USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Trabajo de Fin de Grado

Asociación entre variables cualitativas; ejemplo práctico de análisis de correspondencias

Jorge Castiñeiras Rella

2018/2019

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo de Fin de Grado

Asociación entre variables cualitativas; ejemplo práctico de análisis de correspondencias

Jorge Castiñeiras Rella

Febrero, 2019

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de conocimiento:
Estadística e Investigación Operativa.
Título:
Asociación entre variables cualitativas; ejemplo práctico de análisis de correspondencias.
Breve descripción del contenido:
Una tabla de contingencia permite resumir información de dos variables cualitativas de modo que las filas representan las categorías de una de las variables y las columnas representan las categorías de la otra variable. El análisis de correspondencias permite representar tablas de contingencia y analizar la similitud entre las categorías de cada una de las variables con respecto a las categorías de la otra. El objetivo de este TFG es que la/el alumna/o haga una revisión de esta metodología de análisis multivariante y aplique los conocimientos adquiridos a un conjunto de datos real.

Índice general

Resumen

El análisis de correspondencias es una técnica de la estadística descriptiva aplicada a tablas de contingencia cuyo objetivo es la visualización de una nube de puntos multidimensional asociada a unas ciertas variables en un espacio de menor dimensión. Se generan mapas que plasman los puntos intentando respetar al máximo sus posiciones en el espacio original perdiendo el mínimo de información, lo que nos permite elaborar un análisis exploratorio de ellos rápido y eficaz además de poder establecer asociaciones entre las variables. Las proyecciones habitualmente son sobre espacios de dimensión 2 por nuestra familiaridad con ellos.

Se tratará el contraste de la homogeneidad de la tabla, se introducirá una medida más adecuada para esta técnica (la distancia ji-cuadrado) y se expresará la varianza de la tabla en función de estos conceptos. Esta nueva medida de la varianza es lo que queremos conservar al realizar la proyección de la tabla. Se introducirán diversos tipos de análisis de correspondencias y técnicas para su implementación, así como realizar inferencia y contrastes sobre los datos.

Todo el trabajo se acompaña de varios ejemplos para facilitar la comprensión de qué se está haciendo. Por último, llevaremos a la práctica todo lo introducido a lo largo del trabajo en un estudio médico real con las pretensiones de resumir los datos, establecer un análisis descriptivo y establecer relaciones entre varias enfermedades comunes y la calidad y forma de vida de 820 pacientes.

Abstract

Correspondence analysis is a technique of descriptive statistics applied to contingency tables whose objective is the visualization of a multidimensional point cloud associated to certain variables in a space of smaller dimension. Maps are generated trying to respect as much as possible their positions in the original space, losing the minimum of information, which allows us to elaborate an exploratory analysis of them quickly and efficiently as well as being able to establish associations between the variables. The projections are usually on spaces of dimension 2 because of our familiarity with them.

The contrast of the homogeneity of the table will be treated, a more appropriate measure will be introduced for this technique (the chi-square distance) and the variance of the table will be expressed in terms of these concepts. This new measure of variance is what we want to keep when projecting the table. Various types of correspondence analysis and techniques for its implementation will be introduced, as well as inference and contrasts on the data.

All the work is accompanied by several examples to facilitate the understanding of what is being done. Finally, we will put into practice everything introduced during the work in a real medical study with the aim of summarizing the data, establishing a descriptive analysis and establishing relationships between several common diseases and the quality and way of life of 820 patients.

Introducción

El análisis de correspondencias (AC) es una técnica estadística aplicada sobre tablas de contingencia, tablas de frecuencias o sobre datos que se puedan expresar en una tabla de estas características. Estas tablas y sus características son detalladas en el capítulo 2. Para poder realizar el AC basta con que los individuos o variables puedan ser agrupados o organizados en categorías. La condición necesaria es que expresemos todas las observaciones en la misma escala, ya sean recuentos o medidas en una determinada unidad. Buscamos realizar una descripción gráfica de un conjunto de variables con varias categorías que sea fácil de interpretar para poder realizar un análisis exploratorio y analítico de estos más sencillo y visual.

El objetivo es encontrar un espacio de baja dimensión (en general de dimensión 2 por ser el tipo de gráfica que nos resulta más familiar y por la fácil interpretación de los datos en dos ejes) en el que representar nuestros datos perdiendo el mínimo de información. Más adelante se explicará como medir la perdida de información al "proyectar" los datos y como buscar la solución óptima mediante varios métodos y algoritmos en el capítulo 3.

Cuando tenemos una tabla medianamente grande, ya resulta muy difícil su interpretación a simple vista. Notemos que si queremos representar los puntos de una tabla de tamaño nxm en todas sus dimensiones necesitaríamos al menos un espacio de dimensión $min\{n-1, m-1\}$. Esto se debe al hecho de que convertiremos nuestra tabla en una tabla de frecuencias en la que la suma de las componentes de cada fila es 1 y por tanto la última variable no aporta información extra en la representación (es redundante). Esta propiedad será tratada en el capítulo 2. En consecuencia para tablas de más de 4 filas y columnas, es decir, en general para datos con más de 4 variables y 4 individuos o grupos, la visualización de éstos gráficamente resulta imposible. No suelen ser interesantes tablas más pequeñas, por lo que el análisis de correspondencias será una herramienta de gran valor y utilidad.

Al trabajar con frecuencias, a la hora de medir distancias entre observaciones no podemos trabajar con la euclidea. Esta distancia da el mismo peso a todas las variables, pero al tener algunas mayor frecuencia que otras, debemos considerar una distancia que dé más peso a los valores más frecuentes. Esta es la motivación para definir la distancia χ^2 y trabajar con ella. En el capítulo 3 introduciremos con detalle esta distancia, así como sus propiedades y aplicaciones en el AC. Ya en el grado de matemáticas, en asignaturas como "Inferencia estadística" o "Probabilidad estadística" se ha introducido la distribución χ^2 como prueba para contrastar la independencia o realizar inferencia sobre la varianza de muestras. En el AC le daremos un enfoque similar, siendo el estadístico ji-cuadrado el que nos permita determinar estadísticamente si existe correlación o no entre filas y columnas y a su vez una medida de la dispersión de la tabla.

Introduciremos también en el capítulo 3 el concepto de inercia, una medida de la varianza total de la tabla. Está definida a partir de la distancia ji-cuadrado y nos da una idea de cómo se sitúan los datos. Será lo que intentemos conservar al realizar las reducciones de dimensión. Siempre se pierde algo de inercia al reducir el espacio, pero veremos formas de minimizar de nuevo esta pérdida y que quede representada lo mejor posible esta medida de la variabilidad de la tabla. Es importante intentar explicar el máximo de inercia pues está intimamente relacionada con la correlación entre variables y es la que nos aporta información para interpretar el AC una vez realizado. Las herramientas y técnicas utilizados para construir estos mapas serán explicados en el capítulo 4, en donde trataremos las representaciones de menor dimensión así como mapas alternativos para el AC.

Una vez realizado el AC y construido el mapa de menor dimensión se podrán establecer relaciones entre unas variables y otras a partir de la posición de las variables en el mapa del AC una vez confeccionado.

Muchas veces nos encontraremos ante tablas de gran dimensión o codificadas en escalas y de forma binaria. Estas tablas tendrán que ser tratadas con cuidado para no llegar a conclusiones erróneas y para poder realizar el AC de forma fiable y útil. Estas alternativas del AC se estudian en el capítulo 5.

También podremos una vez realizado el AC interpretar sus resultados para agrupar variables, lo que nos permite diferenciar entre varias agrupaciones de variables que no se consideraban a priori pero latentes en los datos. Algo parecido a los análisis cluster. En el

capítulo 6 tratamos este tema desde el punto de vista estadístico del contraste de hipótesis y se explicarán técnicas para desarrollar las agrupaciones y comprobar estructuras de los datos. Es una forma de clasificar, diferenciar y relacionar las variables unas con otras, lo cual es en muchos casos de gran interés. En las representaciones que hagamos muchas veces ya se apreciarán estas diferencias y agrupaciones entre unas variables y otras, resultando las nubes de puntos alejadas unas de otras (mostrando diferentes grupos) o muy próximas (mostrando agrupación). Esto también lo comprobaremos estadísticamente, pues al ser representaciones de menor dimensión la información perdida podría enmascarar alguna otra diferencia presente en otra dimensión del espacio no representada.

Está claro que los datos de los que dispongamos para cualquier AC siempre conformarán un subgrupo de una población mayor. Por la imposibilidad de realizar mediciones o estudios sobre toda la población, se consideran las conclusiones establecidas en el AC del subgrupo y se extrapolan a la población total. En el capítulo 6 se tratará la inferencia sobre tablas de contingencia y su desarrollo a partir del AC.

Buscaremos alternativas para el estudio del AC y formas de optimizar los resultados, obteniendo la mayor información y mejor representación posibles. Una vez hecho el AC se podrán exigir ciertas condiciones a los resultados como ejes escalados o una solución en un cierto subespacio. Analizaremos distintos tipos de tablas y veremos varias formas de realizar el AC para cada una en el capítulo 5.

Podemos encontrar muchas herramientas en el programa informático R, con el que realizaremos en este trabajo todos los cálculos y análisis. En este software libre encontramos implementado el AC incluyendo representaciones y muchos de los conceptos que se tratarán a lo largo de este trabajo. Utilizaremos principalmente el paquete **ca** que contiene la mayoría de técnicas aplicadas en el AC.

Durante toda la primera parte del trabajo utilizaremos un conjunto de datos, que introducimos en el capítulo 1. Los datos forman parte de un estudio del grupo AEGIS (A Estrada Inflammation and Glycation Study), cuyo investigador principal es Francisco Gude de la Unidad de Epidemiología Clínica del Hospital Clínico Universitario de Santiago de Compostela. Recurriremos a estos datos como herramienta aclaratoria y para explicar de forma más comprensible todos los conceptos que se exponen a lo largo del trabajo, mostrar los resultados y cómo se implementa el AC en el caso práctico.

En el último capítulo se incluirá un AC mucho más detallado y profundo de los datos proporcionados por este investigador, aplicando todo lo explicado y tratado a lo largo del trabajo. Dispondremos de una base de datos con muchas variables de las utilizadas en los ejemplos del trabajo e intentaremos llegar a conclusiones que resulten de interés medico. Muchos de los resultados que se testan en el anexo no lo han sido nunca y pueden llegar a explicar y probar varios fenómenos médicos sobre los que solo hay hipótesis. La introducción a este estudio y sus datos se realizará en el último capítulo.

La parte teórica del trabajo ha sido desarrollada a partir de la consulta de los libros que se incorporan en la bibliografía, siendo el libro de referencia "*La práctica del análisis de correspondencias*" de Michael Greenacre. No mencionamos este libro a lo largo del trabajo pues en general ha sido el que hemos consultado para todos los capítulos. En algunas secciones citamos ciertos libros por su relevancia en esa parte, pero la mayoría de conceptos han sido explicados tras consultar los libros recogidos en la bibliografía, buscando la notación y explicación más adecuada e intuitiva.

Capítulo 1

Datos para los ejemplos

1.1. Base de datos sobre enfermedades y fármacos comunes

Los datos sobre los que elaboraremos todos los ejemplos y cálculos descritos en la primera parte del trabajo se recogen en este capítulo. El primer conjunto de datos cruza 22 de las enfermedades crónicas más habituales con el consumo de 16 fármacos también comunes (ver Cuadro 1.1).

Esta base de datos es fruto de un estudio en profundidad de la medicación que toman y las enfermedades que padecen 1515 pacientes. Disponemos pues de un gran número de individuos por lo que los resultados obtenidos serán significativos de cara a trabajos y estudios posteriores en el campo médico. El hecho de cruzar el uso de fármacos y enfermedades en este trabajo es una demostración práctica de que efectivamente el AC es una técnica que funciona y que nos puede aportar mucha información por sí sola.

El conjunto de datos interesante tanto en el campo médico como matemático será el que se trate en el último capítulo. Por parte de nuestro investigador y de su equipo se espera obtener información respecto a variables sobre las que casi no existe ningún trabajo, tratándose de hecho este trabajo de un estudio completamente novedoso y que no se había realizado hasta ahora. La innovación sumada a la gran cantidad de datos de la que se ha dispuesto abre la puerta a algún descubrimiento interesante.

A lo largo del trabajo utilizaremos estos datos para establecer ejemplos aclaratorios que reflejen la motivación y desarrollo de lo que hacemos en todo momento. Los datos se recogen en el Cuadro 1.1 y en el Cuadro 1.2 mostramos su codificación.

	Benzo	Antidepre	MetGluc	ado1	Antiinf	Insul2	hipSed	Estatinas	cortic	ACO2	betabloq	diureticos	Glucosamin	fenitoina	antipsic
dm	39	42	471	151	24	32	33	103	2	3	30	59	3	1	4
hta	118	93	689	110	84	23	92	209	9	10	78	161	10	2	3
hlp	125	107	696	107	72	27	91	287	5	10	63	113	9	2	10
ci	20	11	159	19	9	3	15	50	1	2	32	27	1	2	4
ic	5	4	74	8	3	4	2	14	0	0	12	21	0	0	0
ap	7	2	55	10	5	5	6	11	0	1	7	12	1	1	0
ir	14	5	78	9	4	6	4	21	1	0	9	19	0	0	0
hep	19	15	81	15	9	1	12	22	2	2	10	19	0	0	0
asma	20	14	35	1	7	0	17	12	2	7	2	7	0	1	0
epoc	10	6	49	6	5	4	8	17	2	0	4	10	2	1	2
saos	8	5	42	4	2	1	4	12	1	0	4	10	0	0	0
depre	126	170	201	32	44	5	90	64	6	10	14	33	4	0	8
reuma	10	8	46	7	17	0	11	14	7	1	5	7	1	0	0
evc	8	7	63	9	4	4	6	22	1	1	5	10	2	1	2
osteo	30	20	55	5	22	1	16	19	3	2	4	15	4	0	0
cancer	16	13	85	14	9	2	11	26	1	6	8	13	2	0	0
eii	5	3	6	1	3	0	2	1	1	2	0	0	0	0	0
psoriasis	10	7	27	3	4	1	9	8	2	5	3	3	0	0	1
derma	2	2	4	0	2	0	1	1	1	0	0	2	0	0	0
tir	27	30	86	12	22	2	28	36	1	3	12	14	0	0	1
ulcus	5	3	19	2	3	1	4	7	0	0	3	4	0	0	0
migraña	15	13	27	3	19	0	18	7	0	5	6	3	2	0	1

Cuadro 1.1: Tabla de contingencia con los datos sobre enfermedades y medicamentos

ABREVIATURA	ENFERMEDAD
dm	Diabetes media
hta	Hipertensión arterial
hlp	Hiperlipemia
ci	Cardiopatía isquémica
ic	Insuficiencia cardíaca
ap	Arteriopatía periférica
ir	Insuficiencia renal
hep	Hepatopatía
asma	Asma
epoc	Bronquitis crónica
saos	-
depre	Depresión
reuma	Reuma
evc	Enfermedad vascular cerebral
osteo	Artosis
cancer	Cancer
eii	-
psoriasis	Psoriasis
derma	Dermatitis
tir	Tiroidismo
ulcus	Úlcera
migraña	Migrañas

ABREVIATURA	MEDICAMENTO
Benzo	Benzocaína
Antidepre	Antidepresivos
MetGluc	Metabolismo de la glucosa
ado1	Antidiabéticos orales
Antiinf	Antiinflamatorios
Insul2	Isulina
hipSed	Hipnótico-sedantes
Nfarm	Nº Fármacos que toma
Estatinas	Estatinas
cortic	Corticoides
ACO2	Anticonceptivos orales
betabloq	Betabloqueante
diureticos	Diuréticos
Glucosamin	Glucosamina
fenitoina	Fenitonina
antipsic	Antipsicótico

Cuadro 1.2: Tabla con la codificación de los datos sobre enfermedades y fármacos

1.2. Base de datos sobre ansiedad y depresión

El otro conjunto de datos que consideraremos es una tabla que recoge las notas para la ansiedad y depresión de 820 pacientes en una escala del 0 al 9, donde 0 corresponde con no tener la correspondiente enfermedad y 9 tener el mayor grado de esta. Estos datos están relacionados con la tabla anterior y forman parte de un conjunto con muchas más variables de interés del cual ya hemos hablado y que será tratado en el último capítulo. La tabla consta de 820 filas (una para cada paciente) por lo que solo mostramos la tabla para los 5 primeros pacientes (Cuadro 1.3), siendo análoga para el resto de individuos considerados.

ID	Ansiedad	Depresión
8	1	0
15	6	7
22	0	0
23	7	6
24	0	2
⋮	⋮	⋮

Cuadro 1.3: Primeras 5 filas de la tabla con las puntuaciones sobre ansiedad y depresión

Durante toda la parte teórica del trabajo trabajaremos sobre las dos tablas expuestas en este capítulo.

Capítulo 2

Tablas de contingencia y perfiles

2.1. Tablas de contingencia

El comienzo de todo AC es una **tabla de contingencia** en la que se recogen recuentos o frecuencias de algún suceso o categoría cruzados con grupos, categorías, sucesos... En esta tabla mediremos la dispersión de los datos, la correlación entre filas y columnas, qué celdas de la tabla son más relevantes en el AC, qué datos son atípicos y hasta podremos ampliar la tabla o considerar subtablas de esta. A partir de las tablas de contingencia se construyen todas las herramientas y rudimentos del AC, de ahí su importancia y que comencemos con ellas.

En el Cuadro 2.1 mostramos una tabla de contingencia para el caso general con k filas y m columnas, tal como muestra el libro *Análisis de correspondencias simples y múltiples* de Santiago de la Fuente Fernández. Sean X e Y dos variables categóricas, respectivamente, con categorías $x_1, ..., x_k$ e $y_1, ..., y_m$. Una variable categórica se caracteriza por solo poder tomar valores en un cierto conjunto, en donde cada valor se asocia a un grupo o categoría. Lo que representamos en las tablas de contingencia es el recuento para cada categoría dentro de la población sobre la que se ha tomado la muestra. La intersección entre una fila y una columna da lugar a una celda, cuya frecuencia observada es n_{ij}. Los valores marginales representan la suma de todos los valores de esa fila o columna, es decir:

$$n_{i\bullet} = \sum_{j=1}^{m} n_{ij} \tag{2.1}$$

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij} \tag{2.2}$$

$$n_{\bullet\bullet} = \sum_{i} n_{i\bullet} = \sum_{j} n_{\bullet j} \tag{2.3}$$

	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	TOTAL FILA
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2m}	$n_{2\bullet}$
$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{in}	$n_{i\bullet}$
$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{km}	$n_{k\bullet}$
TOTAL COL.	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet m}$	$n_{\bullet\bullet}$

Cuadro 2.1: Tabla de contingencia (caso general)

2.2. Perfiles

Desde este momento y a lo largo de todo el trabajo expresaremos los cálculos y definiciones asociados al AC de las filas. Para las columnas se puede realizar todo lo que hagamos de forma exactamente análoga, por lo que no insistiremos en este.

A partir de la tabla de frecuencias construimos un concepto fundamental en el AC, los **perfiles**. Un perfil es el vector de las frecuencias de una fila divididas por su total. En el Cuadro 2.2 mostramos la tabla de frecuencias asociada a la tabla anterior, donde reflejamos los perfiles fila de los datos. Los perfiles columna se hallarían de forma equivalente dividiendo por los marginales de las columnas.

