de eliminar los valores perdidos de las variables demográficas (págs. 308-309). Primero utilizamos la función `mjca()` para obtener la matriz de Burt, a continuación aplicaremos el AC de subgrupos al cuadrante de la matriz de Burt reacomodada sin datos perdidos (tabla de la imagen 21.3). Hacemos la selección definiendo un vector de índices que llamamos `subset`:

```
womenG.B <- mjca(womenG)$Burt
subset   <- c(1:16)[-c(4,8,12,16)]
summary(ca(womenG.B[1:16,1:16], subsetrow = subset,
     subsetcol = subset))

Principal inertias (eigenvalues):

dim      value      %      cum%  scree plot
1        0.263487   41.4   41.4  ************************
2        0.133342   21.0   62.4  ************
3        0.094414   14.9   77.3  *********
4        0.047403    7.5   84.7  *****
5        0.032144    5.1   89.8  ***
6        0.026895    4.2   94.0  ***
7        0.019504    3.1   97.1  **
8        0.013096    2.1   99.1  *
9        0.005130    0.8   99.9
```

```
10       0.000231     0.0   100.0
11       0.000129     0.0   100.0
         --------   -----
Total:   0.635808   100.0
```

```
Rows:
     name    mass  qlt  inr       k=1  cor  ctr      k=2  cor  ctr
1  | Q1.1 |  183  592   25 |    -228  591   36 |     11    1    0 |
2  | Q1.2 |   35  434   98 |     784  345   81 |   -397   88   41 |
3  | Q1.3 |    6  700  119 |    2002  306   88 |   2273  394  224 |
4  | Q2.1 |   13  535  113 |   -1133  236   65 |   1276  299  162 |
5  | Q2.2 |   95  452   69 |    -442  421   71 |   -119   30   10 |
6  | Q2.3 |  120  693   64 |     482  688  106 |    -40    5    1 |
7  | Q3.1 |   28  706  114 |   -1040  412  114 |    878  294  160 |
8  | Q3.2 |  152  481   38 |    -120   91    8 |   -249  390   71 |
9  | Q3.3 |   47  748  106 |     990  681  175 |    312   67   34 |
10 | Q4.1 |  143  731   49 |    -390  702   83 |     80   29    7 |
11 | Q4.2 |   66  583   84 |     582  414   84 |   -371  168   68 |
12 | Q4.3 |    7  702  119 |    1824  312   90 |   2041  391  222 |
```

La secuencia de instrucciones que, por regresión lineal, permiten modificar la escala para obtener el mejor ajuste de las tablas situadas fuera de la diagonal de la matriz de Burt no es fácil de hacer. Como es bastante larga no la incluimos en este apéndice. Sin embargo, la podemos encontrar en la página web. Esperamos incorporarla próximamente en el paquete **ca**.

Capítulo 22: Análisis de tablas cuadradas

Tal como vimos en el capítulo 21, el AC de matrices asimétricas cuadradas consiste en dividir la tabla en una parte simétrica y en una parte antisimétrica, y luego llevar a cabo el AC en la parte simétrica y un AC sin centrar en la parte antisimétrica, con los mismos pesos y distancias χ^2. En la matriz compuesta mostrada en (22.4) podemos realizar simultáneamente ambos análisis. Después de leer la tabla de movilidad en el *data frame* `mob`, las secuencias de instrucciones para formar la matriz compuesta y luego llevar a cabo el AC son las siguientes. (Antes de llevar a cabo el análisis, tenemos que transformar `mob` en una matriz. En caso contrario no podríamos combinar de forma adecuada filas y columnas para crear la matriz compuesta `mob2`.):

```
mob  <- as.matrix(mob)
mob2 <- rbind(cbind(mob,t(mob)), cbind(t(mob),mob))
summary(ca(mob2))
```

```
Principal inertias (eigenvalues):
dim      value      %      cum%  scree plot
 1       0.388679   24.3   24.3  ************************
 2       0.232042   14.5   38.8  ***************
 3       0.158364    9.9   48.7  **********
```

```
 4       0.158364     9.9    58.6  **********
 5       0.143915     9.0    67.6  *********
 6       0.123757     7.7    75.4  ********
 7       0.081838     5.1    80.5  *****
 8       0.070740     4.4    84.9  *****
 9       0.049838     3.1    88.0  ***
10       0.041841     2.6    90.6  ***
11       0.041841     2.6    93.3  ***
12       0.022867     1.4    94.7  *
13       0.022045     1.4    96.1  *
14       0.012873     0.8    96.9  *
15       0.012873     0.8    97.7  *
16       0.010360     0.6    98.3  *
17       0.007590     0.5    98.8  *
18       0.007590     0.5    99.3  *
19       0.003090     0.2    99.5
20       0.003090     0.2    99.7
21       0.001658     0.1    99.8
22       0.001148     0.1    99.9
23       0.001148     0.1    99.9
24       0.000620     0.0    99.9
25       0.000381     0.0   100.0
26       0.000381     0.0   100.0
27       0.000147     0.0   100.0
         --------   -----
Total:   1.599080   100.0
```

Rows:

```
     name   mass   qlt   inr      k=1   cor   ctr      k=2   cor ctr
1   | Arm |   43   426    54 |   -632   200    44 |    671   226   84 |
2   | Art |   55   886   100 |   1521   793   327 |    520    93   64 |
3   | Tcc |   29    83    10 |   -195    73     3 |     73    10    1 |
4   | Cra |   18   293    32 |    867   262    34 |   -298    31    7 |
.       .     .     .     .        .     .     .        .     .    .
.       .     .     .     .        .     .     .        .     .    .
15  | ARM |   43   426    54 |   -632   200    44 |    671   226   84 |
16  | ART |   55   886   100 |   1521   793   327 |    520    93   64 |
17  | TCC |   29    83    10 |   -195    73     3 |     73    10    1 |
18  | CRA |   18   293    32 |    867   262    34 |   -298    31    7 |
.       .     .     .     .        .     .     .        .     .    .
.       .     .     .     .        .     .     .        .     .    .
```

Columns:

```
     name   mass   qlt   inr      k=1   cor   ctr      k=2   cor ctr
1   | ARM |   43   426    54 |   -632   200    44 |    671   226   84 |
2   | ART |   55   886   100 |   1521   793   327 |    520    93   64 |
3   | TCC |   29    83    10 |   -195    73     3 |     73    10    1 |
4   | CRA |   18   293    32 |    867   262    34 |   -298    31    7 |
.       .     .     .     .        .     .     .        .     .    .
.       .     .     .     .        .     .     .        .     .    .
```

```
15 | Arm |  43  426   54 |  -632  200   44 |   671  226  84 |
16 | Art |  55  886  100 |  1521  793  327 |   520   93  64 |
17 | Tcc |  29   83   10 |  -195   73    3 |    73   10   1 |
18 | Cra |  18  293   32 |   867  262   34 |  -298   31   7 |
.      .     .    .    .       .    .    .       .    .   .
.      .     .    .    .       .    .    .       .    .   .
```

Las inercias principales coinciden con los valores de la tabla de la imagen 22.4. Las dos primeras dimensiones corresponden a la parte simétrica de la matriz. Las dimensiones 3 y 4, con valores propios repetidos, corresponden a la parte antisimétrica. Podemos observar que las coordenadas de las dos primeras dimensiones aparecen repetidas en dos bloques. Por defecto la función `summary()` proporciona sólo las dos primeras dimensiones. Si queremos más dimensiones tenemos que especificarlo. Por ejemplo, para obtener las cuatro primeras dimensiones:

```
summary(ca(mob2, nd = 4))
```

```
Rows:
```

```
    name      k=3  cor ctr      k=4  cor ctr
1   | Arm |   -11    0   0 |   416   87  47 |
2   | Art |    89    3   3 |   423   61  62 |
3   | Tcc |  -331  211  20 |   141   38   4 |
4   | Cra |  -847  250  80 |    92    3   1 |
.      .       .    .   .       .    .   .
.      .       .    .   .       .    .   .
15  | ARM |    11    0   0 |  -416   87  47 |
16  | ART |   -89    3   3 |  -423   61  62 |
17  | TCC |   331  211  20 |  -141   38   4 |
18  | CRA |   847  250  80 |   -92    3   1 |
.      .       .    .   .       .    .   .
.      .       .    .   .       .    .   .
```

```
Columns:
```

```
    name      k=3  cor ctr      k=4  cor ctr
1   | ARM |  -416   87  47 |   -11    0   0 |
2   | ART |  -423   61  62 |    89    3   3 |
3   | TCC |  -141   38   4 |  -331  211  20 |
4   | CRA |   -92    3   1 |  -847  250  80 |
.      .       .    .   .       .    .   .
.      .       .    .   .       .    .   .
15  | Arm |   416   87  47 |    11    0   0 |
16  | Art |   423   61  62 |   -89    3   3 |
17  | Tcc |   141   38   4 |   331  211  20 |
18  | Cra |    92    3   1 |   847  250  80 |
.      .       .    .   .       .    .   .
.      .       .    .   .       .    .   .
```

Para las dimensiones 3 y 4 de las filas observamos que también se repiten las coordenadas en dos bloques, pero en este caso con los signos cambiados. Asimismo podemos observar que los valores de las coordenadas de las columnas de la tercera dimensión son los de las filas de la cuarta dimensión cambiadas de signo, y que los de las columnas de la cuarta dimensión son los de las filas de la tercera dimensión también cambiadas de signo. En cualquier caso para obtener el mapa sólo necesitamos uno de los dos conjuntos de coordenadas. Recordemos, como vimos en el capítulo 22, que estos mapas tienen una forma propia de interpretación.