Así, por ejemplo para la fila i-ésima, su perfil fila viene dado por el vector:

$$(n_{i1}, n_{i2}, ..., n_{im})^T / n_{i\bullet} = (n_{i1}/n_{i\bullet}, n_{i2}/n_{i\bullet}, \cdots, n_{im}/n_{i\bullet})^T \tag{2.4}$$

	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	TOTAL FILA
x_1	$n_{11}/n_{1\bullet}$	$n_{12}/n_{1\bullet}$	$\cdots$	$n_{1j}/n_{1\bullet}$	$\cdots$	$n_{1m}/n_{1\bullet}$	1
x_2	$n_{21}/n_{2\bullet}$	$n_{22}/n_{2\bullet}$	$\cdots$	$n_{2j}/n_{2\bullet}$	$\cdots$	$n_{2m}/n_{2\bullet}$	1
$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	$n_{i1}/n_{i\bullet}$	$n_{i2}/n_{i\bullet}$	$\cdots$	$n_{ij}/n_{i\bullet}$	$\cdots$	$n_{im}/n_{i\bullet}$	1
$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	$n_{k1}/n_{k\bullet}$	$n_{k2}/n_{k\bullet}$	$\cdots$	$n_{kj}/n_{k\bullet}$	$\cdots$	$n_{km}/n_{k\bullet}$	1

Cuadro 2.2: Tabla de perfiles fila (caso general)

Un caso especial de perfil es el que llamaremos **vértice**, es un perfil que concentra toda la frecuencia en una sola componente, es decir, es de la forma $(1, 0, ..., 0)^T$. Será de utilidad a la hora de representar los datos.

Una vez hemos definido los perfiles, observamos que los perfiles fila ocupan un espacio de dimensión m-1, ya que la suma de todos los elementos del perfil es 1 y por tanto una componente resulta redundante. En el caso de perfiles con 3 componentes los podríamos representar en un plano sin perdida de información, pero habitualmente no son de interés tablas tan pequeñas.

Como primer ejemplo vamos a tomar la tabla del Cuadro 2.3.

	Insul2	cortic	betabloq	Total
dm	32	2	30	64
ci	3	1	32	36
ic	4	0	12	16
asma	0	2	2	4
psoriasis	1	2	3	6
Total	40	7	79	126

Cuadro 2.3: Tabla de contingencia sobre enfermedades y fármacos de tamaño 5x3

A partir de la tabla de contingencia de los recuentos, podemos expresar la tabla según las frecuencias de cada fila, es decir los perfiles fila. Solo tenemos que dividir cada valor de cada fila por el marginal de la correspondiente fila.

El Cuadro 2.4 muestra los perfiles fila asociados al Cuadro 2.3.

	Insul2	cortic	betabloq	Total
dm	0.500	0.031	0.469	1
ci	0.083	0.028	0.889	1
ic	0.250	0.000	0.750	1
asma	0.000	0.500	0.500	1
psoriasis	0.167	0.333	0.500	1

Cuadro 2.4: Perfiles fila asociados a la tabla del Cuadro 2.3

Esta tabla puede ser representada en un espacio de 3 dimensiones, en donde la componente en cada dimensión sea la frecuencia con la que los individuos con cierta enfermedad toman ese fármaco. Mostramos la representación de los perfiles fila del Cuadro 2.4 en la Figura 2.1. Así por ejemplo, las coordenadas para "psoriasis" en el espacio de 3 dimensiones generado por las tres columnas serían $(0{,}167, 0{,}333, 0{,}5)^T$. Cada eje corresponde con una columna, en nuestro caso con los medicamentos. El valor máximo de cada componente sería 1 (vértices), por lo que todos los perfiles aparecen en un tetraedro de lado 1. En cada eje, a distancia 1 del origen se encuentra el vértice asociado a ese eje. Representando los perfiles de la tabla anterior en un espacio de 3 dimensiones obtenemos la Figura 2.1.

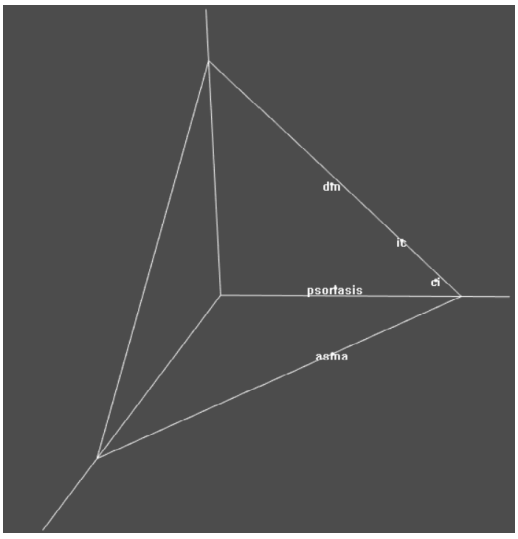

Figura 2.1: Perfiles fila del Cuadro 2.3

Observamos algo que ya se comentó anteriormente, que es que los puntos se sitúan en un espacio de una dimensión menor. En la Figura 2.2 tomamos el plano que contiene a los 5 perfiles fila. Vemos como los puntos se sitúan todos en el mismo plano. Considerando el triángulo en el que se encuentran los puntos, obtenemos:

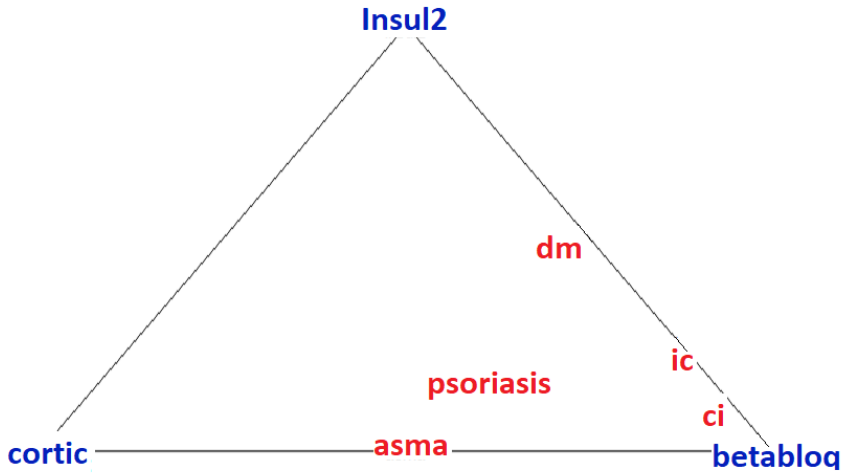

Figura 2.2: Perfiles fila del Cuadro 2.3 en un espacio de dimensión 2

En el espacio de dimensión 2 los perfiles quedan perfectamente representados en un triángulo. Los vértices del triángulo son los valores extremos en los que se concentra toda la frecuencia en un punto, lo que definimos como vértices del AC. Los perfiles no pierden sus posiciones relativas respecto a los vértices ni respecto a ellos mismos y se conservan sus distancias. Esta es una primera aproximación a la utilidad y construcción del AC. Cada perfil se encuentra en las coordenadas definidas por su frecuencia en cada eje. De ahí por ejemplo que asma no tenga componente en la dirección de "Insul2" (su frecuencia es 0), o que "ci" esté tan cerca del vértice asociado a "betabloqueantes" (su frecuencia para este eje es 0,89).

Una vez tenemos la tabla de frecuencias, determinamos el concepto de **masas**. Las masas corresponden a cada una de las filas y se definen como el marginal de la fila dividido por el total de la tabla. Por ejemplo, para la tabla general, la masa asociada al perfil fila i será: $r_i = \frac{n_{i\bullet}}{n_{\bullet\bullet}}$.

Otro perfil que nos será de gran utilidad es el **perfil fila medio**. Lo calcularemos para representar los perfiles y medir el grado de dispersión de los datos. Se define como:

$$c = \left(\frac{\sum_{i=1}^{k} n_{i1}}{n_{\bullet\bullet}}, \frac{\sum_{i=1}^{k} n_{i2}}{n_{\bullet\bullet}}, \cdots, \frac{\sum_{i=1}^{k} n_{im}}{n_{\bullet\bullet}} \right)^T \tag{2.5}$$

Las masas serán muy importantes para establecer los resultados del AC más adelante. En un primer momento, las utilizaremos para calcular el perfil fila medio de los datos. El perfil fila medio es una media de los perfiles fila ponderada por las masas de los perfiles fila. Debe ser ponderada por trabajar con frecuencias, pues no tendría sentido que este perfil medio se hallara a igual distancia de todos los perfiles. Cumplirá entonces que se encuentra más próximo a aquellos perfiles con más peso (masa). Resumiendo, el perfil medio se calcula como:

$$c = \frac{n_{\bullet 1}}{n_{\bullet\bullet}} \cdot y_1 + \frac{n_{\bullet 2}}{n_{\bullet\bullet}} \cdot y_2 + \cdots + \frac{n_{\bullet m}}{n_{\bullet\bullet}} \cdot y_m \tag{2.6}$$

$$r = \frac{n_{1\bullet}}{n_{\bullet\bullet}} \cdot x_1 + \frac{n_{2\bullet}}{n_{\bullet\bullet}} \cdot x_2 + \cdots + \frac{n_{k\bullet}}{n_{\bullet\bullet}} \cdot x_k \tag{2.7}$$

Siendo y_i el vértice asociado a la i-ésima variable. El conjunto de los y_i forman un espacio de dimensión m en donde cada y_i corresponde con el vector unitario que define la i-ésima dimensión.

También podemos interpretar los perfiles como una media ponderada de los vértices, con las masas como las componentes del perfil. La primera interpretación a esto es que los perfiles tenderán a encontrarse más cerca de aquellos vértices para los que su frecuencia es mayor en esta variable.

Para nuestro ejemplo sobre medicamentos y enfermedades, el perfil fila medio es: $(\frac{40}{126}, \frac{7}{126}, \frac{79}{126})^T = (0{,}317, 0{,}056, 0{,}627)^T$, y el perfil columna medio: $(\frac{64}{126}, \frac{36}{126}, \frac{16}{126}, \frac{4}{126}, \frac{6}{126})^T = (0{,}508, 0{,}286, 0{,}127, 0{,}032, 0{,}048)^T$

Los perfiles medios nos resultan muy útiles para comparar el comportamiento de un individuo o grupo respecto a la media global. Así por ejemplo observamos que la frecuencia de pacientes con "ci" (cardeopatía isquémica) que toman betabloqueantes es de 0.889, superior a la media de 0.627, lo que supone una primera aproximación reflejada por los datos de que los pacientes con cardeopatía tienden a tomar betabloqueantes.

Capítulo 3

Distancia ji-cuadrado e inercia

En el capítulo anterior hemos hablado de distancias entre perfiles sin precisar más. Debemos entender que en el AC, como ya explicamos antes, no podemos considerar la distancia euclídea usual para nuestro espacio, pues los perfiles tienen distintas masas y nos llevaría a errores despreciar estas masas. Por ello, en el AC utilizamos la **distancia ji-cuadrado** como herramienta para medir las distancias entre perfiles.

La distancia euclídea usual entre el perfil fila i y el perfil fila j sería:

$$\sqrt{\left(\frac{n_{i1}}{n_{i\bullet}} - \frac{n_{j1}}{n_{j\bullet}}\right)^2 + \left(\frac{n_{i2}}{n_{i\bullet}} - \frac{n_{j2}}{n_{j\bullet}}\right)^2 + \cdots + \left(\frac{n_{im}}{n_{i\bullet}} - \frac{n_{jm}}{n_{j\bullet}}\right)^2}$$

Sin embargo, nosotros utilizaremos una distancia euclídea ponderada, la distancia ji-cuadrado, que repondera las diferencias al cuadrado dividiéndolas por la componente del perfil fila medio correspondiente. Para el caso de la distancia entre el perfil fila i y el j la distancia ji-cuadrado resulta:

$$Distancia\ \chi^2 = \sqrt{\frac{(\frac{n_{i1}}{n_{i\bullet}} - \frac{n_{j1}}{n_{j\bullet}})^2}{c_1} + \frac{(\frac{n_{i2}}{n_{i\bullet}} - \frac{n_{j2}}{n_{j\bullet}})^2}{c_2} + \cdots + \frac{(\frac{n_{im}}{n_{i\bullet}} - \frac{n_{jm}}{n_{j\bullet}})^2}{c_m}} \tag{3.1}$$

La distancia ji-cuadrado cumple el principio de la equivalencia distribucional, que postula que si dos categorías tienen perfiles idénticos pueden ser sustituidas por una sola categoría que sea la suma de sus pesos, sin que con ello se modifique la distancia entre las

filas o columnas. La importancia de esta propiedad reside en que garantiza la estabilidad en los resultados.

Vamos a introducir el concepto de inercia y de estadístico ji-cuadrado de forma intuitiva al principio para comprender el porqué de hacer lo que hacemos y proseguiremos con la descripción formal. Nos interesa comprobar la homogeneidad de la tabla, es decir, si padecer una determinada enfermedad es una característica homogénea según el consumo de ciertos fármacos. En este caso, los perfiles fila asociados a cada enfermedad tenderían a parecerse. Hecha la suposición de homogeneidad, ¿cuales serían los valores esperados de nuestras observaciones? Para la celda (i,j) de la tabla el valor esperado es:

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n_{\bullet\bullet}} \tag{3.2}$$

Continuaremos trabajando sobre el Cuadro 2.3 que cruza 5 enfermedades y 3 tratamientos habituales. Los valores esperados para esta tabla calculados como se ha indicado se recogen en el Cuadro 3.1.

	Insul2	cortic	betabloq
dm	20.32	3.56	40.13
ci	11.43	2.00	22.57
ic	5.08	0.89	10.03
asma	1.27	0.22	2.51
psoriasis	1.90	0.33	3.76

Cuadro 3.1: Valores esperados para el Cuadro 2.3

En este cuadro aparecen los valores esperados para cada celda. Siempre habrá discrepancias entre los valores observados y los esperados, lo que queremos saber es si estas se deben al azar o efectivamente se contradice la suposición de homogeneidad de la tabla. Esta violación de la homogeneidad se traduce en que existe alguna diferencia entre los perfiles de las filas y columnas, es decir, que existe relación entre filas y columnas de algún tipo.

Para saber si estas diferencias son significativamente grandes como para no deberse al azar construimos el estadístico ji-cuadrado definido como:

$$\chi^2 = \sum_{i,j} \frac{(observado_{ij} - esperado_{ij})^2}{esperado_{ij}} \tag{3.3}$$

Para valorar si este estadístico es significativamente grande como para rechazar la hipótesis de homogeneidad recurrimos a la distribución ji-cuadrado con los correspondientes grados de libertad, $(k-1)x(m-1)$. Para la tabla sobre las enfermedades del Cuadro 2.3 los grados de libertad son $(5-1)x(3-1)=8$. El estadístico en nuestro caso será:

$$\chi^2 = \frac{(32-20{,}32)^2}{20{,}32} + \frac{(2-3{,}56)^2}{3{,}56} + \cdots + \frac{(3-3{,}76)^2}{3{,}76} = 46{,}63$$

Notemos que la suma tiene 15 sumandos al estar trabajando con una tabla $5x3$. El p-valor asociado a este valor del $\chi^2 = 46{,}23$ en la distribución ji-cuadrado con 8 grados de libertad es del orden de 0, lo que nos indica que la probabilidad de que las frecuencias observadas provengan de la hipótesis de homogeneidad es muy baja. Por tanto tenemos pruebas significativas de que existen diferencias entre los medicamentos tomados según la enfermedad parecida, es decir, existe asociación entre filas y columnas.

Una forma alternativa de calcular el estadístico sería mediante los perfiles fila observados y esperados. En el libro *Análisis de correspondencias simples y múltiples* de Santiago de la Fuente Fernández aparece expresado de esta forma. En nuestro ejemplo, si dividimos por el cuadrado del total de la fila en los términos correspondientes obtendríamos:

$$\chi^2 = \frac{(\frac{32}{64} - \frac{20{,}32}{64})^2}{\frac{20{,}32}{64^2}} + \frac{(\frac{2}{64} - \frac{3{,}56}{64})^2}{\frac{3{,}56}{64^2}} + \frac{(\frac{30}{64} - \frac{40{,}13}{64})^2}{\frac{40{,}13}{64^2}} + \cdots + \frac{(\frac{1}{6} - \frac{1{,}9}{6})^2}{\frac{1{,}9}{6^2}} + \frac{(\frac{2}{6} - \frac{0{,}33}{6})^2}{\frac{0{,}33}{6^2}} + \frac{(\frac{3}{6} - \frac{3{,}76}{6})^2}{\frac{3{,}76}{6^2}} =$$

$$64 \cdot \frac{(0{,}5-0{,}317)^2}{0{,}317} + 64 \cdot \frac{(0{,}03-0{,}056)^2}{0{,}056} + 64 \cdot \frac{(0{,}47-0{,}63)^2}{0{,}63} + \cdots +$$

$$+ 6 \cdot \frac{(0{,}167-0{,}317)^2}{0{,}317} + 6 \cdot \frac{(0{,}333-0{,}056)^2}{0{,}056} + 6 \cdot \frac{(0{,}5-0{,}627)^2}{0{,}627}$$

De aquí observamos fácilmente que una expresión alternativa para el estadístico ji-cuadrado es:

$$\chi^2 = \sum_{i,j} total\ de\ la\ fila_i \cdot \frac{(perfil\ observados_{ij} - perfil\ esperado_{ij})^2}{perfil\ esperado_{ij}}$$

Como medida de la varianza de la tabla independiente de su tamaño tenemos lo que en AC se llama **inercia** y se expresa como χ^2/n. A la raíz cuadrada de este valor le llamaremos coeficiente phi (ϕ).

Notemos que al dividir el valor esperado para la celda (i,j) entre el marginal de la fila i-ésima obtenemos:

$$\frac{\frac{n_{i\bullet}\cdot n_{\bullet j}}{n_{\bullet\bullet}}}{n_{i\bullet}} = \frac{n_{j\bullet}}{n_{\bullet\bullet}}$$

Lo que coincide con la coordenada j-ésima del perfil fila medio.

A partir de las cuentas realizadas anteriormente podemos declarar la inercia como:

$$Inercia = \sum_i (\text{i-ésima masa}) \cdot (\text{distancia } \chi^2 \text{ del i-ésimo perfil al perfil media})^2 \qquad (3.4)$$

De esta forma la inercia se puede pensar también como una media ponderada de los cuadrados de las distancias χ^2 entre los perfiles fila y su media. Si la inercia es baja los perfiles fila se encuentran muy próximos unos de otros y presentan poca variación. Decimos en ese caso que existe poca correlación o asociación entre las filas y las columnas. Cabe destacar que la inercia siempre será mayor o igual a cero (este sería el caso en que todos los perfiles son iguales) y como máximo igual a la dimensionalidad latente en la tabla. Por ejemplo, para el ejemplo de las enfermedades y fármacos la inercia máxima sería 2.

En el ejemplo que estamos utilizando, la inercia de nuestra tabla es: $46,63/126 = 0,37$.

Podemos definir la inercia formalmente como:

$$\frac{\chi^2}{n} = \sum_i r_i \cdot \|a_i - c\|_c^2 = \sum_i r_i \sum_j \left(\frac{p_{ij}}{r_i} - c_j\right)^2 / c_j \ \text{ (por fila)} \qquad (3.5)$$

$$= \sum_j c_j \cdot \|a_i - r\|_r^2 = \sum_j c_j \sum_i \left(\frac{p_{ij}}{c_j} - r_i\right)^2 / r_i \ \text{ (por columna)} \qquad (3.6)$$

Siendo $r_i = n_{i\bullet}/n$ la masa de la i-ésima fila, $\|a_i - c\|_c = \sqrt{\sum_j (a_{ij} - c_j)^2/c_j}$ la distancia χ^2 entre el i-ésimo perfil fila a_i y el perfil fila medio, $\|a_i - r\|_r = \sqrt{\sum_i (a_{ji} - r_j)^2/r_j}$: distancia χ^2 entre el i-ésimo perfil columna y el perfil columna medio y $p_{ij} = n_{ij}/n$.

Vamos a utilizar toda esta notación introducida para trabajar sobre el cuadro 2.3 de nuevo. Ya calculamos los valores esperados para todas las celdas de la tabla y a partir de ellos el estadístico $\chi^2 = 46{,}63$. Por tanto, la inercia total de los datos es: $\frac{\chi^2}{n} = \frac{46{,}63}{126} = 0{,}37$. Teniendo en cuenta que la inercia máxima para los datos de una tabla de este tamaño sería 2 vemos que es un valor bajo, lo que explica que los datos muestran poca dispersión (algo

que ya observamos en su representación espacial). También comprobamos la hipótesis nula de homogeneidad de la tabla mediante el estadístico χ^2. La distribución ji-cuadrado con 2x4=8 grados de libertad deja un p-valor del orden de 0 para nuestro $\chi^2 = 46{,}63$, por tanto rechazamos esta hipótesis.

Ya representamos los perfiles asociados al ejemplo que cruza enfermedades y fármacos de la tabla 5x3. El problema es que en este espacio las distancias entre puntos corresponden con la distancia euclídea usual, por lo que no nos es de utilidad. Vimos que la χ^2 se puede considerar como una distancia euclídea ponderada, por lo que si escalamos los perfiles de forma adecuada previamente a su representación, las distancias que veremos en el mapa serán las distancias euclídeas entre los perfiles ponderados, es decir, la distancia χ^2. El factor de escala es el que introducimos anteriormente cuando tratamos los paralelismos entre ambas distancias. Bastará dividir cada componente del perfil fila por $\sqrt{c}$, siendo $\mathbf{c}$ el vector de masas de las columnas (perfil medio). Recopilando lo anterior, para el perfil fila a_i, sus coordenadas escaladas en el espacio generado por los vértices son:

$$\left(\frac{a_{i1}}{\sqrt{c_1}}, \frac{a_{i2}}{\sqrt{c_2}}, \cdots, \frac{a_{im}}{\sqrt{c_m}}\right)^T \tag{3.7}$$

Con estos perfiles fila escalados, observemos qué pasaría al calcular la distancia usual entre el perfil fila i y j:

$$Distancia = \sqrt{\left(\frac{a_{i1}}{\sqrt{c_1}} - \frac{a_{j1}}{\sqrt{c_1}}\right)^2 + \left(\frac{a_{i2}}{\sqrt{c_2}} - \frac{a_{j2}}{\sqrt{c_2}}\right)^2 + \cdots + \left(\frac{a_{im}}{\sqrt{c_m}} - \frac{a_{jm}}{\sqrt{c_m}}\right)^2} \tag{3.8}$$

Que coincide con la distancia χ^2 entre los perfiles i y j. De esta forma, las distancias que observaremos en la representación de los perfiles corresponderá con la distancia χ^2 y las podremos interpretar correctamente de forma visual. Al igual que escalamos los perfiles fila, escalaremos los vértices que definen el espacio, pues de lo contrario no tendría sentido la representación. Para el ejemplo sobre las enfermedades que estamos tratando, realizamos este nueva escalado y obtenemos la Figura 3.1.

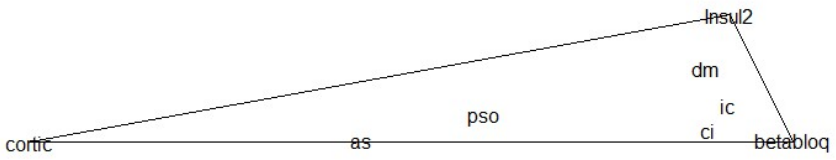

Figura 3.1: Perfiles fila y vértices escalados para los datos del Cuadro 2.3

Se observa, a pesar de que los perfiles ocupan la misma posición unos respecto de otros que las distancias entre ellos se han modificado, dejando patente que efectivamente la anterior representación no era la correcta. Vemos como el vértice asociado a "Corticoides", que era la categoría menos frecuente, es el que más se ha estirado. Esto proviene del hecho de dividir por la raíz cuadrada de la masa para esta categoría y esta ser la que menos masa tiene. De aquí concluimos que las categorías o variables que presenten menos masa serán las que más sufran este efecto de estiramiento.

Las distancias que se observan en este mapa escalado corresponden con la distancia χ^2 entre los perfiles o vértices. Ahora sí que podemos interpretar el mapa y las distancias entre puntos y podemos sacar conclusiones de estas.

Siguen siendo válidos los razonamiento anteriores, donde concluíamos que las enfermedades cardíacas tienden a suponer el consumo de betabloqueantes. No siempre tienen porqué conservarse estas relaciones entre la representación sin escalar y escalada, pero una vez la vemos en el mapa escalado, podemos garantizar esta relación.

Concluimos indicando a partir del mapa de la Figura 3.1 que a parte de lo razonado con las enfermedades cardíacas, las personas que padecen asma tienden más a tomar corticoides y las personas con diabetes a consumir insulina. Es lo que habríamos intuido a priori conforme a lo que sabemos de las dolencias y los medicamentos. La tabla ha confirmado nuestras sospechas iniciales, dejando claro todas estas conclusiones. Este es un ejemplo de para qué sirve y la utilidad práctica del AC.

Capítulo 4

Representación

En el ejemplo anterior ya no resulta evidente visualizar las posiciones de los perfiles al encontrarse en un espacio tridimensional. Cuando trabajemos con tablas mayores (lo cual es habitual) vamos a tener un gran problema para visualizar los datos en un mapa. De ahí que resulte necesario encontrar métodos y alternativas que nos permitan la visualización en espacios de menor dimensión. El AC soluciona esto y nos dará varias herramientas, buscando al mismo tiempo reducir al máximo la pérdida de información. Cuando hablemos de pérdida de información estaremos pensando en la inercia. Esta es la medida de la variación de la tabla y de la relación entre las variables. Nos interesa expresar en nuestro nuevo espacio lo más fielmente posible esta dispersión entre los datos para poder establecer relaciones y diferencias entre grupos y/o categorías. Perder esta información podría ser muy peligroso, conduciéndonos a errores de interpretación y camuflando el verdadero comportamiento de los datos.

Llamemos S a un posible subespacio de nuestro espacio total, preferiblemente de baja dimensionalidad y denotemos por $d_i(S)$ la distancia χ^2 entre el subespacio S y el i-ésimo perfil fila a_i con masa r_i. Consideramos la proximidad de todos los perfiles a S como:

$$\sum_{i=1}^{k} r_i[d_i(S)]^2 \tag{4.1}$$

Multiplicamos cada factor por la masa del perfil fila para que así los perfiles con más masa tengan más peso y forzar a S a pasar cerca de estos perfiles. Claro está que el

subespacio que mejor representa los datos debe pasar "más cerca" de aquellos que más relevancia tienen en términos de masa. El objetivo del AC es encontrar el subespacio que minimice la distancia anterior. Como primera aproximación, se cumple que necesariamente el subespacio S buscado tiene que pasar por el centroide (perfil medio) de los puntos.

4.1. Descomposición en valores singulares

Para implementar la minimización expuesta antes uno de los métodos es recurrir a la **descomposición en valores singulares (DVS)**. Nos apoyaremos en este concepto del álgebra computacional para hallar las coordenadas que tendrían nuestros perfiles en un espacio de menor dimensión. Esta técnica aparece detallada en el libro *Análisis de Correspondencias* de Salvador Figueras, M., de donde hemos extraído los conceptos explicados en este capítulo. Hemos cruzado la información de este libro con el de Luis Joaristi Olariaga y Luis Lizasoain Hernández, *Análisis de correspondencias*, para poder comprender bien esta técnica y poder adoptar la notación y conceptos necesarios para desenvolver el trabajo.

Comenzamos definiendo la matriz C, llamada matriz de residuos estandarizados, del siguiente modo:

$$C = (c_{ij}) \ , \ c_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{4.2}$$

Como trabajamos con tablas de recuentos y frecuencias, los componentes de C serán reales, por tanto la matriz $C^tC \in \mathbb{R}^{mxm}$ será cuadrada, simétrica y semidefinida positiva y por tanto sus autovalores serán todos positivos. Sean $\lambda_1 \geq \lambda_2 \geq \lambda_m$ los autovalores de la matriz C^TC. Entonces los valores $\sigma_i = \sqrt{\lambda_i}$ se denominan los valores singulares de la matriz C.

Sea h el número de dimensiones sobre el que queremos proyectar nuestros datos perdiendo la mínima información (inercia). El AC busca encontrar dos matrices A, B que representen los perfiles fila o columna respectivamente en el espacio de menor dimensión, es decir:

$$A = \begin{pmatrix} a'_1 & \cdots a'_k \end{pmatrix} \quad , a'_i = (a'_{i1}, \cdots, a'_{ih})^T, \text{ con h la dimensión del nuevo espacio.} \tag{4.3}$$

$$B = \begin{pmatrix} b'_1 & \cdots b'_m \end{pmatrix} \quad , b'_i = (b'_{i1}, \cdots, b'_{ih})^T, \text{ con h la dimensión del nuevo espacio.} \tag{4.4}$$

Por lo general h=2.

Para calcular esta A y B el AC descompone la matriz C calculando los valores singulares y las matrices U, D, V tal que:

$$C = UDV^t \tag{4.5}$$

Donde:

$$D = diag(\sigma_1, \cdots, \sigma_H),$$

$$UU^t = V^tV = I,$$

$$U \in \mathbb{R}^{kxH},$$

$$V \in \mathbb{R}^{mxH},$$

$$H = min\{k-1, m-1\}$$

Y una vez calculadas las matrices U, D, V se determinan A y B como:

$$A = D_k^{-1/2}UD \tag{4.6}$$

$$B = D_m^{-1/2}VD \tag{4.7}$$

Siendo:

$$D_k^{-1/2} = diag(n_{1\bullet}, \cdots, n_{m\bullet}),$$

$$D_m^{-1/2} = diag(n_{\bullet 1}, \cdots, n_{\bullet k})$$

4.2. Escalado óptimo

Cuando trabajamos con tablas de contingencias es habitual encontrar variables categóricas a las que es difícil asignar un valor numérico. Por ejemplo, para los datos del cuadro 2.3 sobre el que hemos estado trabajando, podríamos considerar como: 1=Insulina, 2=Corticoides, 3=Betabloqueantes. Podríamos calcular la medicación media como:

$$(1 \cdot 40 + 2 \cdot 7 + 3 \cdot 79)/126 = (0{,}318 \cdot 1) + (0{,}056 \cdot 2) + (0{,}627 \cdot 3) = 2{,}311$$

Donde $(0{,}318, 0{,}056, 0{,}627)^T$ son las componentes del perfil fila medio. El problema de este cálculo, es que los valores 1,2,3 han sido establecidos sin ningún rigor, hemos tomado esta escala como podríamos haber tomado cualquier otra.

Análogamente podemos hallar la media de cada grupo de pacientes con cierta enfermedad usando la escala anterior, obteniendo:

Fila	Media
dm	1.969
ci	2.806
ic	2.500
asma	2.500
psoriasis	2.333

También podemos calcular la varianza de la tabla con esta escala:

$$\frac{64}{126}\cdot(1{,}969-2{,}311)^2+\frac{36}{126}\cdot(2{,}806-2{,}311)^2+\frac{16}{126}\cdot(2{,}5-2{,}311)^2+$$
$$+\frac{4}{126}\cdot(2{,}5-2{,}311)^2+\frac{6}{126}\cdot(2{,}333-2{,}311)^2=0{,}135$$

Los resultados de las medias dependen de la escala entera escogida y puesto que no tenemos ninguna justificación para la elección de estos valores 1,2,3 buscamos una nueva escala más adecuada, que llamaremos **escala óptima**.

En general, llamemos $v_1, v_2, \cdots, v_m$ a la escalada asignada a nuestras variables. La media de medicamento global y para cada fila sería:

$$media\ global = s = \ [(n_{\bullet 1}\cdot v_1)+(n_{\bullet 2}\cdot v_2)+\cdots+(n_{\bullet m}\cdot v_m)]/n$$

$$media\ fila\ i = s_i = \ [(n_{i1}\cdot v_1)+(n_{i2}\cdot v_2)+\cdots+(n_{im}\cdot v_m)]/n$$

Una vez definido esto, establecemos restricciones y condiciones que queremos que cumplan los $v_1, \cdots, v_m$ para despejarlos. Una propiedad deseable sería que las medias de las filas estuvieran lo más espaciadas posibles unas de otras con el fin de distinguir al máximo entre grupos. Igualmente, nos interesa que la varianza sea máxima, para poder tener el máximo de información y poder detectar mejor los agrupamientos o diferencias entre grupos.

La escala óptima vendrá dada por los v_i que maximicen la varianza tal como la indicamos antes. Calculamos estos valores imponiendo las **condiciones de identificación o restricciones**, que son que la media global sea 0, y la varianza 1. Obtenemos entonces los valores óptimos de escala resolviendo el sistema:

$$[(n_{\bullet 1} \cdot v_1) + (n_{\bullet 2} \cdot v_2) + \cdots + (n_{\bullet m} \cdot v_m)]/n = 0 \tag{4.8}$$

$$[(n_{\bullet 1} \cdot v_1)^2 + (n_{\bullet 2} \cdot v_2)^2 + \cdots + (n_{\bullet m} \cdot v_m)^2]/n^2 = 1 \tag{4.9}$$

Tomando la dimensión que mejor ajusta los datos (caso $dim(S) = 1$, donde S es el subespacio considerado al principio de este capitulo), las posiciones de los vértices en esta dimensión resuelven el sistema anterior. La varianza máxima es igual a la inercia en esta dimensión calculada por el AC.

Una vez tenemos los valores de escala óptimos, podemos transformarlos mediante operaciones lineales a una escala más conveniente. En general nos interesa fijar el valor mínimo y máximo de la escala. Para hacer esto, basta con despejar los nuevos valores de escala v'_i tal que:

- $v_{min} = min\{v_i,\ i = 1, \cdots m\}$,
- $rango = Max\{vi,\ i = 1, \cdots m\} - v_{min}$,
- v'_{min} = El nuevo límite inferior, por ej. 0
- $rango'$ = rango que nos interesa, por ej. 0-100,

Entonces, para obtener los valores en la nueva escala con el máximo y mínimo donde nos interesa transformamos los valores:

$$v'_i = \left[(v_i - v_{min}) \cdot \frac{rango'}{rango}\right] + v'_{min} \tag{4.10}$$

Otra alternativa para hallar el escalado óptimo es abordar esta búsqueda desde otra perspectiva. Definidos los vértices, el objetivo será encontrar una escala tal que los perfiles con mayor frecuencia en un vértice estén más cerca de este. Si $h_1, \cdots, h_m$ son los valores de la escala para las columnas y $a_1, \cdots, a_k$ los de las filas, se intenta minimizar la función:

$$\sum_i \sum_j p_{ij} d_{ij}^2 = \sum_{i,j} p_{ij} \|a_i - h_j\|^2 \tag{4.11}$$

Donde $d_{ij} = \|a_i - h_j\|$ es la distancia entre un grupo (en nuestro ejemplo enfermedades) y una categoría (medicamentos). Por conveniencia tomamos el cuadrado de las diferencias en vez de el valor absoluto.

Si consideramos de nuevo las condiciones de identificación, obtenemos de nuevo la solución con el espacio óptimo del AC de dimensión 1.

4.3. Ejes y coordenadas principales

En la anterior sección nos hemos referido a la dimensión que mejor ajusta los datos pero no hemos profundizado en ella. Esta dimensión se suele denotar como **primer eje principal**. Una vez hallado este eje, podemos calcular la inercia de los datos una vez proyectados sobre el. A la inercia explicada por este eje se le llama **primera inercia principal**. Análogamente se definen el segundo eje principal y la segunda inercia principal y así sucesivamente. El segundo eje principal es el que intenta explicar la inercia que no ha sido explicada por el primer eje principal y así con todos los ejes, hasta el punto en que si tomamos todas las dimensiones de la tabla representemos el 100 % de la inercia al estar representando los perfiles en el espacio en que se encuentran.

Dividiendo la inercia principal de un eje entre la inercia original, obtenemos una medida de la inercia explicada en el eje en forma de porcentaje, lo que nos da una idea de lo bien o mal representados que quedan los datos en este eje. Cuanto mayor sea este porcentaje, mejor representados estarán los perfiles en el eje. La inercia que queda sin representar es la que intentaríamos representar con el segundo eje principal. Entre los dos ejes explican la suma de sus inercias principales. El objetivo del AC es encontrar el menor número de ejes que expliquen la mayor parte de la inercia.

A las coordenadas en estos ejes nos referiremos como **coordenadas principales**.

Los conceptos de ejes y coordenadas principales, representación e inercia han sido reforzados en este trabajo con el libro *Análisis de correspondencias* de Luis Joaristi Olariaga y Luis Lizasoain Hernández. Este libro incluye además en el segundo capítulo un detallado ejemplo para mostrar todos estos conceptos, lo que nos ha facilitado su comprensión.

4.4. Representaciones en dos dimensiones

Vamos a trabajar con otros datos extraídos de los introducidos en el primer capítulo. Ahora tenemos una tabla que cruza 6 enfermedades con 4 medicamentos. Presentamos este segundo cruce en el Cuadro 4.1.

	Insul2	Antidepre	Estatinas	ado1
dm	32	42	103	151
ci	3	11	50	19
ic	4	4	14	8
depre	5	170	64	32
hta	23	93	209	110
ap	5	2	11	10

Cuadro 4.1: Tabla de recuentos sobre enfermedades y fármacos

En este ejemplo, el espacio de perfiles queda representado en un espacio de 3 dimensiones. Vamos a proyectar los vértices y los perfiles sobre el plano que mejor se ajuste, es decir, el que más inercia explique. Con los procedimientos expuestos anteriormente y representando el primer eje principal horizontalmente (lo habitual en AC), obtenemos el mapa de la Figura 4.1.

Entre paréntesis al lado de la etiqueta de cada dimensión nos aparecen el porcentaje de inercia explicada en cada eje, por lo que la inercia explicada por el plano es 80,4 % + 17,6 % = 98 % de la inercia total de la tabla. El 2 % que no aparece representado en la tabla corresponde a la tercera dimensión. El porcentaje de inercia explicado es muy elevado, por lo que tenemos una muy buena representación de los datos renunciando a una dimensión. Esto nos indica que los datos pasan todos muy cerca del plano, por ello al explorar las posiciones relativas de los perfiles en el mapa no habrá ningún problema obviando la dimensión no representada. El origen representa el perfil medio.

Se representan los vértices para tener una referencia y establecer comparaciones, lo mismo para el perfil medio. En la tabla vemos una estructura muy clara de los vértices en el primer eje principal. A la derecha se sitúa el perfil y el vértice asociado a enfermedades psíquicas y a la derecha los perfiles y vértices asociados a enfermedades físicas.