Capítulo 23: Recodificación de datos

En el capítulo 23 vimos transformaciones simples de datos y la posterior aplicación del AC habitual. Como ilustración del análisis de datos continuos, utilizaremos los indicadores de la Unión Europea. Supongamos que hemos leído los datos y que los hemos introducido en un objeto llamado `EU`. A continuación convertimos los datos en rangos (utilizando la función `rank()` de R, una vez aplicada la útil función `apply()` para obtener `EUr`). Finalmente realizamos el doblado (para obtener `EUd`) de la siguiente manera:

Conversión a rangos mediante la función `rank()`

```
EUr <- apply(EU, 2, rank)-1
EUd <- cbind(EUr, 11-EUr)
colnames(EUd) <- c(paste(colnames(EU), "-", sep=""),
                   paste(colnames(EU), "+", sep=""))
EUd
```

```
   Unemp- GDPH-  PCH-    PCP-   RULC- Unemp+ GDPH+  PCH+    PCP+   RULC+
Be      6     6     6     6.5    4.5      5     5     5     4.5    6.5
De      4    11    10     0.0    7.0      7     0     1    11.0    4.0
Ge      2    10    11     5.0    6.0      9     1     0     6.0    5.0
Gr      5     1     1     1.0   11.0      6    10    10    10.0    0.0
Sp     11     3     3    10.0    2.0      0     8     8     1.0    9.0
Fr      7     8     8     3.5    4.5      4     3     3     7.5    6.5
Ir     10     2     2    11.0    1.0      1     9     9     0.0   10.0
It      9     7     7     9.0    9.0      2     4     4     2.0    2.0
Lu      0     9     9     3.5    8.0     11     2     2     7.5    3.0
Ho      8     5     4     6.5    3.0      3     6     7     4.5    8.0
Po      1     0     0     8.0    0.0     10    11    11     3.0   11.0
UK      3     4     5     2.0   10.0      8     7     6     9.0    1.0
```

Fijémonos en que hemos introducido los nombres de las columnas con la función `paste()`. Finalmente, ejecutando `ca(EUd)` obtenemos el mapa de la imagen 23.5.

Capítulo 24: Análisis de correspondencias canónico

Con el paquete **ca** no podemos obtener los resultados del capítulo 24. Sin embargo, los podemos obtener utilizando el programa XLSTAT (descrito más adelante) o el paquete **vegan** de Jari Oksanen (véanse los recursos web en el apéndice bibliográfico). Este último recurso no solamente desarrolla el ACC, sino que también permite ejecutar el AC y el ACP (pero sin muchas de las opciones que tenemos en el paquete **ca**). Dado que este paquete se utiliza a menudo en el contexto de

datos sobre ecología, como por ejemplo los del capítulo 24, hablaremos de estaciones o localidades (muestras), «especies» y «variables» (explicativas). La utilización de **vegan** es tan fácil como la de **ca**. La principal función es `cca()`, que la podemos utilizar con los siguientes dos formatos:

```
cca(X, Y, Z)
cca(X ~ Y + condition(Z))
```

donde `X` es la matriz de recuentos de localidades × especies, `Y` es la matriz de localidades × variables de datos explicativos y `Z` es la matriz de localidades × variables de datos condicionados por si queremos hacer (de forma opcional) un ACC parcial. El segundo formato es un formato tipo regresión. Nosotros utilizaremos el primer formato. Si sólo especificamos `X`, se lleva a cabo un análisis del AC (lo podemos comprobar con uno de los ejemplos anteriores. Por ejemplo con `summary(cca(author))` para comparar los resultados con los que hemos obtenido anteriormente —los libros serían las «localidades» y las letras las «especies»—. Por defecto, por ejemplo ejecutando `plot(cca(autor))`, obtenemos el mapa que hemos llamado `"colprincipal"`). Si especificamos `X` e `Y`, realizamos un ACC. Si especificamos `X`, `Y` y `Z`, el análisis es un ACC parcial.

Supongamos que el *data frame* `bio` contiene los datos biológicos que vimos en los capítulos 10 y 24 en una tabla de 13 × 92, y que `env` contiene las variables `logBa`, `logFe` y `logPE` en una tabla de 13 × 3. Podemos llevar a cabo el ACC de la manera siguiente:

```
summary(cca(bio, env))

Call:
cca(X = bio, Y = env)

Partitioning of mean squared contingency coefficient:
Total 0.7826
Constrained 0.2798
Unconstrained 0.5028

Eigenvalues, and their contribution to the
             mean squared contingency coefficient

             CCA1    CCA2     CCA3     CA1     CA2      CA3
lambda     0.1895  0.0615  0.02879  0.1909  0.1523  0.04159
accounted  0.2422  0.3208  0.35755  0.2439  0.4385  0.49161

               CA4      CA5      CA6      CA7      CA8      CA9
lambda     0.02784  0.02535  0.02296  0.01654  0.01461  0.01076
accounted  0.52719  0.55957  0.58891  0.61004  0.62871  0.64245
```

```
Scaling 2 for species and site scores
--- Species are scaled proportional to eigenvalues
--- Sites are unscaled: weighted dispersion equal on all dimensions

Species scores

               CCA1      CCA2      CCA3        CA1       CA2       CA3
Myri_ocul  0.1732392  0.245915 -0.070907  0.6359626 -0.063479  0.031990
Chae_seto  0.5747974 -0.270816  0.011814 -0.5029157 -0.674207  0.093354
Amph_falc  0.2953878 -0.114067  0.075979 -0.2224138  0.041797 -0.005020
Myse_bide -0.5271092 -0.505262 -0.103978 -0.0789909  0.176683 -0.484208
Goni_macu -0.1890403  0.122783 -0.044679 -0.1045244  0.030134  0.111827
Amph_fili -0.9989672 -0.075696  0.107184 -0.3506103  0.076968  0.004931
    .       .          .         .         .          .         .
    .       .          .         .         .          .         .

Site constraints (linear combinations of constraining variables)

         CCA1       CCA2       CCA3
S4   -0.06973    0.75885  -2.29951
S8   -0.35758    1.47282   2.27467
S9    0.48483   -0.72459  -0.66547
S12   0.02536    0.27129  -0.14677
S13   0.30041   -0.01531  -0.80821
S14   0.79386    1.16229   0.24314
S15   0.96326   -0.88970   0.14630
S18  -0.16753    0.25048  -0.77451
S19   0.36890   -0.81800   1.50620
S23  -0.09967   -1.90159   0.06877
S24   0.05478    0.96184  -0.10635
R40  -3.71393   -0.20698   0.53031
R42  -2.96641   -0.18264  -0.67736

Biplot scores for constraining variables

         CCA1       CCA2      CCA3
logBa  0.9957   -0.08413  0.03452
logFe  0.6044   -0.72088  0.33658
logPE  0.4654    0.55594  0.68710
```

Fijémonos en lo siguiente:

— el `mean squared contingency coefficient` es la inercia total;

— en el espacio restringido, los encabezamientos de las inercias principales son `CCA1`, `CCA2`, etc., y en el espacio no restringido, son `CA1`, `CA2`, etc.;

— todos los porcentajes se expresan en relación a la inercia total;

— `Scalin 2` significa filas (localidades) en coordenadas estándares, y columnas (especies) en coordenadas principales, es decir, equivale a las coordenadas de `"colprincipal"` de la función `plot.ca()`;

— las `Species scores` son coordenadas principales de columnas;

— los `Site constraints` son las coordenadas estándares de filas;

— los `Biplot scores for constraining variables` son los coeficientes de correlación ponderados entre las variables explicativas y las coordenadas de las localidades.

Capítulo 25: Consideraciones sobre estabilidad e inferencia

En el capítulo 25 llevamos cabo varios automuestreos de tablas con el fin de investigar su variabilidad, así como pruebas de permutaciones para contrastar hipótesis nulas. Por ejemplo, obtuvimos el mapa del AC del automuestreo parcial de los datos sobre los autores que mostramos en la imagen 25.1 y 25.2 de la siguiente manera (hemos insertado comentarios). Si nos fijamos en la imagen 25.1, sólo hemos dibujado 100 de las 1000 réplicas; como el muestreo es aleatorio, los resultados no serán exactamente iguales:

Muestreo aleatorio multinomial utilizando `rmultinom()`

```
data(author)
author.ca <- ca(author)
nsim <- 1000
# cálculo de la suma de las filas
author.rowsum <- apply(author, 1, sum)
# cálculo de las nsim simulaciones del primer libro
author.sim <- rmultinom(nsim, author.rowsum[1], prob = author[1,])
# cálculo de las nsim simulaciones de los otros libros y
  combinación de columnas
for (i in 2:12) {
    author.sim <- cbind(author.sim,
                  rmultinom(nsim, author.rowsum[i],
                  prob = author[i,]))
    }
# transposición para tener el mismo formato que la
  matriz original
author.sim <-  t(author.sim)
author.sim2 <- matrix(rep(0, nsim*12*26), nrow = nsim*12)
# reorganización de filas para juntar las matrices
for (k in 1:nsim) {
    for (i in 1:12) {
         author.sim2[(k-1)*12+i,] <- author.sim[k+(i-1)*nsim,]
         }
    }
```