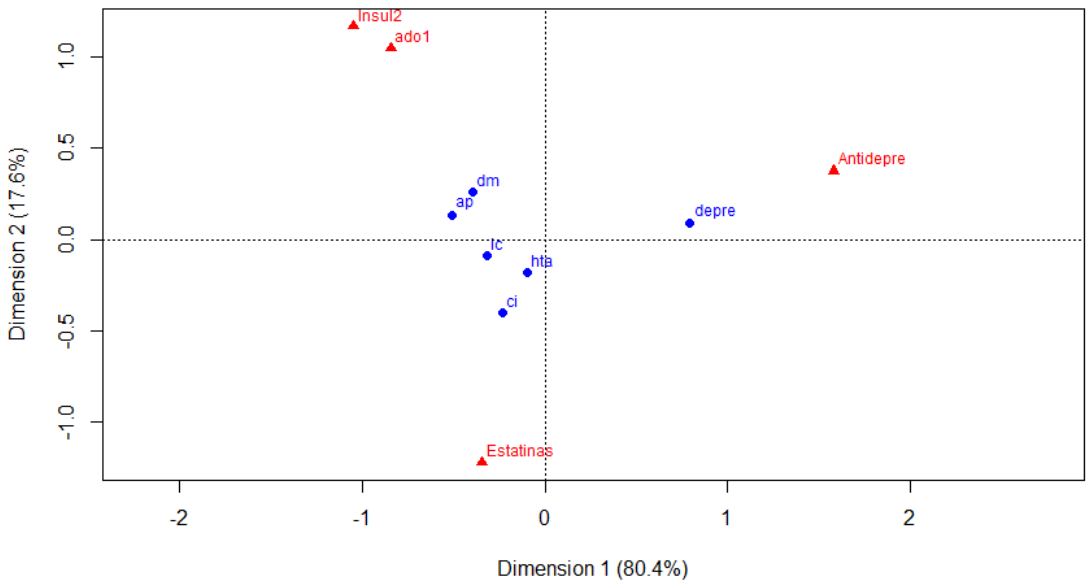

Figura 4.1: Perfiles representados en los dos primeros ejes principales del Cuadro 4.1

Por tanto, un perfil situado a la derecha del origen (perfil medio) se asociará con una enfermedad psíquica. En el segundo eje, también se aprecia una clara tendencia. Los dos vértices superiores a la izquierda corresponden con fármacos para tratar la diabetes y el inferior izquierdo con un medicamento para bajar el colesterol. Por tanto, si un perfil de una enfermedad se encuentra a la izquierda del origen, se asociará con fármacos para la diabetes si está por encima del origen y con tratamientos para la dolencia cardíaca si se encuentra por debajo. También podemos indicar cómo dentro de las enfermedades cardíacas, el caso del "ci" (cardiopatía isquémica) es el que más tiende a ser medicado con estatinas. Al haber una inercia explicada muy alta, las interpretaciones son correctas y vemos como se relacionan las enfermedades y los fármacos fielmente.

En general, si la representación es de calidad (hay un alto porcentaje de inercia explicado en el mapa) podemos establecer conclusiones fiables de esta. Un perfil que se encuentre cerca de un vértice, indica que este perfil se identifica mucho con ese vértice. Podemos sacar conclusiones comparando a que vértice está más próximo cada perfil, entre dos perfiles cual se dispone más cerca de un cierto vértice, si existe alguna ordenación de los perfiles o vértices, etc...

En el AC no existe una norma universal para su interpretación, dentro de cada mapa estudiamos las posiciones de vértices y perfiles y sacamos conclusiones de estas según su proximidad u ordenación.

Decimos que los ejes están **anidados**, ya que de realizar la proyección de los perfiles del plano sobre el primer eje principal, obtendríamos las mismas coordenadas que si hiciésemos el AC para hallar solo el primer eje principal.

Llamamos **coordenadas principales** a las coordenadas de los perfiles en un eje principal y **coordenadas estándares** a las coordenadas en un eje principal de los vértices. En la representación de la Figura 4.1 hemos representado las filas en coordenadas principales y las columnas en coordenadas estándares. A una representación de este tipo se le denomina **mapa asimétrico**. El mismo nombre recibiría un mapa con las filas en coordenadas estándares y las columnas en coordenadas principales. El nombre del mapa viene de que las coordenadas de filas y columnas no están expresadas de la misma forma. La Figura 4.1 también recibe el nombre de **mapa en filas principales**.

Esto motiva la introducción de un nuevo tipo de mapa llamado **mapa simétrico**. Consiste en una representación de filas y columnas en coordenadas principales. No existe un consenso respecto a qué mapa es mejor. Los mapas simétricos suelen ser más útiles cuando la inercia de la tabla es baja, pues al representar filas y columnas en coordenadas principales se aprecia mejor la dispersión de los datos. A cambio, en los mapas simétricos solo se pueden establecer conclusiones a partir de la posición relativa de los perfiles, basándonos en la proximidad entre perfiles. El hecho de no representar los vértices hace que perdamos este punto de referencia. Lo habitual es decantarse entre un mapa u otro dependiendo de los datos y tipo de estudio que nos interese realizar.

4.5. Simetría entre el AC de filas y columnas

Hemos comentado que todos los cálculos y definiciones introducidos para las filas tienen su análogo para columnas. En este capítulo profundizaremos más en la relación que existe entre el AC por filas y columnas. De hecho, como introducción, podríamos pensar el análisis por columnas como un AC de las filas de la tabla traspuesta.

Existe un factor de escala para pasar de las coordenadas de un perfil en un eje principal a las coordenadas del perfil columna:

$$\text{Coordenada del perfil columna} = \text{Coordenada del vértice} \cdot \sqrt{\textit{Inercia principal}} \quad (4.12)$$

Siendo la inercia principal la inercia explicada por el eje principal correspondiente. Por tanto, cuanto mayor sea esta inercia principal, mayor será la dispersión de los perfiles respecto a los vértices en el eje principal. Están relacionados los perfiles y los vértices en el sentido de que en un eje principal, las posiciones de los perfiles tienen las mismas posiciones relativas que los vértices en el análisis análogo, con la salvedad de que se contraen los valores (debido a que el factor de escala es $\sqrt{\textit{inercia principal}}$, que es siempre menor o igual a 1).

Haciendo un AC, estamos realizando también el análogo por filas o columnas. Por tanto, podemos ver el AC como un análisis simultaneo de filas y columnas.

4.6. Contribuciones a la inercia

Ya hemos introducido el concepto de inercia como una medida de dispersión de la tabla y la inercia principal como el porcentaje de esta medida explicada por un eje principal. En esta sección veremos cuánto contribuye cada fila a la configuración de la inercia total, la cual definimos como una media ponderada de las distancias χ^2 entre los perfiles y el perfil medio, i.e.:

$$\textit{Inercia} = \frac{\chi^2}{n} = \sum_i r_i \cdot \|a_i - c\|_c^2 = \sum_i r_i \sum_j \left(\frac{p_{ij}}{r_i} - c_j\right)^2 / c_j$$

La **inercia de filas** es la contribución de cada fila a la inercia. Esta será igual al término del sumatorio asociado a esta fila. Así, la contribución a la inercia de la i-ésima fila es $r_i \cdot \|a_i - c\|_c^2$. Comparamos estos valores de inercia de filas con la total y los expresamos en tantos por mil. Para considerar una contribución como alta, ponemos como límite la media de las inercias. Es decir, en un caso con k filas consideramos como contribución elevada a toda fila con inercia mayor a $\frac{\chi^2}{n}/k = \frac{\textit{Inercia}}{k}$.

Podemos reescribir la inercia como $\sum_i r_i d_i^2$ y la contribución de un perfil i a la inercia principal de un eje k como $r_i f_{ik}^2/\lambda_k$ en tanto por mil. Siendo d_i la distancia del perfil a_i al

perfil fila medio c, f_{ik} la coordenada principal de a_i en el eje principal k y $\lambda_k = \sum_i r_i f_{ik}^2$ la inercia principal k-ésima.

También podemos considerar la contribución de cada celda de la tabla. A estas las llamaremos **contibuciones ji-cuadrado** y corresponden a los valores que aparecen en los términos para calcular el estadístico χ^2. Para la celda (i,j) su contribución a la inercia es:

$$\frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{4.13}$$

La Figura 4.2 (extraída del libro *La práctica del análisis de correspondencias* de Michael Greenacre) muestra la proyección de un perfil sobre un eje principal k. En cada eje principal ocurre lo mismo que en este. Esto es lo que ocurre cuando realizamos el AC y proyectamos un perfil en el nuevo espacio de menor dimensión.

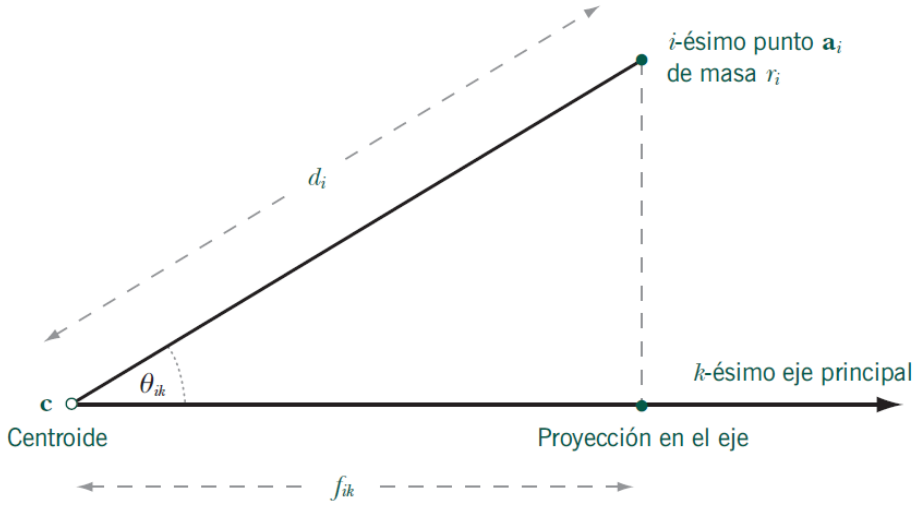

Figura 4.2: Proyección del perfil i-ésimo sobre el eje principal k

Usando el Teorema de Pitágoras deducimos que: $d_i^2 = \sum_k f_{ik}^2$, por lo que la contribución a la inercia de la fila i es:

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

De igual forma, la contribución del eje k a la inercia del i-ésimo perfil es $\frac{r_i f_{ik}^2}{r_i d_i^2} = \left(\frac{f_{ik}}{d_i}\right)^2$.

Podemos hacer también una interpretación geométrica de todo esto. Si nos fijamos, los valores $\left(\frac{f_{ik}}{d_i}\right)^2$ coinciden con $cos(\theta_{ik})^2$, siendo θ_{ik} el ángulo indicado en la Figura 4.2. La contribución del eje k sobre la fila i expresada en tanto por mil siempre está entre 0 y 1. Vemos entonces que cuanto mayor sea la contribución del eje a la inercia de la fila, menor será este ángulo y por tanto el perfil estará mejor representado en la dimensión de este eje. Si la contribución fuese baja, el valor del coseno también y el ángulo se aproxima más a los 90 grados (valor máximo que podría tomar el ángulo en el caso que la contribución del eje a la inercia de la fila sea cero). En este caso, el perfil no se halla cerca de este eje y se encontrará en otras dimensiones del espacio. De esta forma podemos considerar este ángulo como un coeficiente de correlación.

Recapitulando, podemos expresar la contribución de un eje k a la inercia de las filas de dos formas:

- Con relación a la inercia principal del eje. Lo que nos permite diagnosticar qué filas (columnas) han tenido más importancia a la hora de configurar los ejes principales.
- Con relación a la inercia de la fila. Nos permite saber si los perfiles están bien representados. Si están bien representados podemos interpretar el mapa con seguridad de que las conclusiones que saquemos del mapa estarán bien. Si no lo están, debemos interpretar sus posiciones con cuidado pues podemos estar obviando la representación del perfil en otra dimensión que no hayamos considerado y que explique de forma más adecuada el perfil.

4.7. Puntos adicionales

Ya vimos que cada punto influye de distinta forma en la configuración del mapa, dependiendo su influencia de su masa y su posición respecto al centroide de los datos. Los perfiles con más masa ya explicamos que tienen un poder de atracción mayor sobre los ejes del mapa, ya que lo que intentamos minimizar es la distancia entre los ejes y los perfiles ponderada con sus masas. La posición de los perfiles influirá en cuanto los perfiles más alejados del centroide afectan más a la orientación del mapa.

Es conveniente tener recursos para tratar con datos o perfiles que afecten demasiado a nuestro mapa, más aún si son datos atípicos. Estudiaremos tres situaciones importantes:

Punto inherentemente distinto a los demás. Esto ocurre cuando tenemos un dato que por construcción o definición es diferente al resto. Por ejemplo si en el ejemplo que hemos tratado hasta ahora añadiéramos una fila indicando el nivel de estudios estaríamos ante una variable que no tiene la misma medida ni es el mismo tipo de variable que el resto. Esta fila sería inherentemente distinta al resto y además probablemente tuviera unas frecuencias atípicas respecto a las demás filas. El considerar esta fila a la hora de hacer el AC podría cambiar notablemente los resultados, por lo que no se puede considerar por su naturaleza intrínsicamente diferente. Una vez hecho el AC sobre los datos sin el nuevo perfil, podríamos representarlo como un punto sin masa, simplemente calculando sus coordenadas principales y situándolo en el mapa. Esto si que no afectaría al mapa ni al AC en general al ser una representación y consideración a posteriori.

Punto atípico y de poca masa. Como ya comentamos, los puntos atípicos pueden influir sobre el AC y sus resultados. La mera posición de un punto aunque tenga poca masa puede tener el efecto de rotar los ejes principales. Este efecto también se hace latente debido a que la inercia del mapa crece mucho en comparación con la inercia sin considerar este perfil. Estos problemas suelen provenir de perfiles muy alejados del resto de perfiles. Si tenemos un punto de poca masa y que afecta a la inercia significativamente optamos por no considerarlo en el AC. Como en el caso anterior podríamos representar a posteriori el perfil sobre el mapa, o de tener sentido podríamos considerar combinar este perfil con otro.

Agrupaciones. En los casos en que tenga sentido y nos interese, podemos considerar agrupar una o más filas (o columnas) para su estudio conjunto. El problema es que esta agrupación podría influir en la configuración del mapa. Por ello realizamos el AC sobre los datos originales y después representaríamos los perfiles de los nuevos perfiles agrupados. Esto nos da una idea de como se sitúa el nuevo perfil agrupado respecto al resto y si la agrupación es correcta o no.

Las tres interpretaciones y alternativas anteriores las podemos hacer situando los perfiles con relación a los vértices de un mapa asimétrico o considerando el mapa simétrico y también estudiar la calidad de la representación.

4.8. Biplots

Los biplots son un tipo de representación del AC para cada perfil fila y columna conjuntamente a partir del producto escalar de dos vectores convenientemente seleccionados. La idea es poder volver a obtener los datos de la tabla original a partir de productos escalares de vectores de menor dimensión, o al menos, conseguir una aproximación tan exacta como nos sea posible de los datos originales. Así, intentaremos aproximarnos lo máximo posible a cada valor n_{ij} a partir del producto escalar $x_i^T \cdot y_j$, siendo $x_i,\ y_j$ los vectores de menor dimensión comentados antes. Como aclaración, el prefijo "bi" en biplot no significa que sean representaciones en dos dimensiones (aunque por lo general así sea) si no que indica que estamos representando conjuntamente filas y columnas.

Esta sección se ha desarrollado con nuestro libro de referencia y con el trabajo *The biplot graphic display of matrices with application to principal component analisis* de K.R. Grabiel, en donde se detallan las técnicas y rudimentos de los biplots con mucha más profundidad. En este trabajo solo recogemos la idea general, su construcción y su relación con el AC.

El biplot representa los datos originales como:

$$n_{ij} = x_i^T \cdot y_j + \epsilon_{ij}$$

Siendo $x_i,\ i \in \{1, \cdots, m\}$; $y_j,\ j \in \{1, \cdots, k\}$ vectores y ϵ_{ij} el error cometido al aproximar el elemento de la tabla (i,j) por el producto escalar. El objetivo del biplot es minimizar estos errores.

El AC está estrechamente relacionado con el biplot. Introducimos la que llamaremos **fórmula de reconstitución**:

$$p_{ij} = r_i c_j \left(1 + \sum_{k=1}^{K} \sqrt{\lambda_k} \phi_{ik} \gamma_{jk}\right) \tag{4.14}$$

Donde $p_{ij} = \frac{n_{ij}}{n}$, $r_i\ y\ c_j$ son las masas de las filas y columnas respectivamente, λ_k la k-ésima inercia principal, $\phi_{ik},\ \gamma_{jk}$ las coordenadas estándares de las filas y columnas respectivamente y K es la dimensión de la matriz de datos, es decir, el menor número entre el numero de filas menos uno y el de columnas menos uno.

Nos interesa expresar la fórmula de reconstrucción como un producto escalar para poder extrapolarlo al biplot. Reacomodando la expresión (4.14) tenemos:

$$\frac{p_{ij}}{r_i c_j} - 1 = \sum_{k=1}^{K} f_{ik}\gamma_{jk} + e_{ij} \tag{4.15}$$

Siendo $f_{ik} = \sqrt{\lambda_k}\phi_{ik}$, la coordenada principal de la i-ésima fila en el k-ésimo eje.

Aquí se demuestra que el mapa asimétrico de filas es un biplot aproximado de los valores $\frac{p_{ij}}{r_i c_j} - 1$. Multiplicando por $\sqrt{c_j}$ los puntos se aceracarán al origen, acercándose más aquellos puntos con menos masa, lo que nos interesa para mejorar la legibilidad del biplot. A esta variante la llamaremos **biplot estándar**. En este tipo de mapas no podemos interpretar las distancias como en los mapas anteriores, en los biplots estándar interpretamos las proyecciones de los puntos sobre los ejes del biplot (cada eje está determinado por cada categoría). Estas proyecciones estiman los valores estandarizados en el lado izquierdo de la ecuación (4.15).

Vamos a representar en la Figura 4.3, para que quede patente como son, el biplot estándar de la segunda tabla que cruzaba las enfermedades y medicamentos.

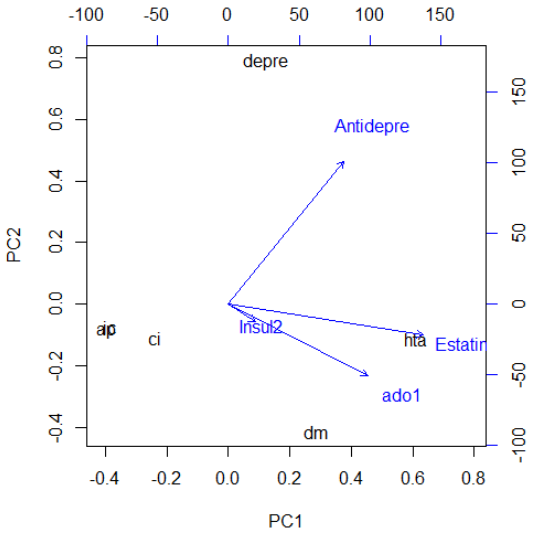

Figura 4.3: Biplot estándar del Cuadro 4.1

Como dijimos, podemos interpretar las proyecciones de cada punto sobre el eje generado por cada columna. De esta forma, vemos como el eje "Antidepre" aparece casi en perpendicular a los otros dos ejes, y como la enfermedad "depre" claramente es la que tiene un valor mayor en este eje. También vemos una situación similar con la diabetes y los dos medicamentos considerados para tratarla y que las enfermedades cardíacas se ordenan de la misma forma respecto al eje de "Estatinas". Podemos obtener los mismos razonamientos que hicimos con el mapa del AC de estos datos pero en este caso con el biplot estándar.