Utilizando la fórmula de transición, a partir de las coordenadas estándares de filas calculamos las coordenadas principales simuladas de las columnas:

```
# obtención de las coordenadas estándares de las filas
author.rowsc <- author.ca$rowcoord[,1:2]
# cálculo de las coordenadas principales de todas las réplicas
  mediante la fórmula de transición
```

```
author.colsim <- t(t(author.rowsc) %*% author.sim2[1:12,])/
                 apply(author.sim2[1:12,], 2, sum)
for (k in 2:nsim) {
    author.colsim <- rbind(author.colsim, t(t(author.rowsc) %*%
                     author.sim2[((k-1)*12+1):(k*12),])/
                     apply(author.sim2[((k-1)*12+1):(k*12),], 2, sum))
# reorganización de las coordenadas de las filas para que
  todas las letras estén juntas
author.colsim2 <- matrix(rep(0, nsim*26*2), nrow = nsim*26)
for (j in 1:26) {
    for (k in 1:nsim) {
         author.colsim2[(j-1)*nsim+k,] <- author.colsim[j+(k-1)*26,]
         }
    }
```

Representación gráfica de los puntos y de los perímetros convexos:

```
# representación de todos los puntos (para etiquetarlos
  utilizamos el primer formato)
plot(author.colsim[,1], -author.colsim[,2], xlab = "dim1",
     ylab = "dim2", type = "n")
text(author.colsim[,1], -author.colsim[,2], letters, cex = 0.5,
     col = "gray")
# representación de los perímetros convexos de cada letra
# en primer lugar calculamos las coordenadas principales de
  las letras de la matriz original
author.col <- t(t(author.rowsc) %*% author)/
              apply(author, 2, sum)
for (j in 1:26) {
    points <- author.colsim2[(nsim*(j-1)+1):(nsim*j),]
# en todos estos mapas invertimos la segunda coordenada
    points[,2] <- -points[,2]
    hpts <- chull(points)
    hpts <- c(hpts,hpts[1])
    lines(points[hpts,], lty = 3)
    text(author.col[j,1], -author.col[j,2],
         letters[j], font = 2, cex = 1.5)
    }
```

Finalmente llevamos a cabo el recorte de todos los perímetros convexos hasta eliminar el 5% de los puntos de las proyecciones de las nubes, luego representamos los perímetros convexos recortados:

```
plot(author.colsim2[,1], -author.colsim2[,2], xlab = "dim1",
    ylab = "dim2", type = "n")
for (j in 1:26) {
    points <- author.colsim2[(nsim*(j-1)+1):(nsim*j),]
# en todos estos mapas invertimos la segunda coordenada
    points[,2] <- -points[,2]
```

```
    repeat {
          hpts <- chull(points)
          npts <- nrow(points[-hpts,])
          if(npts/nsim < 0.95) break
          points <- points[-hpts,]
          }
    hpts <- c(hpts,hpts[1])
    lines(points[hpts,], lty = 3)
    text(author.col[j,1], -author.col[j,2], letters[j],
          font = 2)
    }
```

Para representar las elipses de confianza tenemos que bajar el paquete **ellipse** de la página web de R (www.R-project.org). El texto del programa para representar las elipses de confianza utilizando las réplicas del automuestreo es el siguiente:

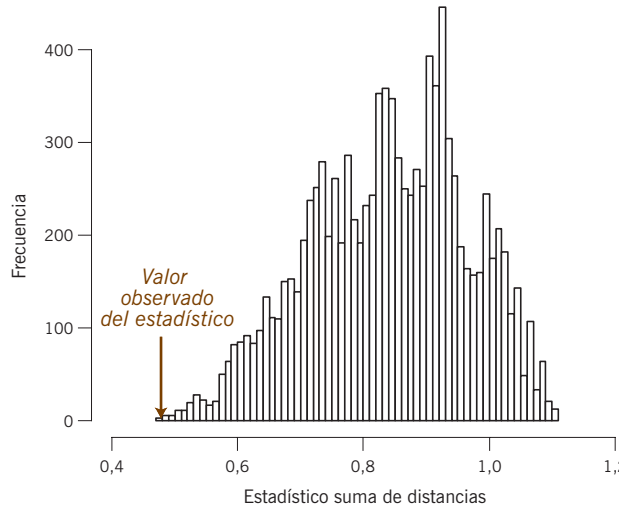
```
# elipses de confianza – necesitamos el paquete 'ellipse'
plot(author.colsim2[,1], -author.colsim2[,2],xlab = "dim1",
    ylab = "dim2", type = "n")
for (j in 1:26) {
    points <- author.colsim2[(nsim*(j-1)+1):(nsim*j),]
    # en todos estos mapas invertimos la segunda coordenada
    points[,2] <- -points[,2]
    covpoints <- cov(points)
    meanpoints <- apply(points, 2, mean)
    lines(ellipse(covpoints, centre = meanpoints))
    text(author.col[j,1], -author.col[j,2], letters[j],
          font = 2)
    }
```

Para reproducir las elipses de confianza del mapa de la imagen 25.3 obtenidas a partir del método Delta, necesitamos la matriz de covarianzas de las coordenadas principales estimadas. La podemos obtener utilizando el programa SPSS. En la página de la red CARME, www.carme-n.org, podemos obtener más detalles y programas adicionales.

Pruebas de permutaciones

Vamos a realizar una prueba de permutación con los datos sobre los autores. Para ello, consultamos un listado de todas las $11 \times 9 \times 7 \times 5 \times 3 = 10395$ combinaciones posibles de los pares libro-autor de los mapas del AC. Luego calcularemos las sumas de las distancias entre los pares del mismo autor en el mapa del AC. En la página web podemos encontrar el programa R para la obtención de todas las combinaciones posibles. El programa hace un listado de los 11 primeros pares posibles {(1,2), (1,3), (1,4), ..., (1,12)}, luego un listado de los 9 pares posibles de cada uno de ellos, por ejemplo para (1,2) son {(3,4), (3,5), ..., (3,12)}, luego un listado de los 7 pares posibles de cada uno de éstos últimos, y así suce-

Imagen B.6:
Distribución exacta, suponiendo cierta la hipótesis nula, del estadístico suma de distancias en la prueba de permutaciones para contrastar la aleatoriedad de las posiciones de los pares de textos del mismo autor en el mapa del AC. El valor observado es el segundo más pequeño de todos los 10395 valores posibles



sivamente. En la imagen B.6 mostramos la distribución de las 10395 distancias posibles en la que hemos señalado la distancia correspondiente a la combinación observada 0,4711. Como vimos en el capítulo 25, no hay otra combinación de pares libro-autor en el mapa del AC bidimensional con una suma de distancias menor. Por tanto, el valor *p* asociado con este valor es igual a 1/10395, es decir, $p < 0{,}0001$. Hicimos una prueba similar en los mapas del AC de subgrupos, que mostramos en las imágenes 21.1 (sólo de consonantes) y 21.2 (sólo de vocales) y obtuvimos 47 y 67 combinaciones a la izquierda del valor observado, por lo que los valores *p* son 48/10395 = 0,0046 y 68/10395 = 0,0065, respectivamente.

Pruebas de permutaciones en ACC

En el ACC nos centramos en la parte del espacio de las variables respuesta (en general en ecología, las especies), que está relacionado linealmente con un determinado conjunto de variables explicativas (en general variables ambientales), como puede verse en el capítulo 24. Pero, ¿cómo saber si las variables respuesta están realmente relacionadas con las variables explicativas? Una medida de la relación existente entre ambos conjuntos de variables es la inercia del espacio restringido. Podríamos situar esta inercia en la distribución de inercias del espacio restringido bajo el supuesto de que no hay relación alguna. Podemos obtener esta distribución permutando al azar los casos (filas) en la matriz de variables explicativas (o variables respuesta). Al permutar aleatoriamente las filas tendrían que perderse las posibles relaciones de éstas con las filas en la matriz de respuestas. Repetimos el ACC y volvemos a calcular la inercia del espacio restringido. Haciendo esto 999 veces (o el número de veces necesarias para poder calcular un valor *p* con suficiente precisión), podemos situar el valor de inercia observa-

do en la distribución para ver si este valor es inusualmente elevado. Si se halla en el 5% de los valores más elevados, consideraremos que la relación entre las variables respuesta y las variables explicativas es estadísticamente significativa. Como antes, podemos estimar el valor de *p* haciendo un recuento del número de valores de la distribución de permutaciones mayores que el valor observado (para ser significativo el valor observado tiene que ser suficientemente elevado). El paquete **vegan** incorpora esta prueba, que podemos obtener aplicando la función `anova()` a `cca()`:

```
anova(cca(bio, env))

Permutation test for cca under reduced model

Model: cca(X = bio, Y = env)
          Df     Chisq         F  N.Perm     Pr(>F)
Model      3    0.2798    1.6696    1300    0.03462  *
Residual   9    0.5028
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En realidad, el estadístico utilizado no es la inercia sino un «seudo» estadístico F, como el del análisis de la varianza (para más detalles podemos consultar la documentación sobre **vegan**). Debemos fijarnos en el valor de *p* que nos va a dar la salida. Así, en este caso el valor del estadístico F para el espacio restringido es significativamente elevado ($p = 0{,}03462$).

Programas R de Fionn Murtagh

En su reciente libro *Correspondence Analysis and Data Coding with Java and* R (véase el apéndice bibliográfico), Fionn Murtagh proporciona muchos programas en R para AC, especialmente para la recodificación de datos. Están disponibles en Internet y se pueden bajar de www.correspondances.info. En concreto, en sus páginas 21 a 26, encontramos el único programa en R que permite hacer la agrupación jerárquica utilizando el método de Ward, con la incorporación de pesos, que es exactamente lo que necesitamos para el capítulo 15. Suponiendo que hayamos sido capaces de bajar el programa del web mencionado, que hemos leído la tabla de datos de la imagen 15.3 y que la hemos guardado en el *data frame* `food`, podremos realizar el análisis de grupos de los perfiles de las filas que mostramos en la figura de la imagen 15.5, utilizando la función `hierclust()` de Murtagh de la siguiente manera:

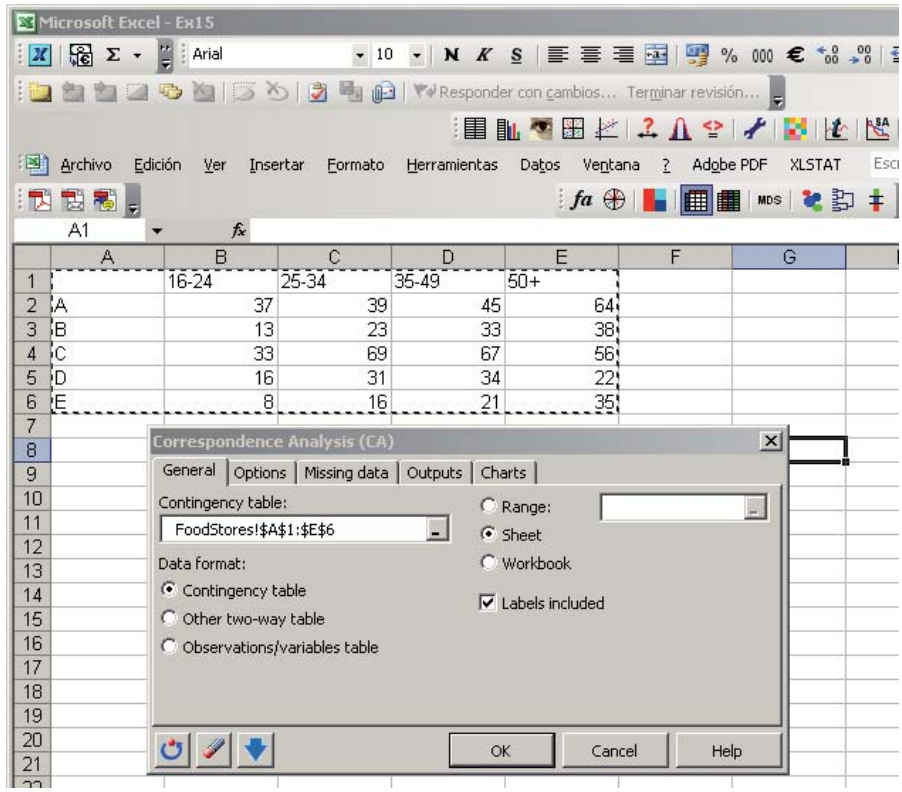
```
food.rpro <- food/apply(food,1,sum)
food.r <- apply(food,1, sum)/sum(food)
food.rclust <- hierclust(food.rpro, food.r)
plot(as.dendrogram(food.rclust))
```

XLSTAT

Para llevar a cabo los análisis de este libro (y análisis adicionales), una de las mejores alternativas es el programa estadístico XLSTAT (www.xlstat.com) derivado de Excel. En XLSTAT, los programas para el AC y el ACM incluyen los ajustes de inercia del ACM y el análisis de subgrupos del AC y del ACM. También incluye un programa para el ACC que incorpora la prueba de la permutación para contrastar que las variables explicativas están significativamente relacionadas con lo ejes principales de la solución restringida. Otros programas para el análisis multivariante de XLSTAT son el análisis de componentes principales, el análisis factorial, el análisis discriminante, el análisis de grupos, la regresión de mínimos cuadrados parcial y el análisis de Procrustes generalizado. Dado que el programa opera en el entorno Excel es muy fácil de utilizar. Por ejemplo, para ejecutar el AC con los datos `food` que hemos utilizado anteriormente en el análisis jerárquico de grupos, clicamos sobre el icono de AC y seleccionamos la tabla (con etiquetas para filas y columnas) que queremos analizar (imagen B.7).

El menú de «Options» permite seleccionar puntos adicionales o subgrupos de datos. El menú «Missing data» permite varias opciones para el tratamiento de los

Imagen B.7:
Menú de XLSTAT para ejecutar el AC en una tabla seleccionada en Excel



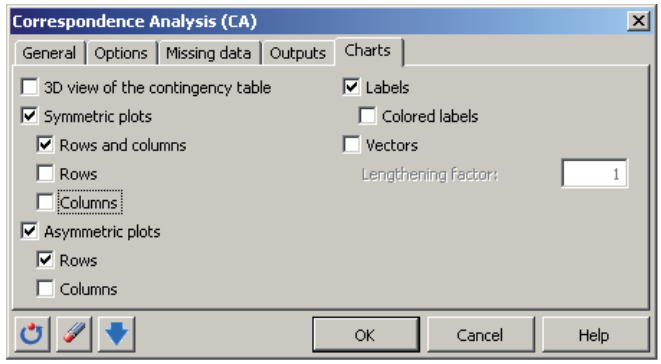


Imagen B.8:
Menú de XLSTAT para seleccionar opciones gráficas del AC

valores perdidos. El menú «Outputs» permite seleccionar varias tablas numéricas (perfiles, distancias χ^2, coordenadas principales, coordenadas estándares, contribuciones, correlaciones al cuadrado, etc.). El menú «Charts» permite llevar a cabo diversos mapas del AC. En la imagen B.8 mostramos cómo llevar a cabo un

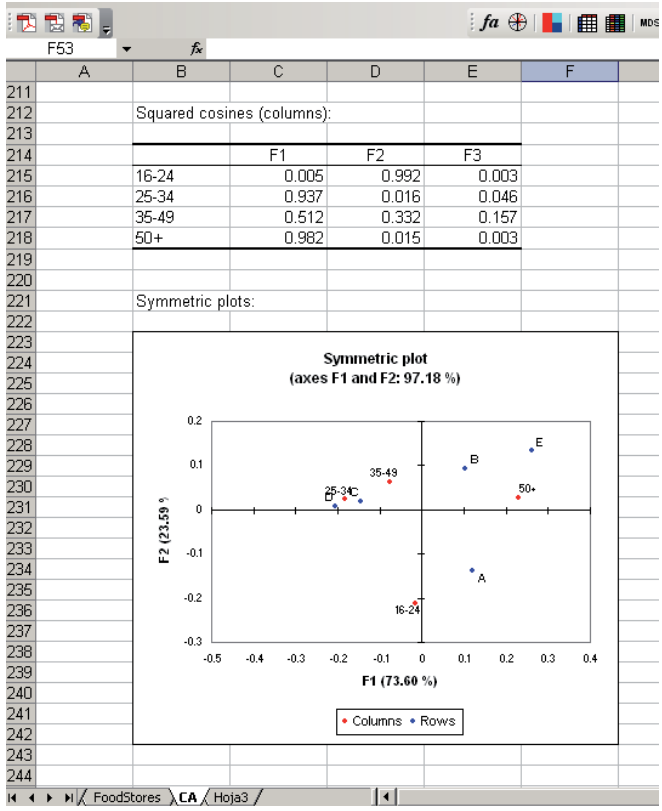


Imagen B.9:
Parte del resultado del programa de AC de XLSTAT, que se proporciona en una hoja de cálculo adicional

mapa simétrico de filas y de columnas, y un mapa asimétrico de filas (es decir, la opción `"rowprincipal"` del paquete **ca**). En la imagen B.9 mostramos parte de estos resultados.

En el módulo de análisis de grupos de XLSTAT podemos desarrollar el análisis de grupos que vimos en el capítulo 15, ya que permite asignar pesos a los puntos; por tanto, podemos realizar la agrupación de Ward de perfiles ponderados con sus masas.

Opciones gráficas

Crear un mapa de AC con determinadas características listo para ser publicado, no es trivial. En esta sección describimos los tres procedimientos utilizados para la obtención de los gráficos de este libro.

Gráficos con LaTeX

La composición tipográfica de la edición en inglés de este libro fue realizada con LaTeX. Con LaTeX —y algunas macros de este programa— podemos crear directamente mapas sin tener que utilizar otros paquetes gráficos. Así, realizamos la mayor parte de los mapas del libro utilizando la macro PicTeX. Como ejemplo de mapa creado en LaTeX, a continuación mostramos el programa que utilizamos para crear el mapa asimétrico correspondiente a los datos de los fumadores que mostramos en la imagen 9.2:

```
\beginpicture
\setcoordinatesystem units <2.5cm,2.5cm>
\setplotarea x from -2.40 to 1.70, y from -1.6 to 2.25
\accountingoff
\gray
\setdashes <5pt,4pt>
\putrule from 0 0 to 1.7 0
\putrule from 0 0 to -1.4 0
\putrule from 0 0 to 0 2.25
\putrule from 0 0 to 0 -1.6
\put {+} at 0 0
\black
\small
\put {Axis 1} [Br] < -.2cm,.15cm> at 1.70 0
\put {0.0748 (87.8\%)} [tr] < -.2cm,-.15cm> at 1.70 0
\put {Axis 2} [Br] < -.1cm,-.4cm> at 0 2.25
\put {0.0100 (11.8\%)} [Bl] <.1cm,-.4cm> at 0 2.25
\setsolid
\putrule from 1.3 -1.3 to 1.4 -1.3
\putrule from 1.3 -1.32 to 1.3 -1.28
\putrule from 1.4 -1.32 to 1.4 -1.28
\put {\it scale} [b] <0cm,.25cm> at 1.35 -1.3
\put {0.1} [t] <0cm,-.2cm> at 1.35 -1.3
\multiput {$\bullet$} at
```