Este ejemplo muestra como efectivamente existe gran relación entre el AC y los biplots y como interpretar la información que nos dan estos.

Los biplots funcionan igual de bien para tablas con inercias bajas o altas, por lo que serán especialmente útiles cuando nos encontremos con tablas de datos de baja inercia como la que teníamos. De hecho en el biplot vemos una mayor diferenciación entre los perfiles asociados a la diabetes y su tratamiento y el resto.

4.9. Transformación de tablas

Pueden interesarnos casos en los que intervengan más de dos variables en la determinación de los individuos o de las categorías por la naturaleza inherente de los datos, por una estructura o diferenciación que nos interese comprobar o simplemente para refinar lo calculado en un AC anterior. Otra posibilidad sería la de pretender realizar el AC a una tabla con muchas variables en filas, columnas o ambas. En este caso consideraremos el análisis de tablas concatenadas, en el que solaparemos varias tablas con distintas variables pero relacionadas y en la misma escala para que el solapamiento tenga sentido. Así, al realizar el AC podemos estudiar el comportamiento global de la tabla solapada y el de cada subtabla.

En el ejemplo del Cuadro 4.1 sobre las enfermedades podríamos introducir una tercera variable como el sexo. De esta forma, obtendríamos 2x5 variables en las filas en lugar de las 5 que teníamos anteriormente. Una vez hecho esto, nos interesará saber si el efecto del género se mantiene a lo largo de todos los grupos, es decir, que la relación hombres-mujeres se mantenga uniformemente dentro de los 5 grupos. De haber algún o algunos grupos en los que no se mantuviese esa homogeneidad, en estos se presentaría un efecto de interacción.

Para contrastar esto, construiríamos una nueva tabla con 10 filas representando (2 géneros)x(5 enfermedades), realizaríamos el AC y veríamos en el mapa obtenido en qué grupos existe el fenómeno de interacción y en cuales no. Si no se produjese interacción en ningún grupo veríamos una distribución en el mapa similar en los puntos de los hombres y mujeres, encontrándose cerca enfermedades iguales en hombres y mujeres. De esta forma intuiríamos la no interacción entre género y enfermedades.

En general, si tenemos dos variables categóricas con I, J categorías respectivamente, podemos codificar ambas en una tabla de múltiples entradas codificando una nueva variable con IxJ categorías. Cruzamos esta nueva variable con otra que nos interese, realizamos el AC e interpretamos este de la manera habitual, con la posibilidad además de estudiar el efecto de interacción entre las dos variables categóricas que codificamos como una y la variable cruzada.

En el párrafo anterior hablamos del caso de dos variables categóricas con I, J categorías respectivamente, lo que conduce a una tabla con una variable con IxJ categorías. Pero en el caso de tener más variables categóricas la tabla crecería de tamaño hasta un punto en que no podamos interpretar los puntos del mapa. Para el ejemplo de las enfermedades cruzadas con los fármacos, si añadiésemos las categorías sexo, casado/soltero, nacionalidad, ... llegaríamos a una tabla enorme y consecuencia de esto obtendríamos una elevada cantidad de puntos en el mapa que no nos permitirían diferenciar unos de otros y dificultaría la interpretación.

En estos casos consideramos codificar los datos en forma de tablas concatenadas. En estas tablas cruzamos las categorías con la variable indicada en la tabla pero considerando las subtablas de cada categoría. Primero realizamos este análisis y después consideramos todas las tablas concatenadas. Este método no nos permite identificar interacciones, pero si lo podemos considerar como un AC medio de las subtablas que consideremos. Recalcamos que con estas interpretaciones no podemos establecer relaciones entre las categorías de las subtablas, solo entre las categorías dentro de cada subtabla con las categorías con las que se cruzan. Para realizar estas comparaciones, tendríamos que cruzarlas y realizar el AC a la tabla grande.

Hemos expresado el ejemplo para el caso en que aumentamos el número de filas. Pero podría razonarse análogamente para las columnas e incluso para el caso en que tengamos más categorías para filas y columnas. Pongamos que nuestra tabla concatenada tenga

ahora S y Q categorías para filas y columnas respectivamente. Dentro de cada categoría encontramos subcategorías que conforman la subtabla. Para el ejemplo de las enfermedades, al añadir la categoría sexo encontraríamos una tabla con 2x5 filas, donde S=5 y cada tabla tendría dos subcategorías, hombre-mujer. Otro razonamiento sería el de una tabla con S=2 y 5 subcategorías referentes a cada enfermedad. Esta elección suele venir motivada por lo que nos interese contrastar. La tabla concatenada es esencialmente la misma, solo estamos intercambiando filas, por lo que los resultados del AC no varían.

Como aclaración, la tabla concatenada resultante tendría la forma de la tabla de la Figura 4.4.

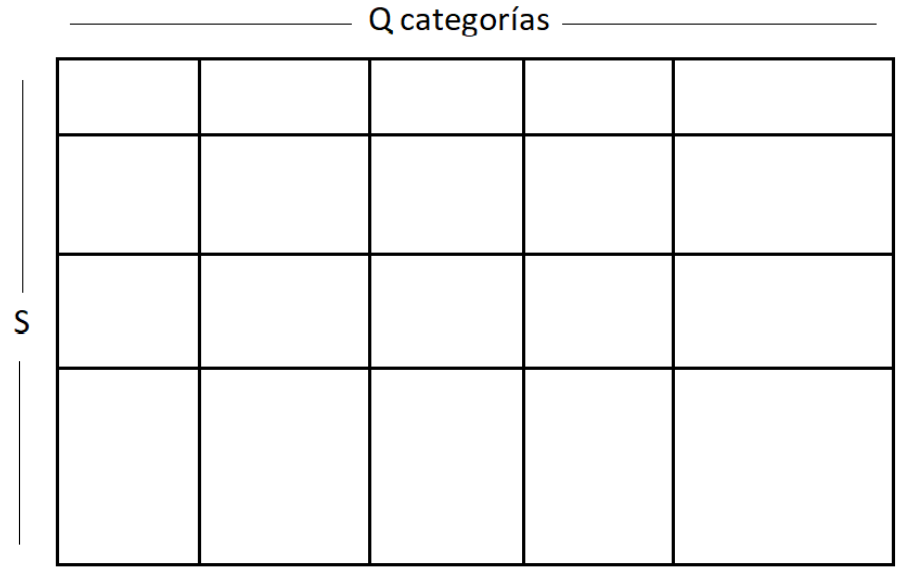

Figura 4.4: Diagrama tabla concatenada con S y Q categorías

La inercia total de la tabla concatenada es la media de las inercias de cada subtabla. Si tenemos S y Q variables categóricas respectivamente que componen la tabla concatenada, la inercia de esta es:

$$inercia(N) = \frac{1}{QS} \sum_{q=1}^{Q} \sum_{s=1}^{S} inercia(N_{qs}) + \epsilon \tag{4.16}$$

Donde N_{qs} representa la subtabla de contingencia (q,s) y ϵ es un error que se comete al realizar esta aproximación. El error proviene de que no todas las subtablas tienen porque

tener las mismas frecuencias marginales debido a valores perdidos, no respondidos... por lo que la media no concuerda exactamente con la inercia global. Si todas tuvieran los mismos valores marginales, $\epsilon = 0$.

Si trabajamos con variables expresadas en una escala de grados (por ejemplo una encuesta de satisfacción donde las respuestas estén entre 1-5) precisamos realizar una transformación de esta escala que aporte información al AC. Esta transformación se conoce como **doblado de los datos**. Consiste en expresar los datos en relación a su diferencia con el valor más bajo y más alto de la escala. Lo preferible es situar en cero el valor más bajo de la escala. Para el ejemplo de la encuesta de satisfacción procederíamos restando uno a todas las variables y trabajando con la escala 0-4. Después expresaríamos los valores doblados de cada respuesta como:

$$1 \rightarrow 0 \quad 4$$

$$2 \rightarrow 1 \quad 3$$

$$3 \rightarrow 2 \quad 2$$

$$4 \rightarrow 3 \quad 1$$

$$5 \rightarrow 4 \quad 0$$

Por ejemplo el valor 2 tiene 1 (1) valor por debajo y 3 (3,4,5) por encima. De ahí su codificación. De esta forma construimos una nueva tabla que dobla el tamaño de las columnas, apareciendo para cada fila su valor doblado en cada respuesta para cada variable. Esta tabla mide la asociación de cada encuestado con los extremos de la escala.

Aplicando el AC a esta tabla doblada, obtenemos información de los datos y podemos interpretar los mapas obtenidos. Los valores doblados en el mapa se hallarán en puntos opuestos respecto al origen. Podemos trazar una linea entre cada perfil doblado e interpretar los cosenos entre cada linea como una correlación como hacíamos antes. De esta forma, los perfiles cuyas rectas muestren poco ángulo entre ellas se asociarán con variables con mucha correlación, y los que muestren ángulos más próximos a 90^o se asociarán con variables incorreladas.

Podemos realizar todo lo anterior ante cualquier tipo de datos que se pueda codificar doblando los datos, ya sean grados de preferencia, ordenar algún producto,... Cuando estas respuestas aparezcan de forma continua podemos hacer el doblado de los datos de igual forma. Basta expresar primero la escala continua en rangos.

4.10. Regresión

Hemos hablado de cómo representar las filas y columnas en un mapa a partir del AC y como interpretarlo. En esta sección trataremos las relaciones existentes entre filas y columnas de forma matemática y no solo gráfica.

Unas vez representado un punto en coordenadas estándares, podemos representar este punto como unas coordenadas cartesianas en un plano de la forma (x,y). La recta de regresión de Y sobre X (filas sobre columnas) tendrá como pendiente el coeficiente de correlación entre X e Y, es decir $\sqrt{\lambda_1}$, siendo λ_1 la primera inercia principal. La recta de regresión de X sobre Y tendrá como pendiente la inversa de este valor, es decir $1/\sqrt{\lambda_1}$.

Cuanto más parecidas sean las dos rectas de regresión mayor será la inercia, por lo que se cumple que las coordenadas estándares de las filas y columnas minimizan el ángulo entre la regresión de filas sobre columnas y la de columnas sobre filas.

Recordando que las coordenadas principales corresponden a perfiles y las estándares a vértices, podemos establecer la relación entre filas y columnas como:

$$f_{ik} = \sum_j \left(\frac{p_{ij}}{r_i} \right) \gamma_{jk} \tag{4.17}$$

Siendo f_{ik} la k-ésima coordenada principal de la fila i y γ_{jk} la k-ésima coordenada estándar de la j-ésima columna.

Análogamente:

$$g_{ik} = \sum_i \left(\frac{p_{ij}}{c_j} \right) \phi_{ik} \tag{4.18}$$

Siendo g_{jk} la k-ésima coordenada principal de la columna j y ϕ_{ik} la k-ésima coordenada estándar de la i-ésima fila.

El coeficiente k hace referencia a la dimensión en la que nos interesa interpretar los datos, por lo que habitualmente k=2.

Ya habíamos visto la relación entre las coordenadas principales y estándares:

$$f_{ik} = \sqrt{\lambda_k}\phi_{ik} \tag{4.19}$$

$$g_{ik} = \sqrt{\lambda_k}\gamma_{jk} \tag{4.20}$$

A partir de las expresiones (4.17), (4.18), (4.19), (4.20) llegamos a las siguientes ecuaciones para relacionar las coordenadas principales de filas y columnas:

$$f_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_j \left(\frac{p_i j}{r_i} \right) g_{ik} \tag{4.21}$$

$$g_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_i \left(\frac{p_i j}{c_j} \right) f_{ik} \tag{4.22}$$

Y análogamente para relacionar las coordenadas estándares de filas y columnas:

$$\phi_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_j \left(\frac{p_i j}{r_i} \right) \gamma_{jk} \tag{4.23}$$

$$\gamma_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_i \left(\frac{p_i j}{c_j} \right) \phi_{ik} \tag{4.24}$$

Podemos utilizar cualquiera de las ecuaciones anteriores para estimar por regresión las coordenadas de los puntos en el mapa a partir de los perfiles (variables explicativas). Realizaríamos una regresión lineal sobre los valores que tenemos buscando los coeficientes que mejor ajusten cualquiera de los modelos expresados en las ecuaciones (4.21), ..., (4.24). En regresión suponemos que los datos se ajustan a un modelo lineal y buscamos la estimación de los coeficientes que mejor ajustan esta recta. Aquí hacemos lo mismo estimando los coeficientes que hemos expresado antes.

Capítulo 5

Análisis de correspondencias alternativos

5.1. Análisis de correspondencias múltiple

Hasta ahora hemos considerado unas variables categóricas fila y otras columna diferentes entre sí. En este capítulo consideraremos el caso en que realizamos el análisis sobre variables similares, el **análisis de correspondencias múltiples (ACM)**. Analizaremos la relación entre dos o más variables en el contexto de un solo fenómeno de interés. Lo necesario es que las variables sean notablemente similares en su naturaleza o definición.

5.1.1. ACM a partir de la tabla binaria

Podemos pensar el ACM como un AC sobre la **matriz binaria** de los datos. Expresaremos esta matriz como lo hace Santiago de la Fuente Fernández en su libro *Análisis de correspondencias simples y múltiples.*

La matriz binaria es una matriz $Nx\left(\sum_{q=1}^{Q} J_q\right)$, donde N son los individuos o grupos, Q las variables y J_q el número de categorías dentro de la q-ésima variable.

Denotamos esta matriz por Z y se construye como:

$$z_{ij} = 1 \text{ , si el i-ésimo individuo pertenece a la categoría j}$$

$$z_{ij} = 0 \text{ , si el i-ésimo individuo NO pertenece a la categoría j}$$

Las respuestas son excluyentes entre sí, en el sentido de que para un individuo en una variable, la matriz binaria solo puede tomar el valor 1 en una sola de las J_q categorías y 0 en el resto dentro de cada variable. Definida esta matriz, realizamos el AC sobre ella y podemos interpretar el mapa de forma análoga a lo hecho hasta ahora. La diferencia reside en que ahora interpretamos las relaciones entre variables similares y no tomamos de referencia los vértices, simplemente interpretamos las posiciones de los puntos y la proximidad de unos a otros.

Podemos considerar la matriz binaria como una tabla concatenada donde solapamos Q subtablas, donde en cada subtabla para cada fila hay un 1 y el resto 0. Es decir, los marginales de las filas de cada subtabla es una columna toda de unos.

Vamos a introducir otro ejemplo para exponer este tipo de tablas. En este ejemplo (de nuevo a partir de los datos proporcionados por nuestro investigador) se recoge una puntuación para el nivel de ansiedad y el nivel de depresión en una escala de 0-9 para 820 pacientes. Construimos la matriz binaria asociada a estos datos y obtenemos la tabla del Cuadro 5.1 (en la que solo se representan las 5 primeras filas).

ID	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
3	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
5	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Cuadro 5.1: Primeras 5 filas de la tabla binaria con las notas de ansiedad y depresión

Notemos que en este ejemplo, $J_1 = J_2 = 10$, $Q = 2$ y $N = 820$ y que la tabla cumple las condiciones que hemos expuesto para ser considerada binaria. Ahora simplemente realizamos el AC como hemos introducido a lo largo de todo el trabajo, con la única salvedad de que la interpretación es distinta. Ahora no representamos vértices (no tendría sentido) y lo que interpretamos son las distancias y posiciones de los perfiles usando los propios perfiles

como referencia. Así diremos que un vértice está relacionado con otro si se encuentran cerca o que no lo están si se encuentran distanciados uno de otro.

El mapa del AC para el ejemplo del Cuadro 5.1 lo incluimos en la Figura 5.1.

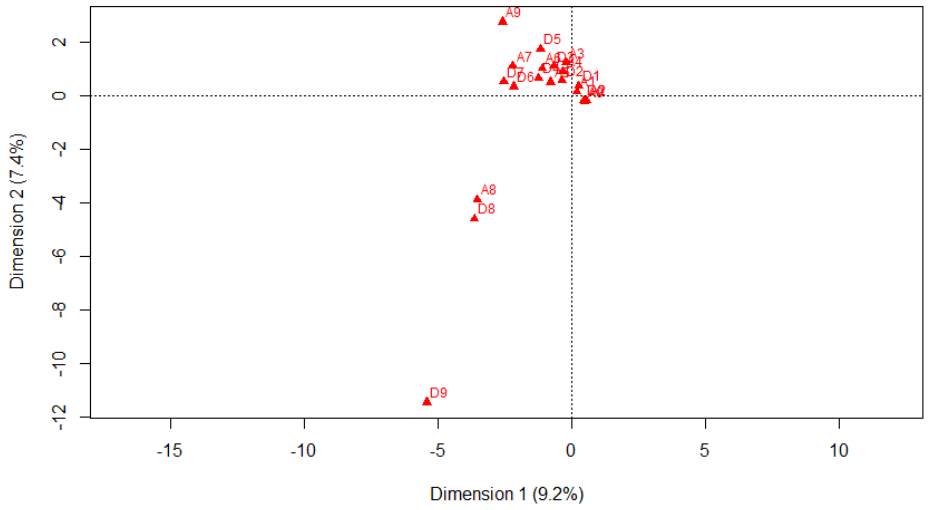

Figura 5.1: Mapa del AC para la tabla binaria del Cuadro 5.1

Este mapa explica un bajo porcentaje de la inercia por lo que no lo podemos interpretar con seguridad. Aquí la incluimos para mostrar como se razona este tipo de AC. La interpretación no resulta muy fácil al encontrarse muchos perfiles muy juntos, por tanto solo concluimos de esta tabla que los altos grados de depresión (8 y 9) no están relacionados con la puntuación de la ansiedad, salvo el nivel 8 que parece seguir relacionado. Podríamos razonar que la cercanía entre los perfiles indica que en los niveles bajos de depresión y ansiedad las dos enfermedades están bastante relacionadas salvo para el peor grado de depresión, para el cual probablemente afecten más factores que solo la ansiedad al tratarse del peor nivel para la enfermedad.

La inercia de la matriz binaria solo depende del número de variables y no de sus valores. Si $J = \sum_q J_q$, la matriz binaria es de tamaño QxJ. Los marginales de las filas para cada subtabla de tamaño NxJ_q son todos 1, por lo que estamos en el caso extremo en que todos los perfiles se corresponden con los vértices. Por tanto, como ya explicamos en el capítulo 2, la inercia de la subtabla será igual a su dimensionalidad, es decir, $J_q - 1$.

Utilizando el resultado anterior sobre tablas concatenadas donde definíamos la inercia de la tabla concatenada como la media de las subtablas, obtenemos:

$$inercia(Z) = \frac{1}{Q}\sum_{q} inercia(Z_q) = \frac{1}{Q}\sum_{q} J_q - 1 = \frac{J-Q}{Q} \tag{5.1}$$

5.1.2. ACM a partir de la tabla de Burt

Realizar el AC sobre la matriz de Burt es una forma alternativa de realizar el ACM. Esta matriz está compuesta por todos los posibles cruces dos a dos de las variables de interés. Se define como:

$$B = Z^T Z \tag{5.2}$$

Las tablas de la diagonal corresponden con el cruce de las variables consigo mismas. Es decir, en la diagonal obtenemos tablas con valores solo en la diagonal que se identifican con las frecuencias marginales de cada variable. Otra propiedad de la matriz B es que es simétrica y cuadrada, por lo que las soluciones para filas y columnas son las mismas.