```
 0.06577 0.19373
-0.25896 0.24330
 0.38059 0.01066
-0.23295 -0.05775
 0.20109 -0.07891
 /
\sf
\put {SM} [l] <.15cm,0cm> at   0.06577  0.19373
\put {JM} [r] < -.15cm,0cm> at -0.25896  0.24330
\put {SE} [bl] <.15cm,0cm> at  0.38059  0.01066
\put {JE} [r] < -.15cm,0cm> at -0.23295 -0.05775
\put {SC} [tl] <.15cm,0cm> at  0.20109 -0.07891
\gray
\multiput {$\circ$} at
 1.4384  0.3046
-0.3638 -1.4094
-0.7180 -0.0735
-1.0745  1.9760
 /
\sl
\put {none} [b] <0cm,.2cm> at     1.4384  0.3046
\put {light} [b] <0cm,.2cm> at   -0.3638 -1.4094
\put {medium} [T] <0cm,-.3cm> at -0.7180 -0.0735
\put {heavy} [b] <0cm,.2cm> at   -1.0745  1.9760
\black
\endpicture
```

El programa anterior permite darnos cuenta de que crear un mapa como el de la imagen 9.2 es bastante laborioso. Hay que situar con mucha precisión cada uno de los puntos y de las líneas del mapa. Una ventaja es que podemos asegurar que la razón de escalas del mapa es exactamente 1. Por ejemplo, en este mapa hemos establecido que las unidades en los ejes de coordenadas vertical y horizontal sean exactamente iguales (`2,5 cm`).

Gráficos en Excel

Muchos de los mapas nuevos de esta segunda edición los hemos hecho en Excel, a partir de los resultados del análisis estadístico en XLSTAT. Sin embargo, para asegurar que la razón de escalas de los mapas de los capítulos 17 a 19 fuera correcta, tuvimos que efectuar algunos ajustes. Así tuvimos que redefinir los valores máximos y mínimos de los ejes y alargar vertical u horizontalmente los mapas. Luego los copiamos en metaarchivos y los pegamos en el *Adobe Ilustrator*. Al realizar esta operación se modifica algo la razón de escalas del mapa, las unidades verticales aumentan algo más que las horizontales. Por tanto, de nuevo, tuvimos que retocar los mapas. Posteriormente los guardamos en formato *PostScript Encapsulado* (EPS), para finalmente incorporarlos al texto como archivos LaTeX utilizando la instrucción `\includegraphics`, por ejemplo:

```
\begin{figure}[h]
\center{\includegraphics[width=10cm,keepaspectratio]{Ex18_5.eps}}
\caption{\sl MCA map of Burt matrix of four questions on women
         working, showing first and second dimensions;
         total inertia = 1.145, percentage inertia in map: 65.0\%.}
\end{figure}
```

Gráficos en R

En este libro también hemos creado muchos mapas en R. Por ejemplo, los del capítulo 25. Para incorporarlos al texto, primero los copiamos como metaarchivos, luego los pegamos en el *Adobe Ilustrator*, los terminamos de ajustar para asegurar que la razón de escalas fuera correcta y finalmente los guardamos en formato EPS.

Bibliografía sobre análisis de correspondencias

El principal objetivo de este libro es la enseñanza del AC. Para dar mayor énfasis a la orientación didáctica, en el texto de los capítulos no hemos incluido referencias bibliográficas. Sin embargo, para que el lector pueda seguir avanzando en el conocimiento del AC, en esta sección reseñamos las principales fuentes bibliográficas. También hemos indicado dónde podemos hallar revisiones bibliográficas más amplias, así como referencias históricas sobre este método.

La escuela de Benzécri de análisis de datos

A pesar de que la teoría del AC data de principios del siglo XX, el enfoque sobre el AC que presentamos en este libro tiene su origen en el trabajo de Jean-Paul Benzécri y colaboradores en Francia en los años sesenta, que fue publicado en los dos volúmenes de *Analyse des Données* (Análisis de datos).

- Benzécri J.P. et al. *Analyse des Donnés. Tôme 1: La Classification. Tôme 2: L'Analyse des Correspondances.* París: Dunod, 1973.

Sin embargo, estos libros son poco accesibles para lectores que no estén familiarizados con la particular notación de Benzécri, distinta de la notación matricial habitual más práctica. Para dar a conocer a la comunidad de habla inglesa las ideas de Benzécri, apareció una traducción al inglés de estos libros que, no obstante, tuvo poco éxito. El libro de Le Roux y Rouanet que siguió a esta traducción indicaba de forma clara la aproximación de Benzécri al análisis de conjuntos de datos grandes, el «análisis de datos geométrico», pero los autores siguieron manteniendo una compleja notación que dificultaba su comprensión.

- Le Roux B. y H. Rouanet. *Geometric Data Analysis: From Correspondence Analysis to Structured Data.* Dordrecht: Kluwer, 2004.

El libro de Fionn Murtagh, también estudiante de Benzécri, es una de las mejores publicaciones en inglés para comprender el trabajo del maestro. No solamente da a conocer buena parte de la filosofía de Benzécri (incluye un prefacio del propio Benzécri traducido al inglés), sino que es un libro con una innovadora aproximación, muy orientado hacia el cálculo, proporcionando muchas aplicaciones interesantes, así como muchos detalles sobre la programación en R.

- Murtagh F. *Correspondence Analysis and Data Coding with Java and R.* Londres: Chapman & Hall/CRC, 2005.

El trabajo de Brigitte Escofier, uno de los líderes y miembros más innovadores del grupo de Benzécri, se publicó póstumamente con una colección de sus artículos más importantes:

- Escofier B. *Analyse des Correspondances: Recherches au Coeur de l'Analyse des Données.* Rennes, Francia: Presses Universitaires des Rennes, 2003.

Los dos libros en inglés de 1984

En 1984 aparecieron, casi simultáneamente, dos libros en inglés sobre AC que, gracias a la utilización de una notación más convencional, expresaban de forma más comprensible el trabajo de Benzécri.

- Lebart L., A. Morineau y K. Warwick. *Multivariate Descriptive Statistical Analysis.* Chichester: Wiley, 1984.

- Greenacre M.J. *Theory an Applications of Correspondence Analysis.* Londres: Academic Press, 1984.

Aunque ambos libros están agotados, es recomendable consultarlos. El libro de Lebart y colaboradores proporciona una descripción menos detallada sobre el AC. Sin embargo, ofrece una amplia panorámica sobre su utilización en el contexto de las encuestas a gran escala. El libro de Greenacre trata de dar a la vez una visión teórica y práctica del método. Estos dos libros incluyen una amplia revisión bibliográfica sobre el trabajo realizado hasta aquel momento.

El sistema Gifi

Con el seudónimo de Albert Gifi se encuentra el grupo de los Países Bajos, dirigido por Jan de Leeuw. Es el grupo que fuera de Francia ha llevado a cabo un desarrollo más importante sobre el AC. Todavía hoy sigue siendo el grupo más activo. Este grupo ha explorado, principalmente, la utilización del ACM, llamado *análisis de homogeneidad,* como una técnica de cuantificación que partiendo del análisis multivariante clásico lleva a cabo una generalización no lineal de los métodos multivariantes. Su trabajo se describe ampliamente en el libro:

- Gifi A. *Nonlinear Multivariate Analysis.* Chichester: Wiley, 1990.

Como excelente resumen del «método Gifi» podemos consultar:

- Michalidis G., J. de Leeuw. «The Gifi system for descriptive multivariate analysis». *Statistical Science* **13** (1998): 307-336. (Disponible en Internet, consúltese Google.)

La escuela japonesa

Fundada por Chikio Hayashi, este grupo desarrolló en paralelo con las escuelas francesa y holandesa, un sistema equivalente de «cuantificación de datos cualitativos», el *escalado dual,* muy impregnado de sus propios referentes culturales. Varios

libros de Shizuhiko Nishisato describen esta aproximación, muy concentrada en las propiedades algebraicas de los valores de escala cuantificados. De todas formas, el último libro de Nishasato contiene muchas representaciones gráficas:

- Nishisato S. *Multivariate Nonlinear Descriptive Analysis*. Londres: Chapman & Hall/CRC, 2006.

El libro de Nishisato contiene muchas referencias históricas y una extensa lista de referencias sobre la literatura relacionada con el AC. Sin embargo, no contiene detalles de cálculo.

Los libros de Barcelona y de Colonia

En 1991, 1995, 1999 (en el «Archivo Central para Investigación Social Empírica, en Colonia»), en 2003 (en la Universidad Pompeu Fabra, en Barcelona) y en 2007 (en la Universidad Erasmus de Róterdam) tuvieron lugar conferencias internacionales cuyo tema principal fue el AC. Como resultado de estas conferencias, se escribieron tres libros colectivos con la participación de estadísticos y científicos sociales, que reflejaban el desarrollo de la teoría y la práctica del AC y métodos relacionados:

- Greenacre M.J. y J. Blasius, J. (eds.). *Correspondence Analysis in the Social Sciences*. Londres: Academic Press, 1994.

- Blasius J. y M.J. Greenacre (eds.). *Visualizing Categorical Data*. San Diego: Academic Press, 1998.

- Greenacre M.J. y J. Blasius, J. (eds.). *Multiple Correspondence Analysis and Related Methods*. Londres: Chapman & Hall/CRC, 2006.

Para profundizar en el tema, recomendamos vivamente estos tres volúmenes, a los que han contribuido más de 100 autores. El tercer volumen está especialmente orientado a las necesidades de cálculo y, muchos autores proporcionan recursos de software para el cálculo.

La relación con R

El objetivo de esta segunda edición de *Correspondence Analysis in Practice* no es sólo presentar de forma didácticamente estructurada un texto sobre el AC, sino también permitir a los lectores llevar a cabo sus propios análisis, principalmente utilizando el sistema de programación R. Situados a principios de siglo XXI, el software R se ha convertido en el sistema estándar de cálculo estadístico. En comparación con la primera edición, aparte de la inclusión de temas adicionales, el cambio más significativo es el apéndice de cálculo de 46 páginas. Entre los muchos libros sobre programación en R que podríamos recomendar para alguien que empiece, descatamos:

- Crawley M. *Statistics: An Introduction using R*. Chichester: Wiley, 2005.

Además, contiene una magnífica introducción a la metodología estadística moderna.

En el siguiente artículo se ve en detalle el paquete **ca** para R:

- Nenadić O. y M.J. Greenacre. «Correspondence Analysis in R, with Two- and Three-dimensional Graphics. The **ca** Package». *Journal of Statistical Sofware* 20, 3 (Feb 2007). (Disponible en Internet en http://www.jstatsoft.org.)

Recursos en la red

Para obtener más información y más software sobre AC y métodos relacionados, podemos consultar las siguientes páginas web (no comerciales):

http://www.carme-n.org

(una red sobre análisis de correspondencia y métodos relacionados, con programas en R y datos de la segunda edición de *Correspondence Analysis in Practice*),

http://gifi.stat.ucla.edu

(la página web de Jan de Leeuw para el sistema Gifi y funciones en R),

http://www.correspondances.info

(la página web personal de Fionn Murtagh para su libro, con programas en R y datos),

http://www.math.yorku.ca/SCS/friendly.html

(la página web personal de Michael Friendly para gráficos de datos categóricos),

http://www.imperial.ac.uk/bio/research/crawley/statistics

(contiene material del libro de Michael Crawley, *Statistics: An Introduction using* R),

http://www.gesis.org/en/za

(la página web del «Archivo Central para Investigación Social Empírica de Colonia», con enlaces a varias encuestas sociales incluyendo las del ISSP-International Social Survey Program, el Programa Internacional sobre Encuestas Sociales),

http://www.r-project.org

(el proyecto R para cálculo estadístico),

http://cc.oulu.fi/jarioksa/softhelp/vegan.html

(la página web de Jari Oksaken para el paquete **vegan** en R),

http://www.people.few.eur.nl/groenen/mmds/datasets

(la página web con datos del libro: *Modern Multidimensional Scaling* de Ingwer Borg y Patrick Groenen).

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.

Epílogo

En este libro hemos presentado el análisis de correspondencias (AC) como un método versátil para la visualización de datos, aplicable a una amplia variedad de situaciones. Este epílogo tiene como objeto avanzar algo más en el análisis de algunos aspectos de este método que aparecen con frecuencia en discusiones sobre AC, así como aportar algunas consideraciones personales.

Mapas simétricos

La interpretación de los mapas simétricos, aunque es una opción más de los mapas de AC, sigue siendo uno de los aspectos más controvertidos de este método. Este tipo de mapas expresan tanto las filas como las columnas en coordenadas principales; es decir, a pesar de que las proyecciones de los perfiles fila y los perfiles columna ocupan espacios distintos, mostramos sus proyecciones en un mismo mapa. Hemos visto (por ejemplo, en los capítulos 9 y 10) que la diferencia entre los mapas simétricos y los asimétricos (en los que todos los puntos se hallan en el mismo espacio) es el factor de escala de los ejes principales, la raíz cuadrada de sus respectivas inercias principales. Por tanto, las direcciones, indicadas por los puntos en coordenadas principales y por sus homólogos en coordenadas estándares, son casi iguales cuando las raíces cuadradas de las inercias principales no son muy distintas; así, podemos ver un ejemplo en el mapa de la imagen 13.4 en la que los ejes del biplot, que pasan a través de los vértices, casi coinciden con los puntos correspondientes a los perfiles. En tales casos, la forma de interpretar los mapas simétricos y los asimétricos como si fueran biplots es válida. Sin embargo, si las raíces cuadradas de las inercias principales son muy distintas, al interpretar los mapas simétricos como si fueran un biplot pueden aparecer problemas; lo podemos ver, por ejemplo, en las diferentes direcciones definidas por las categorías de fumadores en los mapas de las imágenes 9.2 y 9.5. Aun así, como se pone de manifiesto en el artículo de Gabriel que mencionamos a continuación, la distorsión que se produce al interpretar los mapas simétricos como si fueran verdaderos biplots, no es demasiado grande.

- Gabriel K.R. «Goodness of Fit of Biplots and Correspondence Analysis». *Biometrika* 89 (2002): 423-436.