Por la construcción de la matriz simétrica y de Burt arrojaran los mismos resultados, obteniendo las mismas coordenadas estándares para las variables categóricas. Las inercias principales del análisis de Burt son los cuadrados de los de la matriz binaria. Las coordenadas principales son las coordenadas estándares multiplicadas por la raíz de las inercias principales, por esto el mapa de la matriz simétrica y el de Burt proporcionan las mismas posiciones de los perfiles pero cambiando la escala. Esta escala encoje las coordenadas obtenidas por la matriz de Burt respecto a las de la matriz simétrica por el hecho de que, al ser siempre la $inercia \leq 1 \Rightarrow \sqrt{inercia} \leq inercia$. Esta es la explicación de que los perfiles del análisis de Burt se encuentren más próximos y que la matriz de Burt explique mayores porcentajes de inercia.

La inercia de la matriz de Burt se calcula como para la matriz simétrica. Hay QxQ tablas de dimensión $J_q - 1$ donde estamos en el caso extremo de asociación fila-columna, por tanto tendremos la inercia máxima, que es igual al número de dimensiones.

Para el ejemplo de la ansiedad y la depresión podemos construir la matriz de Burt como hemos indicado. Una vez tenemos esta matriz, de nuevo basta realizar el AC y proceder a

su análisis como se hizo para el caso de la matriz binaria. El mapa del AC que obtenemos a partir de la matriz de Burt se muestra en la Figura 5.2.

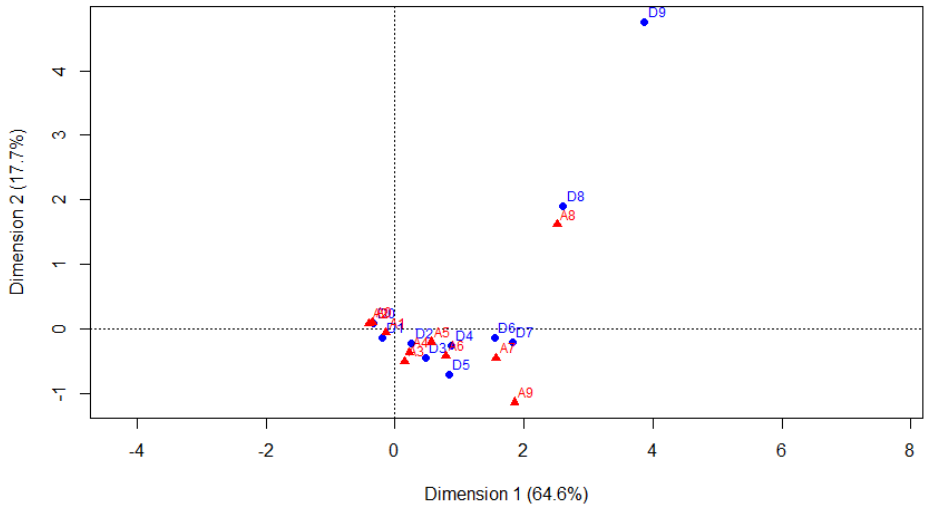

Figura 5.2: Mapa del AC para la tabla de Burt del Cuadro 5.1

Vemos como se conservan las posiciones relativas y como lo único que ha variado es su escala y su orientación. Esto resulta irrelevante pues lo que interpretamos en los mapas de estas matrices son las posiciones relativas entre los perfiles, que es lo que se conserva. Con un razonamiento análogo al hecho en la sección anterior llegamos exactamente a las mismas conclusiones.

Lo que ha cambiado sustancialmente son los porcentajes de inercia explicados por los ejes. El mapa construido a partir de la matriz binaria explicaba solo un 16,6 % de la inercia, mientras el mapa de la matriz de Burt explica el 82,3 %, por lo que este mapa si que es fiable. Esto concuerda con lo explicado anteriormente.

5.1.3. Escalado óptimo del ACM

Supongamos que tenemos Q variables, $J = \sum_q J_q$ categorías, S individuos y que las Q variables toman los valores de a_{11} a a_{1J_1}, de a_{21} a a_{2J_2}, ..., de a_{Q1} a a_{QJ_q}.

Para un cierto individuo le asignaríamos los valores a_{1i}, a_{2j}, ..., a_{Qs}. Así plasmaríamos todas las variables para cada individuo. Sean ahora a_1, ..., a_Q las respuestas de toda la muestra, donde a_i simboliza los S valores de la i-ésima variable.

Definimos la **puntuación** de cada individuo como $a_{1i} + a_{2j} + ... + a_{Qs}$ y la de toda la muestra como $a_1 + a_2 + ... + a_Q$. El criterio de búsqueda de los valores de escala óptimo será análogo al hecho en el AC simple, pero en este caso buscamos maximizar la media de las correlaciones al cuadrado entre las puntuaciones de las variables y la suma de las puntuaciones:

$$\text{correlaciones al cuadrado} = \frac{1}{4}[cor^2(a_1,\ a_1 + a_2 + ... + a_Q) + \cdots$$

$$+cor^2(a_2,\ a_1 + a_2 + ... + a_Q) + ... + cor^2(a_Q,\ a_1 + a_2 + ... + a_Q)] \tag{5.3}$$

De nuevo, necesitaremos las condiciones de identificación. Recordemos que son que la media $\sum_{i=1}^{Q} a_i = 0$ y $var(\sum_{i=1}^{Q} a_i) = 1$. Obtenemos la solución a este problema de maximización con las coordenadas estándares de las categorías de las variables en el primer eje principal del ACM.

Ya vimos que cada variable contribuye en un cierto porcentaje a la inercia principal, siendo la suma de estas el 100 %. Cada correlación individual al cuadrado de la fórmula (5.3) viene dada por:

(porcentaje de contribución a la inercia principal)x(Inercia principal)xQ

Y la correlación individual es la raíz del valor anterior. Si una variable presenta poca correlación con la puntuación total nos podemos plantear eliminar esta variable.

El ACM, al ser una aplicación del AC nos proporciona un mapa asimétrico. En este caso, es mejor interpretar las inercias principales y las correlaciones y no la geometría como hacíamos antes. Podemos de todas formas establecer estas interpretaciones y seguirán siendo válidas. En el mapa del ACM, cada individuo o caso se encontrará en la media de sus valores en las Q categorías.

En el caso en que una variable tenga poca correlación con las demás decimos que esta variable empeora la **fiabilidad** de la puntuación total. Una medida de la fiabilidad nace de

suponer una estructura en los datos. Definimos la **alfa de Cronbach** como una medida estándar de la fiabilidad, dada por:

$$\alpha = \frac{Q}{Q-1}\left(1 - \frac{\sum_q s_q^2}{s^2}\right) \tag{5.4}$$

Donde s_q^2 es la varianza de la puntuación de la variable q-ésima, q=1, ..., Q; y s^2 la varianza de las variables media. Si aplicamos esta fórmula a la primera dimensión hallada por el ACM:

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

Donde λ_1 es la primera inercia principal de la matriz binaria. De aquí deducimos que cuanto mayor sea esta primera inercia principal mayor será la fiabilidad.

En el ACM no obtenemos una solución con ejes anidados como en el AC simple. Es decir, la solución bidimensional no contiene necesariamente la mejor solución unidimensional como eje principal.

5.2. Análisis de correspondencias conjunto

En la sección anterior hemos tratado el difícil objetivo de pasar de una tabla de contingencia de dos entradas a una de más variables. Introducimos la matriz binaria y de Burt y realizamos el AC sobre estas matrices. El problema es que la inercia y las posiciones de los perfiles en estos AC no eran del todo bien interpretables y no estaban muy claros.

Al realizar el AC sobre la matiz de Burt los resultados se ven alterados por los valores en la diagonal al ser las frecuencias extremas que corresponden con los marginales de cada variable o grupo. Las celdas de la diagonal presentan una gran inercia y resulta innecesario intentar visualizar estas tablas al no aportarnos ninguna información extra. Suprimiendo estas tablas de la diagonal mejoramos la inercia explicada por el mapa

El **análisis de correspondencias conjunto (ACCo)** es un algoritmo del AC aplicado sobre la matriz de Burt ignorando las celdas de la diagonal. Realizamos el AC para la tabla de Burt y a continuación se sustituyen los valores de la diagonal por valores estimados a

partir de la fórmula de reconstrucción (4.14). Para la solución bidimensional y teniendo en cuenta que la matriz de Burt es simétrica la formula se trasnforma en:

$$\hat{p_{ij}} = c_i c_j \left(1 + \sqrt{\lambda_1}\gamma_{i1}\gamma_{j1} + \sqrt{\lambda_2}\gamma_{i2}\gamma_{j2}\right) \tag{5.5}$$

Donde $\hat{p_{ij}}$ es el valor estimado para la (i,j)-ésima celda de la matriz de Burt.

Esta fórmula la aplicamos a los elementos de la diagonal y construimos una matriz de Burt modificada, realizamos el AC sobre esta nueva matriz y obtenemos una nueva solución. A partir de esta matriz construimos una nueva tabla de Burt modificada utilizando la fórmula (5.5). Repetimos este algoritmo hasta llegar a la convergencia.

En cada interacción mejoramos el ajuste de las tablas fuera de la diagonal al eliminar el efecto que producían las frecuencias extremas de la diagonal original. Ya comentamos que en el ACM no se producen ejes anidados, pero en el ACCo se produce un anidamiento aproximado. Más adelante intentaremos "mejorar" el anidamiento de los ejes obtenidos por el ACCo.

5.3. Análisis de correspondencias de subgrupos

Ya comentamos que nos puede interesar en algunos casos suprimir o obviar una parte de los datos por algún motivo práctico. También puede darse la situación de que nos interese estudiar solo un subgrupo de los datos por la naturaleza o características de estos. Por ejemplo, estudiar sobre nuestro ejemplo de las enfermedades las diferencias entre las enfermedades vasculares.

Podríamos pensar en realizar el AC sobre la subtabla directamente, pero al proceder de esta manera podrían cambiar los valores marginales y en consecuencia los perfiles, las masas y las distancias. Por esto, utilizaremos los datos de la matriz original en el AC de subgrupos.

Para evitar que la perdida de información redunde en un mapa erróneo lo que hacemos es realizar el AC sobre la tabla original y representar solo los perfiles que nos interesan. Esta es una alternativa muy útil para solucionar algunos problemas como tener muchos

perfiles y no poder interpretarlos bien o no poder colocar adecuadamente etiquetas en el mapa. También nos ayudará para visualizar mejor algunas relaciones al solo ver los perfiles que nos interesan.

5.4. Análisis de tablas cuadradas

Este caso se produce cuando filas y columnas hacen referencia a las mismas variables. Por la naturaleza de la tabla, los valores más elevados en las frecuencias se hallan en la diagonal, lo cual puede enmascarar efectos menos fuertes entre los cruces fuera de la diagonal. Suelen presentar altos valores de inercia debido a esta fuerte relación entre las filas y columnas. Para evitar que esto suceda, dividimos el análisis de estas tablas en dos partes: un análisis de la parte simétrica (donde aparece la diagonal) y otro de la parte antisimétrica que revela las asociaciones fuera de la diagonal.

Por la alta inercia de la tabla utilizaremos de forma adecuada los mapas asimétricos. El problema es la diagonal que enmascara efectos más sutiles entre filas y columnas. Esto sucede a raíz de que si descomponemos la inercia, la de la diagonal será mucho más elevada que la de los elementos fuera de la diagonal, es decir, los elementos fuera de la diagonal están pobremente explicados en comparación.

Sea N la matriz de los datos. Un resultado algebraico nos proporciona una descomposición de cualquier matriz como:

$$N = \frac{1}{2}(N + N^T) + \frac{1}{2}(N - N^T) = S + T \tag{5.6}$$

Siendo S una matriz simétrica y T antisimétrica. A estas matrices las llamamos respectivamente parte simétrica y antisimétrica de N.

Debido a la naturaleza de la tabla debemos buscar un algoritmo alternativo para realizar el AC. Lo primero que hacemos es obtener una nueva matriz conformada de la siguiente forma:

$$\begin{pmatrix} N & N^T \\ N^T & N \end{pmatrix} \tag{5.7}$$

Una vez tenemos esta matriz realizamos el AC sobre ella. Si la matriz N es de tamaño SxS, la nueva tabla tendrá una dimensionalidad de $2S-1$. Cuando descomponemos las inercias principales asociadas a todos los $2S-1$ ejes nos encontraremos con valores de inercia que aparecen a pares y otros individualmente. Los valores que aparezcan a pares serán los asociados a la parte antisimétrica de la tabla y los que aparezcan solos a la simétrica.

El AC de la parte simétrica lo interpretaremos de la manera habitual, es decir, el mapa mostrará la asociación entre variables de la tabla.

El AC de la parte antisimétrica tiene una interpretación especial debido a las diferencias en la naturaleza de la tabla con las vistas hasta ahora. La primera diferencia se produce por la igualdad de las inercias principales, lo que provoca que las coordenadas puedan girar por el mapa. Por este motivo, en este AC no dibujamos los ejes. La segunda se debe a la característica antisimétrica de la matriz, lo que hace que los resultados aparezcan repetidos con signos opuestos. Solo representaremos uno de los dos pares de signos.

Llamaremos **flujo** al fenómeno que expresa la relación entre las filas y las columnas de la tabla original, en el sentido de que si existe un flujo entre una determinada fila i y columna j estaremos expresando que la frecuencia del punto (i,j) de la tabla varía mucho respecto al (j,i). La interpretación una vez configurado el mapa con las sutilezas comentadas anteriormente no se realiza a partir de las distancias. Lo que se hace es definir regiones triangulares definidas por dos puntos y el origen. Los triángulos grandes expresarán un flujo entre las filas y columnas que definen los triángulos y los triángulos pequeños indicarán la ausencia de flujo entre esas filas y columnas.

Recurrimos de nuevo al ejemplo sobre la ansiedad y depresión. Hecho el AC sobre la matriz de recuentos de las respuestas de la ansiedad y depresión (donde en cada celda (i,j) aparece el número de individuos con puntuación i-1 en depresión y j-1 en ansiedad) construimos la matriz como la indicada en la fórmula (5.7). Una vez realizado el AC sobre esta tabla obtenemos las inercias principales en las 2x10-1 dimensiones de la tabla. Las inercias principales para cada eje se muestran en el Cuadro 5.2.

Del Cuadro 5.2 deducimos que el primer eje está asociado a la parte simétrica de la tabla, la segunda y la tercera a la parte antisimétrica y así sucesivamente. Una vez identificados qué ejes corresponden con cada parte, solo nos queda interpretar los resultados.

Dimensión	Inercia Principal
1	0.3718
2	0.0799
3	0.0799
4	0.0561
5	0.0345
6	0.0345
7	0.0253
8	0.0253
9	0.0142
10	0.0124
11	0.0095
12	0.0052
13	0.0052
14	0.0038
15	0.0027
16	0.0016
17	1.8x10-50
18	1.8x10-50
19	3x10-6000
TOTAL	0.7622

Cuadro 5.2: Tabla con las 19 inercias principales

Representando los primeros ejes de la parte simétrica obtenemos el mapa de la Figura 5.3 y representando los primeros ejes para la parte antisimétrica obtenemos el mapa de la Figura 5.4.

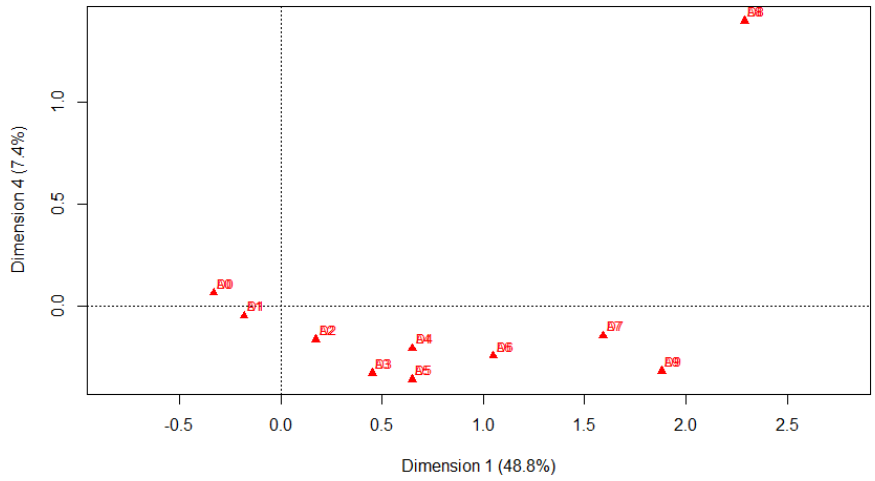

Figura 5.3: Mapa de la parte simétrica para la tabla cuadrada de ansiedad y depresión

Como ya hemos comentado, el análisis de la parte simétrica se realiza de la manera habitual y por tanto no realizaremos la interpretación de este mapa. El caso nuevo es el mapa asociado a la parte antisimétrica de la tabla. En este mapa consideraremos los flujos entre perfiles. Hemos representado en el mapa un triángulo que une el eje, el perfil de A8 y D8. Es uno de los triángulos de mayor área que podemos trazar considerando como vértices del triángulo el origen y dos perfiles, uno para cada variable. Interpretamos triángulos con un área grande como perfiles entre los que existe un flujo.

En este caso, no hay triángulos muy grandes, por lo que interpretamos que las frecuencias de los pacientes con un nivel de ansiedad y depresión se mantienen en toda la tabla. El único punto para el que podría haber flujo es para el mayor nivel de ansiedad, mostrando que las frecuencias del mayor grado de ansiedad son distintas dependiendo del grado de depresión.

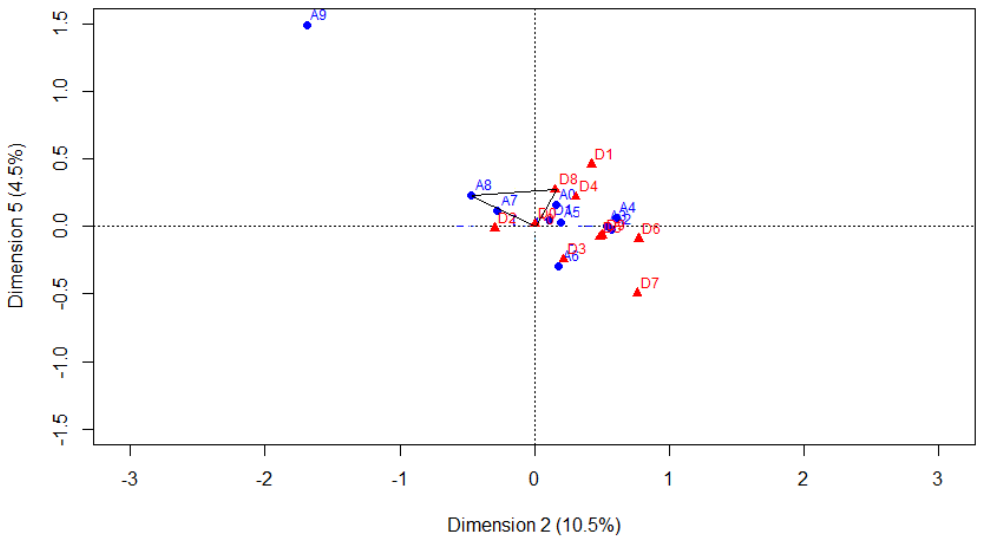

Figura 5.4: Mapa de la parte antisimétrica para la tabla cuadrada de ansiedad y depresión

Capítulo 6

Inferencia

Hasta ahora hemos considerado en algunos casos la eliminación de puntos por ejemplo. Lo que no hemos tratado es cómo afecta esta eliminación al AC. Ya vimos en el capítulo 4 las contribuciones a la inercia de los puntos.