Esto significa que el debate sobre las diferencias de escala es más bien un tema académico. Toda la discusión que ha generado este tema tiene poco interés cuando se trata de aplicar el AC. En mi opinión, el mapa simétrico sigue sien-

do, por defecto, el mejor mapa. De hecho, es la opción que aparece por defecto en nuestro paquete **ca** para R. Si interpretamos de forma asimétrica la matriz de datos, en la que la filas representen «unidades observacionales» (como, por ejemplo, individuos en estudios sociales, localidades de muestreo en ecología o en arqueología, o textos en lingüística, etc.) y las columnas representen «variables» (como, por ejemplo, las respuestas categóricas en sociología, las especies en ecología, los artefactos en arqueología, o los indicadores de estilo en lingüística, etc.), el biplot estándar del AC es una buena alternativa. Representa de forma óptima las distancias entre unidades y permite una interpretación tipo biplot válida de las unidades proyectadas sobre las direcciones de las variables. Además, las longitudes de los vectores (variables) tienen una interpretación clara.

«No puedes comerte un pastel y, al mismo tiempo, conservarlo»

Desgraciadamente, en el contexto que nos ocupa, se cumple este dicho inglés. Podemos decir lo mismo de la expresión: «En la vida, no lo puedes tener todo». Sería maravilloso que en un solo mapa pudiéramos representar de forma óptima e interpretar los tres elementos siguientes:

1. Las distancias entre perfiles fila.
2. Las distancias entre perfiles columna.
3. Los productos escalares entre filas y columnas, que reconstruyen los datos originales (es decir, el biplot).

Sin embargo, la realidad es que, al mismo tiempo y como máximo, podemos tener representados óptimamente sólo dos de los tres elementos anteriores. Los mapas simétricos representan óptimamente las distancias ji-cuadrado entre los perfiles fila y entre los perfiles columna. Por tanto, podemos interpretar las distancias entre filas y las distancias entre columnas (es decir, se cumplen los puntos 1 y 2). No podemos interpretar de forma óptima las relaciones entre filas y columnas. Sin embargo, teniendo en cuenta las observaciones del párrafo anterior, las podemos interpretar con una seguridad razonable. En los mapas asimétricos representamos de forma óptima, por ejemplo, los perfiles fila, mientras que los vértices columna proporcionan los perfiles extremos como puntos de referencia. Sus proyecciones sobre los ejes del biplot nos permiten interpretar de forma óptima las relaciones entre filas y columnas (es decir, se cumplen los puntos 1 y 3). Los biplots estándares del AC son una variante de los mapas asimétricos que muestran, por ejemplo, los perfiles fila, al mismo tiempo que acercan los vértices columna, multiplicando por la raíz cuadrada de sus masas, para mejorar la representación conjunta (es decir, se cumplen 1 y 3). En este último biplot, podemos relacionar las proyecciones de los vectores columna sobre los ejes del biplot con sus contribuciones a los ejes principales (capítulo 13).

Normalización simétrica en SPSS

Aparte del programa libre R, y del programa comercial XLSTAT que hemos descrito en el apéndice de cálculo, todavía no hemos comentado nada sobre otros softwares que incluyen el AC. Entre estos programas encontramos Minitab, Stata, Statistica, SPAD, SAS y SPSS. Dado que SPSS es ampliamente utilizado, es conveniente que hagamos algunos comentarios sobre esta opción. En el módulo *Categories* del programa de AC del SPSS, se proporciona un biplot llamado *symmetrical normalization* que no hemos visto en este libro. Podríamos confundir dicho biplot con el mapa simétrico que sí hemos descrito. Sin embargo, no se trata de lo mismo, ya que el primero presenta las coordenadas estándares multiplicadas por las raíces cuadradas de los valores singulares (es decir, la raíz cuarta de las inercias principales) y no por los valores singulares. Dicho de otro modo —con relación a los pasos (A.8) y (A.9) del algoritmo básico de cálculo del AC que vimos en la página 267—, este procedimiento calcula $\mathbf{\Phi D}_\alpha^{\frac{1}{2}}$ y $\mathbf{\Gamma D}_\alpha^{\frac{1}{2}}$ en vez de $\mathbf{\Phi D}_\alpha$ y $\mathbf{\Gamma D}_\alpha$ como en los mapas simétricos. Por tanto, la «normalización simétrica» del SPSS proporciona una representación óptima de los productos escalares, pero no proporciona una representación óptima de distancias, ya que ni filas ni columnas se expresan en coordenadas principales. Por tanto, esta representación gráfica proporciona sólo uno de los tres elementos mencionados anteriormente (se cumple 3, pero ni 1 ni 2). A pesar de que la diferencia entre esta representación gráfica y el mapa simétrico es sólo un tema de factores de escala en los dos ejes —que en la mayoría de casos son difícilmente distinguibles para un observador no experimentado—, no recomendamos la utilización de este mapa ya que no aporta beneficio alguno (en realidad representa una pérdida) con relación a las otras opciones existentes. Si las inercias principales de los dos ejes son similares, entonces, como vimos anteriormente, las posiciones relativas de los puntos en la «normalización simétrica» son prácticamente idénticas a las del mapa simétrico. Sin embargo, es preferible el mapa asimétrico ya que representa las distancias ji-cuadrado en su verdadera escala. El mapa con «normalización simétrica» lo denominamos *symmetric biplot*, y es una de las posibilidades de nuestro paquete **ca** de R. Para obtenerlo escribiremos: `map="symbiplot"` (págs. 304-305). Curiosamente, en las últimas versiones de SPSS no era posible representar un mapa simétrico, una de las representaciones gráficas más populares entre los investigadores franceses. Sigue siendo imposible en las últimas versiones del programa obtener un mapa conjunto de filas y columnas en coordenadas principales. La mejor opción es seleccionar la normalización «principal», que proporciona los valores numéricos de las coordenadas principales de filas y columnas. Sin embargo, el programa siempre rechaza el realizar un mapa conjunto con estos datos, prefiere mapas separados. A no ser que los datos originales del usuario se hallen en formato SPSS, como decíamos, no recomendamos el programa del AC de SPSS. Sin embargo, dentro del módulo *Categories*, resultan muy útiles para ciencias sociales el programa de optimización de escalas para análisis de correspondencias múltiples (llamado, en versiones anteriores, análisis de homogeneidad) y el de análisis de componentes principales no lineal (CatPCA).

Categorías poco frecuentes

El efecto de las categorías poco frecuentes sobre la distancia χ^2 y sobre el resultado del AC es también un tema que ha generado mucha discusión, especialmente entre los investigadores en ecología, casi siempre sin justificación. Por ejemplo, según C.R. Rao, «la distancia ji-cuadrado que utiliza proporciones marginales en el denominador otorga al medir las afinidades entre perfiles, demasiada importancia a las categorías con bajas frecuencias» (en pág. 42 del siguiente artículo):

- Rao C.R. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Quëstiió* 19 (1995): 23-63. Disponible en Internet en:

  www.idescat.net/sort/questiio/questiiopdf/19.1,2,3.1.radhakrishna.pdf

Sin embargo, la realidad es que en AC ponderamos cada categoría proporcionalmente a su masa, lo que reduce el papel de las categorías de baja frecuencia. Lo podemos ver de forma muy simple analizando las contribuciones numéricas de las distintas categorías a los ejes principales. Así, podemos constatar que las categorías poco frecuentes tienen, en general, poca influencia sobre la solución hallada; es decir, la solución sería casi la misma si eliminásemos estas categorías del análisis.

Consideremos, a título ilustrativo los datos sobre abundancia de especies del capítulo 10 (pág. 109) con los que calculamos la abundancia relativa de las 10 especies más frecuentes y la de las 10 menos abundantes, y lo comparamos con sus contribuciones relativas, en porcentaje, a los dos primeros ejes del mapa de AC de la imagen 10.5. Los resultados son los siguientes:

		Contribución a los ejes	
Especies	*Abundancia relativa*	Eje 1	Eje 2
10 más abundantes	74,6%	77,3%	89,3%
10 menos abundantes	0,4%	0,8%	0,5%

Estos cálculos ilustran que las especies poco frecuentes no contribuyen demasiado a la solución bidimensional, pues las contribuciones se hallan mucho más en la línea con las abundancias de cada grupo de especies. Según nuestra experiencia, sólo de vez en cuando, las categorías poco frecuentes contribuyen de forma excesiva a los ejes principales. En tales casos, debemos eliminarlas o combinarlas con otras categorías. Esta situación se da en estudios sociológicos, en los que las categorías de baja frecuencia, como los valores perdidos, coinciden en el mismo grupo de encuestados. Estas categorías pueden dominar la solución del ACM, a menudo definiendo el primer eje. Lo vimos, en los mapas de las imágenes 18.2 y 18.5. Podemos rectificar esta situación mediante un análisis de subgrupos o

combinando, de forma razonable, las respuestas correspondientes a categorías de baja frecuencia con otras similares. En ecología se produciría una situación análoga cuando determinadas especies poco frecuentes se hallaran simultáneamente en las mismas muestras. Sin embargo, no se trata de una situación común; en general, las especies poco frecuentes ocurren de forma aleatoria en distintas muestras.

Las categorías de baja frecuencia son, a menudo, observaciones atípicas

A menudo, las filas y columnas con frecuencias bajas son observaciones atípicas con extraños perfiles. Probablemente por este motivo llaman la atención y dan la impresión de que pueden afectar, de forma importante, al análisis. Sin embargo, como hemos dicho, tienen en general poca influencia sobre la solución del AC debido a su escasa masa. Además, según hemos mostrado en el capítulo 13 y mencionamos anteriormente, el biplot estándar del AC podría solucionar este problema, ya que «acerca» estos puntos a razón de la raíz cuadrada de sus masas, lo que en la práctica implica una eliminación de las observaciones atípicas de poca frecuencia. Ello también constituye una ilustración gráfica de su escaso efecto sobre la configuración de los ejes principales.

La distancia χ^2 es una distancia de Mahalanobis

Este apartado es algo técnico, aunque resulta útil para que el lector formado estadísticamente pueda comprender que la distancia ji-cuadrado, aparte de ser la clave de todas las propiedades del AC, es una distancia estadística apropiada. Matricialmente, podemos expresar la distancia euclídea ponderada como:

$$\text{distancia euclídea ponderada} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\mathbf{D}_w(\mathbf{x}-\mathbf{y})} \qquad \text{(E.1)}$$

donde $\boldsymbol{x}$ e $\boldsymbol{y}$ son vectores con elementos x_j y y_j, $j = 1, ..., J$, $^{\mathsf{T}}$ indica la transposición de una matriz o de un vector, y $\mathbf{D}_w$ es la matriz digonal que contiene los factores de ponderación w_j. Podemos suponer que las filas de una tabla de contingencia corresponden a una variable aleatoria *multinomial*. La distribución multinomial es una generalización de la distribución binomial. Constituye un modelo para la descripción del comportamiento de datos muestreados de poblaciones con probabilidades p_j, $j = 1, ..., J$ para cada uno de los J grupos. Por ejemplo, los tres tipos de lectores del capítulo 3 (tabla de la imagen 3.1). A partir de la hipótesis nula de que hemos muestreado los datos en la misma población, los cinco niveles educativos de este conjunto de datos serían muestras multinomiales de la población con probabilidades p_1, p_2, p_3 en las que las estimaciones de p_j de los tres grupos son los elementos del perfil medio $\hat{p}_1 = c_1 = 0{,}183$, $\hat{p}_2 = c_2 = 0{,}413$ y $\hat{p}_3 = c_3 = 0{,}404$ (última fila de la tabla de la imagen 3.1). La *distancia de Mahalanobis* es la distancia clásica utilizada para datos multivariantes agrupados. Se basa en la inversa de la matriz de covarianzas de las variables:

$$\text{distancia de Mahalanobis} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\,\Sigma^{-1}(\mathbf{x}-\mathbf{y})} \qquad \text{(E.2)}$$

excepto por el hecho de que implica una matriz cuadrada completa de pesos Σ^{-1}, y no una matriz diagonal, tiene el aspecto de una distancia euclídea ponderada (E.1). Para una distribución multinomial, la matriz de covarianzas Σ tiene una forma simple. Por ejemplo, en nuestro caso trinomial $J = 3$ (los resultados serían similares para cualquier número de grupos):

$$\Sigma = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{bmatrix} = \mathbf{D}_p - \mathbf{p}\mathbf{p}^\top \qquad \text{(E.3)}$$

donde $\mathbf{p}$ es el vector de las p_j y $\mathbf{D}_p$ la correspondiente matriz diagonal. Estimamos (E.3) sustituyendo las probabilidades p_j por sus estimaciones c_j. No es posible invertir la matriz de covarianzas Σ de la forma habitual, ya que se trata de una matriz singular. Por tanto no podemos hallar una matriz Σ^{-1} tal que $\Sigma\Sigma^{-1} = \mathbf{I}$. Una manera de sortear este problema es eliminar una de las categorías y seguir con sólo $J - 1$ categorías. No obstante, calquiera que sea la categoría que se omita, la distancia de Mahalanobis será la misma. Una aproximación alternativa más elegante, completamente equivalente pero que utiliza las J categorías, consiste en utilizar la *generalización inversa*, simbolizada como Σ^-, que tiene la propiedad de que $\Sigma\Sigma^-\Sigma = \Sigma$ (la *inversa de Moore-Penrose*). La inversa generalizada de Moore-Penrose de (E.3) es igual a:

$$\Sigma^- = \begin{bmatrix} 1/p_1 & 0 & 0 \\ 0 & 1/p_2 & 0 \\ 0 & 0 & 1/p_3 \end{bmatrix} = \mathbf{D}_p^{-1} \qquad \text{(E.4)}$$

Es decir, la distancia χ^2 estima de forma exacta la distancia de Mahalanobis (E.2). Aquí la situación es similar a la del análisis discriminante lineal: para maximizar la discriminación entre grupos, suponemos que los grupos tienen matrices de covarianzas iguales, lo que en el caso multinomial equivale a asumir el modelo de independencia y que los vectores se hallan en un espacio de Mahalanobis, que equivale a un espacio χ^2.

Rotación de las soluciones

En este libro no hemos visto nada sobre rotaciones debido a que raramente se justifican o se necesitan en AC. Debemos tener en cuenta que el espacio de perfiles no es un espacio de vectores real ilimitado, es un espacio delimitado por puntos unidad o vértices, que definen un simplex en un espacio multidimensional. La idea de alinear los puntos de las distintas categorías en ejes que formen ángulos rectos no tiene, en nuestro contexto, el mismo significado que en el análisis factorial en que los ángulos rectos indican que las correlaciones entre variables son cero (recordemos que en AC, la suma de los elementos del perfil es 1; por tanto, la posición de un determinado punto viene determinada por las de los restantes puntos). Las rotaciones pueden ser apropiadas en algunos contextos como el

ACM y en ACP no lineal (que no hemos visto en este libro) cuando analizamos varias variables simultáneamente. Por ejemplo, en ACM ocurre con frecuencia que los puntos correspondientes a las no respuestas se hallan juntos —mostrando así una elevada asociación dentro del conjunto de datos— y que, sin embargo, su posición no coincida con ningún eje principal. En tal caso podría tener interés hacer girar los ejes para separar el efecto de los puntos de no respuesta de los restantes. De todas formas, podemos solucionar mejor este problema haciendo un análisis de subgrupos (capítulo 21), que permite ignorar las no respuestas y concentrar el analisis en las respuestas sustantivas. En cualquier caso, si queremos llevar a cabo una rotación, deberemos tener en cuenta las masas de las categorías. Una posibilidad podría ser una versión ponderada de la rotación varimax del análisis factorial cuyo (para el caso de las columnas) criterio de maximización sería:

$$\sum_j \sum_k c_j^2 \left(\tilde{y}_{jk}^2 - \frac{1}{J} \sum_{j'} \tilde{y}_{j'k}^2 \right)^2 \qquad \text{(E.5)}$$

donde $\tilde{y}_{jk}$ es la coordenada estándar rotada, es decir, el *(j,k)*-ésimo elemento de $\tilde{\mathbf{Y}} = \mathbf{YQ}$, siendo $\mathbf{Q}$ una matriz ortogonal de rotación. Fijémonos en que las masas c_j se hallan al cuadrado ya que la función objetivo implica la cuarta potencia de las coordenadas. Dado que $c_j \tilde{y}_{jk}^2 = (c_j^{\frac{1}{2}} \tilde{y}_{jk})^2$, sugerimos una alternativa casi idéntica, que deriva de una pequeña modificación del usual criterio varimax: llevar a cabo una rotación (sin ponderar) con las coordenadas estándares recalibrabas $c_j^{\frac{1}{2}} y_{jk}$, que son exactamente las mismas utilizadas en el biplot estándar del AC. Es decir, rotar la solución para concentrar (o, concretar, en terminología del análisis factorial) las contribuciones de las categorías sobre los ejes rotados.

AC y modelización

En el capítulo 13 vimos el AC en K^* dimensiones como una descomposición que se puede expresar de la siguiente manera [véanse (13.4) y (A.14) en el apéndice teórico]:

$$p_{ij} = r_i c_j + r_i c_j \left(\sum_{k=1}^{K^*} \sqrt{\lambda_k} \phi_{ik} \gamma_{jk} \right) + e_{ij} \quad i = 1,\ldots,I; \quad j = 1,\ldots,J \qquad \text{(E.6)}$$

Obtenemos la solución del AC minimizando la suma ponderada de los cuadrados de los residuos e_{ij}. La primera parte de la descomposición, $r_i c_j$, es el valor esperado según el modelo de independencia, de manera que la segunda parte explica las desviaciones del modelo de independencia como la suma de K^* términos bilineales (esta parte bilineal tienen una interpretación geométrica en K^* dimensiones, lo que constituye la mayor parte del tema de este libro). Sin embargo, podemos sustituir el modelo de independencia por cualquier otro modelo a elección del usuario. Por ejemplo, en el artículo que mencionamos a continuación, los autores consideran para tablas de contingencia, modelos log-lineales, así que utilizan

el AC como una manera para explorar la esructura de las posibles desviaciones del modelo log-lineal.

- Van der Heijden P.G.M., A. de Falguerolles y J. de Leeuw. J. «A Combined Approach To Contingency Table Analysis and log-Linear Analysis (with Discussion)». *Applied Statistics* 38 (1989): 249-292.

También podemos utilizar esta estrategia en tablas de contingencia de múltiples entradas, utilizando una modelización de las tablas de contingencia que primero tenga en cuenta los efectos principales y determinadas interacciones para, a continuación, calcular los residuos del modelo para analizarlos mediante AC. Sin embargo, dado que los datos ya se han centrado con relación al modelo, no se trata de una aplicación directa del AC. Por tanto, al realizar el AC no debemos llevar a cabo el centrado, y en el ajuste de mínimos cuadrados ponderado debemos utilizar los valores marginales originales de la tabla.

AC y mapas espectrales

El análisis de correspondencias presenta una gran afinidad con los *mapas espectrales*, un método desarrollado originalmente por Paul Lewi en los años setenta y que en el desarrollo de nuevos medicamentos se ha utilizado ampliamente en el análisis biológico de espectros de actividad. Una referencia reciente es:

- Lewi P.J. «Analysis of Contingency Tables». En: B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi y J. Smeyers-Verbeke (eds.). *Handbook of Chemometrics and Qualimetrics*: *Part B*. Amsterdam: Elsevier, 1998: 161-206.

En los mapas espectrales trabajamos con los logaritmos de los valores de la tabla. Sin embargo, llevamos a cabo la ponderación de filas y de columnas como en el AC —utilizamos las masas de filas y de columnas de la tabla original—. Antes de realizar la DVS, llevamos a cabo un centrado con relación a las medias ponderadas de filas y columnas, como en el AC. Si la inercia de los datos es baja, el mapa espectral y el mapa del AC son casi iguales. La diferencia entre los dos métodos es más acusada para inercias mayores. En los mapas espectrales representamos los cocientes de los logaritmos de los datos, lo que hace que este procedimiento tenga propiedades para el diagnóstico del modelo muy interesantes. Además de cumplir el principio de equivalencia distribucional (pág. 60), es *subcomposicionalmente coherente*. Es decir, los cocientes entre valores permanecen constantes aunque se eliminen filas o columnas del análisis. Una propiedad que refuerza este tipo de análisis; pues nos permite analizar con seguridad grupos de filas o de columnas. Por el contrario, en el AC cuando analizamos subgrupos los perfiles y las distancias se ven afectados. Es decir, el AC no es subcomposicionalmente coherente. De ahí la necesidad de desarrollar el AC de subgrupos que vimos en el capítulo 21. Para más detalles y referencias, podemos consultar el documento de trabajo aceptado en el Journal of Classification:

• Greenacre M.J. y P.J. Lewi. «Distributional Equivalence and Subcompositional Coherence in the Analysis of Contingency Tables, Ratio-Scale Measurements and Compositional Data». *Working paper* no. 908, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, 2005. Disponible en Internet: www.econ.upf.edu/en/research/onepaper.php?id=908.

La dimensión de conjuntos de datos categóricos multivariantes

Para finalizar este epílogo, vamos a plantear un problema sin resolver. Sabemos que en AC la dimensión de una tabla $I \times J$, es $(I-1, J-1)$. Para una matriz de Burt $J \times J$ obtenida a partir de Q variables categóricas, el número de dimensiones es $J-Q$. Sin embargo, sabemos que $J-Q$ dimensiones es mucho más de lo que necesitamos para reproducir de forma exacta las tablas que se hallan fuera de la diagonal. Podríamos definir la dimensionalidad de un conjunto de datos con Q variables como el número de dimensiones necesarias para reproducir exactamente la tabla de contingencia $\frac{1}{2}Q(Q-1)$. Es decir, el número de dimensiones necesarias en un AC conjunto para explicar el 100% de la inercia. La pregunta es: ¿podemos determinar las dimensiones de antemano? o, por el contrario, sólo podemos determinarlas empíricamente. Dar respuesta a esta cuestión sería muy útil. Por ejemplo, en el ACM ajustado en el que consideramos sólo las K^* dimensiones para las cuales $\sqrt{\lambda_k} > 1/Q$. En estudios empíricos, la inercia explicada utilizando este número (K^*) de dimensiones se acerca mucho al 100%, aunque no es una prueba suficiente de que la dimensionalidad sea K^*. ¡Quizá con el tiempo se llegue a publicar una tercera edición de este libro, en la que este problema ya esté resuelto!

Índice de imágenes

Índice alfabético

Nota sobre el autor

Michael Greenacre, catedrático de Estadística en la Universidad Pompeu Fabra, en Barcelona, se educó en Sudáfrica, su país de origen, y se doctoró en 1978 en la Universidad de París VI con el profesor Jean-Paul Benzécri, creador del análisis de correspondencias. Se ha especializado en la visualización de conjuntos de datos grandes, especialmente en ciencias sociales y ambientales. Ha realizado años sabáticos en la Rothamsted Experimental Station (Reino Unido), en los Bell Laboratories y en la Universidad de Rochester (Estados Unidos) y en la École des Mines (Francia), así como diversas estancias de investigación en Akvaplan-niva y el Norwegian Polar Institute en Tromsø (Noruega). Desde que se trasladó a París como estudiante en 1973, ha trabajado en análisis de correspondencias durante treinta y cinco años, en los que ha publicado tres libros y ha coeditado con Jörg Blasius otros tres sobre análisis de correspondencias y visualización de datos.