Si un punto tiene mucho peso en la configuración del mapa puede modificarlo sustancialmente, así como un punto con menos influencia puede ser obviado en cuanto a que no varían mucho los resultados. Para comprobar si la supresión de uno o varios puntos afecta significativamente a la configuración del mapa efectuamos el AC omitiendo los puntos y vemos como se ven afectados los resultados.

Otro tema importante es la variabilidad de la muestra que tenemos en la tabla de contingencia. ¿Tenemos datos sobre una parte de la población elegida de forma aleatoria o siguiendo algún criterio de selección? Lo idílico sería poder realizar el AC muchas veces sobre muchas muestras, pero por lo general es imposible. A raíz de esto nos formulamos la pregunta, ¿caracteriza el AC de esta muestra a la población o es resultado del azar?

6.1. Bootstrapping

Esta sección ha sido desarrollado a partir de la consulta al libro *An introduction to the bootstrap* de Efron, B. y Tibshirani, R. J., donde se introduce el concepto de bootstrap y su

desarrollo en profundidad. Nuestro libro de referencia no incide mucho en este apartado, por lo que el trabajo de Efron, B. y Tibshirani, R. J. nos ha sido de gran utilidad para la comprensión y posterior explicación del bootstrap.

No podemos realizar remuestreos sobre la población, por lo que recurrimos al **bootstrapping**. Consiste en considerar los datos que tenemos como si fuesen la población total y crear una nueva muestra tomada de esta. Realizaríamos la toma del mismo número de datos que para el muestreo original tomándolos de uno en uno con reemplazamiento de los datos originales. De esta manera las frecuencias de la nueva muestra tenderán a parecerse bastante a las que ya teníamos. Podemos repetir este automuestreo tantas veces como queramos, siendo lo habitual entre 100 y 1000 veces.

Otra alternativa es recurrir al **automuestreo multinomial**, que consiste en tomar los nuevos datos de una supuesta población en donde cada variable tiene como probabilidad de ser tomada su frecuencia en la tabla original.

Realizar el AC sobre las 100-1000 tablas de las que disponemos ahora y compararlas sería muy laborioso y no tendría una fácil interpretación. La alternativa más sencilla consiste en el **automuestreo parcial**. Consiste en representar los perfiles de todas las tablas que hayamos construido en el mismo mapa. Una vez representados los perfiles de todas las tablas, representamos **perímetros convexos** que contienen todos los perfiles de una cierta fila en cada tabla. Por ejemplo, si remuestreamos hasta 100 tablas, para la fila i tendríamos 100 perfiles fila distintos, cada uno asociado a las coordenadas calculadas para esa fila en una de las 100 tablas. Podemos encerrar los 100 perfiles fila dentro de un perímetro convexo delimitado por los perfiles más externos.

Habitualmente se eliminan valores atípicos, construyendo perímetros convexos que contengan el 95 % de los datos más internos. Este perímetro constituye entonces una región de confianza al 95 % para el perfil fila correspondiente.

La interpretación después de representar todos los perímetros convexos (uno para cada fila) es que si dos perímetros convexos no se solapan tenemos bastante seguridad de que las filas son significativamente distintas. No podemos hacer el razonamiento análogo en el caso de que dos perímetros se solapen, pues al estar representando una "proyección" de los datos, podrían no estar solapados en otra dimensión no representada y por tanto no estar relacionadas las filas.

Una alternativa para construir los perímetros convexos es a partir del **método Delta**. Con esta herramienta calculamos las varianzas y covarianzas de las coordenadas de todas las tablas y suponiendo una distribución normal bivariante calculamos elipses de confianza en el plano. Estas elipses tenderán a parecerse a los perímetros convexos de los que hablamos antes.

Podemos realizar más contrastes sobre la tabla de contingencia y los resultados del AC que los expuestos anteriormente. Resultan de gran interés para establecer un análisis más profundo y llegar a un mejor análisis de los datos. Existen herramientas para justificar estadísticamente las conclusiones que hayamos extraído del AC, aquí destacamos a mayores del bootstraping:

6.2. Prueba de distribución asintótica

Es utilizada para contrastar la significación de la primera inercia principal. Utilizamos la misma prueba que la expuesta en el capítulo 3. Hallamos el estadístico $\chi^2 = (Inercia) \cdot N$ y contrastamos con la distribución χ^2 con los grados de libertad correspondientes si se cumple la hipótesis nula de que la primera inercia principal no es significativa, lo cual será análogo a la homogeneidad de la tabla.

6.3. Test de permutaciones

En este test comenzamos definiendo una medida que nos interese contrastar. Por ejemplo, para contrastar la agrupación de varios perfiles medimos la distancia entre ellos (agrupación en el sentido de encontrarse cerca todos los perfiles). A continuación, generamos todas las posibles tablas permutando las agrupaciones y para estas calculamos de nuevo las distancias entre los perfiles. Con todas estas medidas se construye la distribución del estadístico de contraste para el test de permutación. Tomamos como p-valor el porcentaje de permutaciones que produzcan una distancia menor entre los perfiles que la que tenemos en la tabla de la agrupación contrastada. Si el p-valor de contraste es bajo, aceptamos la hipótesis alternativa de que las variables están efectivamente agrupadas.

La linea de planteamiento del ejemplo anterior se reproduce para cualquier contraste que queramos hacer. Basta definir una medida que nos interese y contrastar la hipótesis nula que queramos comprobar testando su significación frente al resto de posibles permutaciones de los datos.

6.4. Simulación de Monte Carlo

El test de permutaciones es un test exacto, en el sentido de que realizamos todas las comparaciones y el p-valor es el porcentaje entre todas las posibilidades. El problema de este test es que cuando el tamaño muestral es grande resulta complicado de llevar a cabo. Para solucionar esto nace la simulación de Monte Carlo, que en vez de tomar todas las posibles muestras considera solo varias tomadas de forma aleatoria. Resulta mucho más sencillo con el coste de que deja de ser un test exacto.

La prueba no es insesgada, por lo que hay que hacer la corrección:

$$p - valor = \frac{r+1}{t+1}$$

Donde r es el número de permutaciones que superan el valor de nuestro estadístico observado y t el número de permutaciones usadas en la simulación.

En el AC se utilizará habitualmente para contrastar la hipótesis nula de que no existe asociación entre filas y columnas. De ser así, como ya vimos, los valores de las celdas deberían parecerse bastante al producto de los valores marginales correspondientes. Construimos una nueva tabla con los valores esperados a partir de los datos originales, y con esta tabla obtendremos muestras simuladas del mismo tamaño que la original.

Sea t el número de tablas que construimos. Realizamos el AC para estas t tablas y calculamos las inercias principales de cada una. Para el caso de la primera inercia principal, sea r el número de tablas cuya primera inercia principal es mayor que la de la tabla original. Estimamos el p-valor utilizando la fórmula anterior $\frac{r+1}{t+1}$. A un nivel de significación α, rechazamos la hipótesis nula de que la primera inercia principal no es significativa si $\frac{r+1}{t+1} < \alpha$. Podemos proceder de forma análoga con la inercia total para testar la hipótesis nula de que no exista asociación entre filas y columnas o con la segunda inercia principal para testar si esta es significativa.

6.5. Agrupaciones

Durante todo el trabajo hemos considerado la representación e interpretación de los datos en los mapas. En este capítulo nos centraremos en la herramienta estadística que nos permite contrastar las igualdades o diferencias entre grupos.

Ya se explicó que agrupar puntos conlleva una perdida de inercia y la separación en varios subgrupos de un grupo su aumento. Comencemos introduciendo el concepto de **inercia intergrupos**, una medida de la variabilidad entre los grupos que estemos considerando (una vez agrupados los datos originales de alguna forma).

Llamamos **inercia intragrupos** a la diferencia entre la inercia total y la intergrupos. Este concepto nos dará una idea de la variabilidad que perdemos al realizar la agrupación.

Un resultado importante sobre las medidas anteriores es que:

$$\text{Inercia total} = \text{Inercia intergrupos} + \text{Inercia intragrupos} \tag{6.1}$$

Definimos la inercia en la fórmula (3.4) como:

$$\sum_i r_i d_i^2$$

Siendo r_i la masa y d_i la distancia χ^2 entre el i-ésimo perfil fila y el centroide c de los datos.

La inercia intergrupos se calcula de forma análoga:

$$\sum_g \overline{r_g}\overline{d_g^2} \tag{6.2}$$

Siendo los coeficientes los mismos que para el calculo de la fórmula (3.4) pero de la tabla con los datos una vez agrupados. Los perfiles $\overline{a_g}$ tienen el mismo centroide c que los perfiles originales.

Por tanto, la inercia de cada grupo g la podemos calcular como:

$$\sum_{i \in g} r_i d_{ig}^2 \tag{6.3}$$

Donde d_{ig} es la distancia χ^2 entre el perfil de la fila i del grupo g al centroide. Sumando los valores de la fórmula (6.3) para todos los grupos, obtenemos la inercia intragrupos.

Recopilando todo lo anterior, obtenemos lo que queríamos demostrar:

$$\sum_g \overline{r_g}\overline{d_g^2} + \sum_g \sum_{i \in g} r_i d_{ig}^2 \; = \; \sum_i r_i d_i^2 \tag{6.4}$$

De esta forma, podemos calcular la inercia intragrupos realizando el AC para la tabla original y la tabla agrupada y restando estos valores. Notemos que la inercia intragrupos de cualquier variable agrupada solo consigo misma es cero.

El objetivo entonces será agrupar los datos de forma que obtengamos la máxima inercia intergrupos y la mínima intragrupos. Para ello utilizamos el **algoritmo de agrupación**, que agrupa en cada paso al par de variables cuya agrupación conlleve a la menor reducción de la inercia intergrupos, es decir, a la menor reducción del estadístico χ^2. El algoritmo termina cuando se agrupan todas las filas y tenemos un solo grupo.

El concepto de que dos filas se agrupen por conllevar la menor reducción del estadístico χ^2 se traduce en que estas dos filas son bastante parecidas y cuanto menor sea la reducción del estadístico, mayor será este parecido. Ahora lo que nos queda es decidir qué filas es razonable agrupar y en que punto del algoritmo debemos parar. Está claro que el caso en que agrupemos todos los grupos de la tabla en uno no será habitual.

Como herramienta para decisión utilizaremos las representaciones en árbol. Comenzamos calculando el valor de la distribución ji-cuadrado de la tabla original para un cierto valor de confianza. Si el estadístico nos permite rechazar la hipótesis nula de homogeneidad de la tabla continuamos con el proceso, si aceptamos la hipótesis de homogeneidad no tiene sentido continuar pues no existen diferencias significativas entre perfiles y no llegaremos a ninguna agrupación nueva. A continuación realizamos el AC para todas las tablas con los datos agrupados a partir del algoritmo. Tomamos como umbral el valor del cuartil a un cierto nivel de confianza de la distribución χ^2 con los grados de libertad de la tabla original. Si el estadístico χ^2 para alguna tabla reducida es mayor que el el cuartil considerado, tendremos una prueba estadística de que estos grupos presentan similitudes significativas entre ellos, por lo que tiene sentido la agrupación.

El diagrama de árbol es de la forma del mostrado en la Figura 6.1.

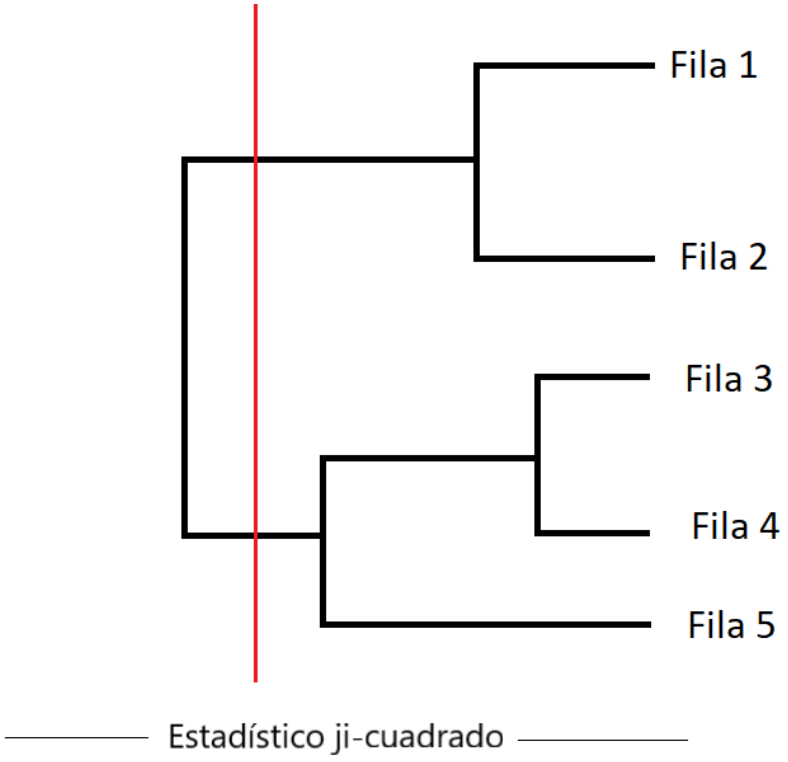

Figura 6.1: Diagrama en árbol para la agrupación de 5 filas

Donde la linea roja marca el valor del cuartil de la distribución ji-cuadrado de la tabla original. En el ejemplo de la Figura 6.1 solo tendríamos evidencias estadísticamente significativas para la agrupación de las filas (1,2) y (3,4,5), pues es la única agrupación que mantiene un valor del estadístico χ^2 superior al cuartil de la distribución ji-cuadrado al nivel de confianza considerado.

La tabla original mostraba diferencias significativas entre los perfiles al tener un estadístico χ^2 mayor que el cuartil para el nivel de confianza considerado. Por tanto, en algún punto de la tabla tenían que existir estas diferencias. Al realizar todas las agrupaciones vemos que una tabla agrupada es significativamente heterogénea, por lo que estos eran los perfiles que provocaban la alta inercia.

Capítulo 7

Aplicación práctica del AC en un estudio

Vamos a aplicar todo lo introducido hasta ahora en un caso práctico. Trataremos el AC aplicado a un estudio del grupo AEGIS (A Estrada Inflammation and Glycation Study), cuyo investigador principal es Francisco Gude de la Unidad de Epidemiología Clínica del Hospital Clínico Universitario de Santiago de Compostela. La base de datos introducida en el primer capítulo de este trabajo era un subconjunto de todos los datos recogidos en este estudio. A las enfermedades comunes, fármacos y notas de ansiedad y depresión se le suman las respuestas a un test de resiliencia, la edad, sexo, nivel de estudios y estado civil de los pacientes.

Todos los cálculos, construcciones y mapas serán generados con el software estadístico R. Principalmente utilizaremos el paquete **ca** para realizar todos los cálculos del AC, la función "pchisq", que nos devuelve el p-valor de nuestro estadístico para los grados de libertad correspondientes y el propio programa para construir las variables de interés, los cruces y las regiones convexas.

Construimos las tablas de contingencia para los cruces multiplicando la subtabla traspuesta asociada a las variables de interés con otra subtabla. Vamos a trabajar con tablas en las que aparece un 1 si el paciente pertenece a ese grupo y un 0 si no. De ahí que este producto de matrices nos proporcione los recuentos para cada grupo en cada variable.

Respecto a los datos, comenzamos definiendo la resiliencia, la cual consiste en la capacidad de una persona de superar adversidades sobreponiéndose a ellas e incluso saliendo reforzado de éstas. Una persona con alta resiliencia tendrá la capacidad de superar los problemas que se le presenten y sabrá lidiar con ellos. Al contrario, una persona con baja resiliencia se verá superado por las adversidades y tendrá dificultades para superarlas. Suele asociarse una alta resiliencia a personas que han pasado por muchas situaciones difíciles o se han visto obligadas a superarse y a personas que de por sí lidian mejor con los problemas.

La base de datos dispone de 820 pacientes considerados. Tras un análisis exploratorio de los datos, lo primero que haremos será agrupar varias enfermedades por categorías para corregir la baja frecuencia que presentan algunas. En el trabajo ya se comentó el inconveniente de estos perfiles, por ello realizamos la agrupación. También eliminamos las variables de los anticonceptivos orales y de la fenitoína por ser fármacos sin relación con los demás y no ser de interés. Tampoco consideramos la variable del nivel de estudios pues no contiene ningún tipo de información relevante ni está relacionada con las demás variables de interés.

En el Cuadro 7.1 presentamos un resumen de todos los datos que utilizaremos a lo largo de toda esta parte práctica. Sexo se codifica como 1 para hombre, 0 para mujer y sobre estado civil no tenemos referencias (esta variable no influirá en los estudios posteriores por lo que no es un problema no conocer la codificación). En la variable "Cardio" se han agrupado todas las enfermedades cardíacas (insuficiencia cardíaca, cardiopatía isquémica, enfermedad vasculo cerebral y hipertensión arterial), en "Arterial" la hiperlipemia y la arteriopatía periférica, en "Pulmon" el asma y la bronquitis crónica, en "Cuerpo" las enfermedades de reuma y osteoporosis, en "Piel" la psoriasis y la dermatitis, en "Hormona" el hipertiroidismo y la diabetes, y el resto de enfermedades son las introducidas ya en el primer capítulo.

En los fármacos, el número corresponde con la dosis que toma el paciente de ese fármaco. Nosotros en este estudio solo consideraremos el hecho de tomar o no determinado fármaco sin tener en cuenta la dosis. El número de tomas de los medicamento no es una variable tan relevante como tomar el medicamento en sí, por lo que no tendremos problema al realizar el estudio de esta forma.

Las variables "ER" corresponden con las 10 preguntas del test de resiliencia. La variable "Resiliencia" se construye sumando las puntuaciones da las respuestas a las 10 preguntas, que están en una escala de 1 a 5 y transformaremos en una escala de 0 a 4. Así, la mayor puntuación posible de resiliencia es 40 y la más baja 0.

	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
Edad	Min. :18.00	1st Qu.:35.00	Median :48.00	Mean :48.61	3rd Qu.:61.25	Max. :88.00
Sexo	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.4341	3rd Qu.:1.0000	Max. :1.0000
Estado civil	Min. :1.000	1st Qu.:1.000	Median :1.000	Mean :2.587	3rd Qu.:5.000	Max. :6.000
Cardio	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.2683	3rd Qu.:1.0000	Max. :1.0000
Arterial	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.2561	3rd Qu.:1.0000	Max. :1.0000
Pulmon	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.07317	3rd Qu.:0.00000	Max. :1.00000
Cuerpo	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.03659	3rd Qu.:0.00000	Max. :1.00000
Piel	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.04878	3rd Qu.:0.00000	Max. :1.00000
Hormona	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.1585	3rd Qu.:0.0000	Max. :1.0000
Ins.Ren	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.01585	3rd Qu.:0.00000	Max. :1.00000
Hepat.	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.04756	3rd Qu.:0.00000	Max. :1.00000
Depre	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.1378	3rd Qu.:0.0000	Max. :1.0000
Cancer	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.0378	3rd Qu.:0.0000	Max. :1.0000
Benzo	Min. :0.000	1st Qu.:0.000	Median :0.000	Mean :0.139	3rd Qu.:0.000	Max. :2.000
Antidepre	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.1317	3rd Qu.:0.0000	Max. :2.0000
MetGluc	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.5537	3rd Qu.:1.0000	Max. :7.0000
ado1	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.07927	3rd Qu.:0.00000	Max. :1.00000
Antiinf	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.08902	3rd Qu.:0.00000	Max. :1.00000
Insul2	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.01341	3rd Qu.:0.00000	Max. :1.00000
hipSed	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.1195	3rd Qu.:0.0000	Max. :4.0000
Estatinas	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.1646	3rd Qu.:0.0000	Max. :1.0000
cortic	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.006098	3rd Qu.:0.000000	Max. :1.000000
betabloq	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.05366	3rd Qu.:0.00000	Max. :1.00000
diureticos	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.09878	3rd Qu.:0.00000	Max. :1.00000
Glucosamin	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.007317	3rd Qu.:0.000000	Max. :1.000000
antipsic	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.008537	3rd Qu.:0.000000	Max. :1.000000
Nota Ansiedad	Min. :0.000	1st Qu.:0.000	Median :0.000	Mean :1.578	3rd Qu.:3.000	Max. :9.000
Nota Depresión	Min. :0.000	1st Qu.:0.000	Median :0.000	Mean :1.138	3rd Qu.:1.000	Max. :9.000
ER1	Min. :1.000	1st Qu.:4.000	Median :5.000	Mean :4.326	3rd Qu.:5.000	Max. :5.000
ER2	Min. :1.000	1st Qu.:3.000	Median :4.000	Mean :3.979	3rd Qu.:5.000	Max. :5.000
ER3	Min. :1.000	1st Qu.:3.000	Median :3.000	Mean :3.388	3rd Qu.:4.000	Max. :5.000
ER4	Min. :1.000	1st Qu.:3.000	Median :4.000	Mean :3.667	3rd Qu.:5.000	Max. :5.000
ER5	Min. :1.000	1st Qu.:4.000	Median :4.000	Mean :4.152	3rd Qu.:5.000	Max. :5.000
ER6	Min. :1.000	1st Qu.:3.000	Median :4.000	Mean :3.734	3rd Qu.:5.000	Max. :5.000
ER7	Min. :1.00	1st Qu.:2.00	Median :3.00	Mean :3.21	3rd Qu.:4.00	Max. :5.00
ER8	Min. :1.000	1st Qu.:2.000	Median :3.000	Mean :3.195	3rd Qu.:4.000	Max. :5.000
ER9	Min. :1.000	1st Qu.:3.000	Median :4.000	Mean :3.776	3rd Qu.:5.000	Max. :5.000
ER10	Min. :1.000	1st Qu.:3.000	Median :3.000	Mean :3.484	3rd Qu.:4.000	Max. :5.000
Resiliencia	Min. : 1.00	1st Qu.:22.75	Median :27.00	Mean :26.91	3rd Qu.:32.00	Max. :40.00

Cuadro 7.1: "Summary" de los datos que consideraremos en este estudio

El objetivo a priori es el de establecer asociaciones entre las enfermedades más comunes, el nivel de ansiedad y depresión, la resiliencia y la edad. Se trata de un estudio novedoso en cuanto a que no existen prácticamente trabajos con estos contrastes ni se han testado de esta forma. Además la extensa base de datos hace que sea un estudio muy fiable y del que podemos obtener conclusiones muy valiosas de cara a la investigación médica de las enfermedades más comunes, sabiendo qué variables son factores de riesgo y se asocian a determinadas enfermedades. Por ejemplo, trataremos de encontrar y demostrar la asociación que existe entre la resiliencia de una persona y las enfermedades que padezca. De llegar a alguna asociación, empezaría a considerarse la resiliencia como un factor de protección ante el padecimiento de determinadas enfermedades.

A partir de las notas de ansiedad y depresión podemos considerar si una persona tiene ansiedad o depresión a partir de unos valores que se consideran críticos. Desde el punto de vista médico, se considera que una persona padece de depresión si su nota es mayor o igual a 4. De igual forma, se considera que una persona tiene ansiedad si su nota es mayor o igual a 2. Con estos umbrales, construimos dos nuevas variables que utilizaremos más adelante, que son las que indican si una persona padece o no ansiedad o depresión. En el Cuadro 7.2 mostramos los recuentos para estas variables.

	Recuento
NO ANS.	638
ANS	182
NO DEPRE	616
DEPRE	204

Cuadro 7.2: Tabla con los recuentos para los pacientes con ansiedad y depresión

Hacemos con la resiliencia lo mismo que con la ansiedad y depresión. El problema en este caso es que no existe un consenso o escala universal para determinar si un nivel es alto o bajo. En este trabajo consideraremos como nivel bajo de resiliencia si la nota es inferior a la media de la resiliencia de todos los pacientes menos la desviación típica. Análogamente, consideramos un nivel alto de resiliencia si su valor está por encima de la media total más su desviación típica y un nivel medio si se encuentra entre los dos intervalos anteriores. El rango que consideraremos de esta forma es: $[0, 26'91-7'22]$, $[26'91-7'22, 26'91+7'22]$, $[26'91+7'22, 40] = [0, 19'69]$, $[19'69, 34'13]$, $[34'13, 40]$. Así, por ejemplo, un paciente con una nota de resiliencia de 15 se considerará que tiene baja resiliencia. En el Cuadro 7.3 mostramos la tabla con los recuentos para los niveles de resiliencia.

	Recuento
Resiliencia baja	113
Resiliencia media	590
Resiliencia alta	117

Cuadro 7.3: Tabla con los recuentos para los niveles de resiliencia

7.1. AC del estado civil respecto al resto de variables

Una vez tenemos todos nuestros datos codificados de manera conveniente comenzamos a realizar el AC para distintos cruces de los datos. Comenzamos con el cruce del estado civil con las enfermedades y los niveles de resiliencia. Construimos las tablas de contingencia asociadas a estos cruces tal como indicamos al comienzo del capítulo. Tomamos la fila asociada al estado civil y construimos su matriz binaria asociada. En esta matriz aparece un 1 si el paciente pertenece a ese estado civil y un 0 si no. Transponiendo esta tabla y multiplicándola matricialmente por la subtabla de las enfermedades, obtenemos la tabla del Cuadro 7.4. Para la resiliencia también construimos la matriz binaria, donde para cada paciente se indica con un 1 que tiene ese nivel de resiliencia. De nuevo multiplicamos la tabla binaria del estado civil traspuesta por la binaria de la resiliencia y obtenemos la tabla del Cuadro 7.5.

	Cardio	Arterial	Pulmon	Cuerpo	Piel	Hormona	Ins.Ren	Hepat.	Depre	Cancer
EC1	135	131	36	19	24	79	11	25	68	19
EC2	6	9	0	2	1	2	0	0	3	0
EC3	16	23	7	3	4	11	0	4	7	5
EC4	10	8	2	2	2	5	0	3	3	3
EC5	4	4	3	0	0	5	0	0	0	0
EC6	49	35	12	4	9	28	2	7	32	4

Cuadro 7.4: Tabla de contingencia para el cruce del estado civil con las enfermedades

	resi.baja	resi.media	resi.alta
EC1	69	359	72
EC2	5	14	2
EC3	9	61	18
EC4	3	23	3
EC5	2	11	3
EC6	25	122	19

Cuadro 7.5: Tabla de contingencia para el cruce del estado civil con los niveles de resiliencia

Una vez construidas las tablas de contingencia de los cruces contrastamos la hipótesis nula de homogeneidad de la tabla. En el caso de cumplirse esta hipótesis nos indicaría que las frecuencias de las enfermedades estudiadas o de los niveles de resiliencia son similares entre todos los estados civiles. De incumplirse la hipótesis de homogeneidad, nos indicaría que la frecuencia de alguna enfermedad o nivel de resiliencia varía dependiendo del estado civil. El contraste se realiza como indicamos en el capítulo 3, construyendo el estadístico ji-cuadrado y buscando su p-valor asociado en la distribución ji-cuadrado con tantos grados de libertad como $(n^o filas - 1)x(n^o columnas - 1)$.

Con el comando **ca** de R realizamos al AC para las dos tablas y nos devuelve el valor de la inercia total. Multiplicando la inercia por n (el número de casos considerados para la tabla del Cuadro 7.4 y 7.5 respectivamente) obtenemos el estadístico ji-cuadrado para cada una de las dos tablas. Para la tabla del Cuadro 7.4, $\chi^2 = 0{,}044719 * 886 = 39{,}62$, con un p-valor asociado en la distribución ji-cuadrado con $9 * 5 = 45$ grados de libertad de 0,7. Para la tabla del Cuadro 7.5 tenemos $\chi^2 = 0{,}009061 * 820 = 7{,}43$, con un p-valor asociado en la distribución ji-cuadrado con $5*2 = 10$ grados de libertad de $0,68$. Por tanto, en ambas tablas aceptamos la hipótesis nula de homogeneidad y concluimos que el estado civil no influye en el padecimiento de las enfermedades ni en los niveles de resiliencia.

Por completitud realizamos también el AC para la tabla que cruza el estado civil con el padecer o no ansiedad y depresión. La tabla de contingencia de este cruce se muestra en el Cuadro 7.6. Para este cruce volvemos a concluir que el estado civil no influye en tener o no ansiedad o depresión. Las frecuencias de ambas enfermedades se mantienen homogéneas en todos los estados civiles, lo que deducimos de que el estadístico $\chi^2 = 8{,}58$ deje en la distribución ji-cuadrado con $5 * 3 = 15$ grados de libertad un p-valor de 0,9.

	NO ANS.	ANS	NO DEPRE	DEPRE
EC1	382	118	380	120
EC2	17	4	17	4
EC3	69	19	65	23
EC4	23	6	21	8
EC5	15	1	15	1
EC6	132	34	118	48

Cuadro 7.6: Tabla de contingencia para el cruce del estado civil y la ansiedad y depresión

7.2. AC de la ansiedad y depresión

Ahora vamos a realizar un cruce más interesante, con el que queremos estudiar la relación entre las notas de ansiedad y la depresión. No consideramos ahora el tener o no ansiedad y depresión, consideramos las propias notas medidas de 0-9. Debido a las bajas frecuencias en los altos niveles de ansiedad y depresión agrupamos las notas 8 y 9 para ambas enfermedades. Seguimos teniendo una inercia alta con estas agrupaciones y el estadístico ji-cuadrado deja un p-valor del orden de 0 para el contraste de la dependencia de las variables tras la agrupación, por lo que no habrá ningún problema al hacer estas agrupaciones y resolvemos el problema de las bajas frecuencias. La tabla de contingencia para este cruce se recoge en el Cuadro 7.7.

	A0	A1	A2	A3	A4	A5	A6	A7	A89
D0	423	83	12	12	18	23	10	5	1
D1	14	8	1	2	1	1	2	0	0
D2	14	2	0	3	3	5	3	1	1
D3	15	6	1	2	5	2	4	5	3
D4	8	6	0	2	3	3	8	6	3
D5	4	7	0	4	2	4	3	5	3
D6	3	1	0	0	0	7	2	6	4
D7	4	0	0	0	2	2	2	7	7
D89	0	1	0	0	0	1	1	2	6

Cuadro 7.7: Tabla de contingencia para el cruce de las notas de ansiedad y depresión

Antes de comenzar con el AC debemos comprobar si la tabla presenta forma homogénea o no. De ser homogénea no serviría de nada realizar el AC y construir lo mapas pues

ya la propia tabla nos indica la no relación entre unas variables y otras. Para la tabla mostrada en el Cuadro 7.7, $\chi^2 = 467{,}40$, que en la distribución ji-cuadrado con $8 * 8 = 64$ grados de libertad deja un p-valor del orden de 0. Tenemos pruebas significativas por tanto de la no homogeneidad de la tabla, por lo que existirán notas de ansiedad que tengan asociadas distintas notas de depresión, es decir, algunas notas de ansiedad y depresión estarán relacionadas.

Una vez comprobamos que la tabla no es homogénea, construimos el mapa del AC. En este caso vamos a fijarnos en el mapa simétrico (ver Figura 7.2), pues lo que vamos a interpretar son la distancia entre los perfiles de filas y columnas y no precisaremos los vértices como referencia. El mapa asimétrico (ver Figura 7.1) proporciona la misma interpretación de los datos, pero al considerar tantas variables se aprecian menos las relaciones y es más difícil el estudio de este mapa. Mostramos ambos mapas como ejemplo para reflejar la similitud entre ambos mapas que ya comentamos en el capítulo 4.

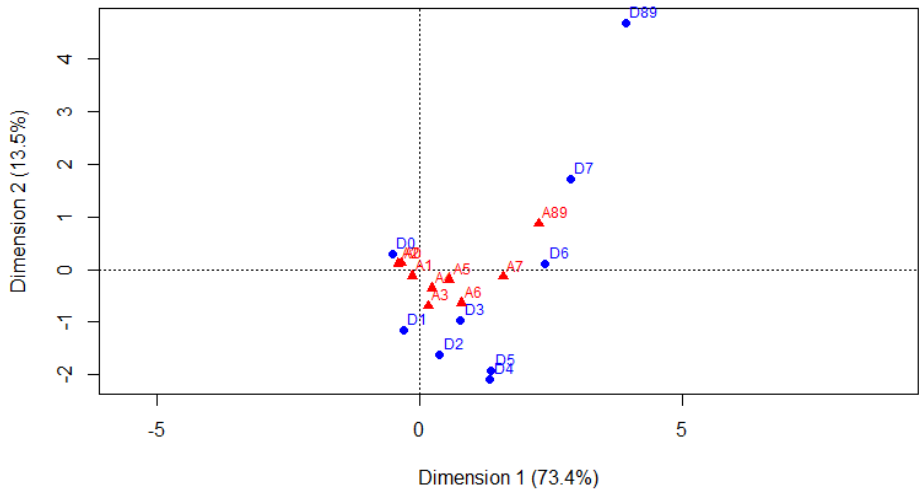

Figura 7.1: Mapa asimétrico del AC asociado a la tabla del Cuadro 7.7

Los mapas muestran una clara estructura. En primer lugar, en el mapa asimétrico en el eje horizontal (primer eje principal) se produce una ordenación de menor a mayor en las notas en depresión de izquierda a derecha. Por tanto un perfil fila de ansiedad situado a la derecha indicará una mayor nota en depresión. Los propios perfiles fila de la ansiedad

también se ordenan en el eje horizontal de menor a mayor, por lo que concluimos que existe una relación directa entre las notas de depresión y ansiedad de forma que pacientes con altos niveles de ansiedad tienden a tener mayores niveles de depresión. En el eje vertical (segundo eje principal) se observa como los vértices asociados al 8 y 9 en depresión se alejan del resto. Podemos interpretar esto indicando que los altos niveles de depresión no siguen una estructura similar a los demás, lo que probablemente se deba a que niveles tan altos de depresión no se expliquen solo respecto a la ansiedad.

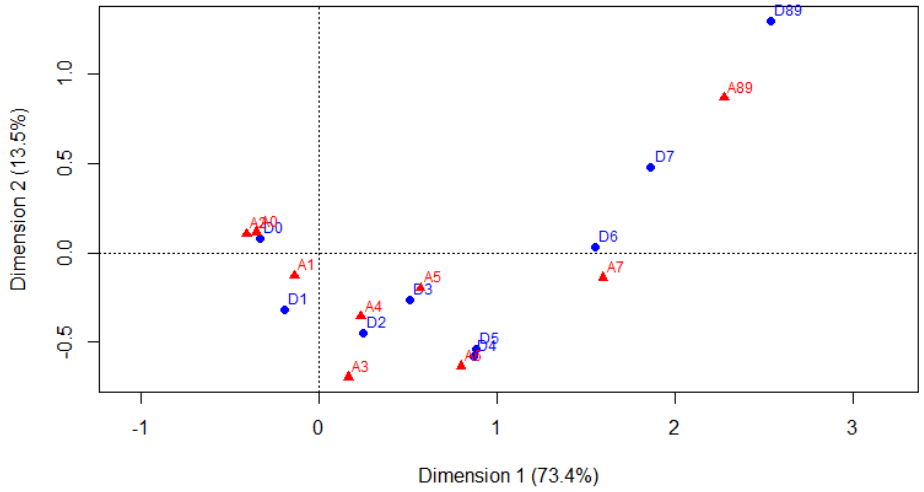

Figura 7.2: Mapa simétrico del AC asociado a la tabla del Cuadro 7.7

En el mapa simétrico de la Figura 7.2 observamos lo mismo que hemos comentado antes pero más claramente. No tenemos los vértices como referencia pero vemos como los perfiles de ansiedad y depresión tienden a estar juntos según sus notas, estando más cerca los perfiles de ansiedad y depresión con la misma nota. Al igual que antes, los niveles altos de depresión se desmarcan del resto de donde se deduce que debe existir algún otro factor que conlleve un nivel de depresión extremo. En cada eje se muestra el porcentaje de inercia que explica cada uno como ya explicamos en el capítulo 4. El primer eje principal del mapa asimétrico explica el 73,4 % de la inercia total de la tabla del Cuadro 7.7 y el segundo eje principal el 13,5 %. La suma es el 86,9 %, que es el porcentaje explicado por el mapa. El mapa simétrico explica un 86,9 % también. Tenemos un gran porcentaje de inercia explicado, por lo que las conclusiones obtenidas son fiables y los perfiles están bien representados en el mapa respecto a su posición real (en un espacio de 9 dimensiones).

Una vez que hemos visto que existe una relación directa entre los niveles de ansiedad y depresión vamos a contrastar qué relación existe entre la ansiedad y la depresión y la resiliencia. En este caso cruzamos el hecho de tener o no ansiedad y depresión y los niveles bajo-medio-alto de resiliencia. El cruce de estas dos variables se muestra en la tabla del Cuadro 7.8.

	NO ANS.	ANS	NO DEPRE	DEPRE
resi.baja	74	39	57	56
resi.media	459	131	451	139
resi.alta	105	12	108	9

Cuadro 7.8: Tabla de contingencia para el cruce de ansiedad y depresión con niveles de resiliencia

De nuevo, lo primero que hacemos es contrastar la hipótesis nula de homogeneidad de la tabla. Para la tabla de contingencia del Cuadro 7.8 el estadístico de contraste vale $\chi^2 = 75{,}45$, que deja un p-valor de $3{,}09e{-}14$ en la distribución ji-cuadrado con $3{*}2 = 6$ grados de libertad, por lo que tenemos pruebas significativas de que varían las frecuencias de ansiedad y depresión respecto al nivel de resiliencia del individuo. Para la representación, notemos que todos los perfiles se pueden representar perfectamente en un espacio de dimensión 2, por lo que para este cruce representaremos el mapa simétrico (Figura 7.3), donde se representan los perfiles sin perdida de información.

La interpretación en el mapa de la Figura 7.3 es muy clara. En el primer eje principal se sitúan a un lado del origen (que corresponde con el perfil medio como ya se comentó en el capítulo 4) los perfiles de no tener depresión, no tener ansiedad y los perfiles de resiliencia alta y media. A la derecha del origen se sitúan los perfiles asociados a la depresión, ansiedad y resiliencia baja. El primer eje principal explica el 99,3 % de la inercia, por lo que considerando solo la interpretación de este eje ya obtendremos conclusiones fiables. Se ve muy claramente como los pacientes con una resiliencia alta o media tienden a no padecer de depresión ni ansiedad y como los pacientes con baja resiliencia padecen de estas enfermedades. Por tanto asociamos un nivel bajo de resiliencia como un factor de riesgo ante la posibilidad de padecer ansiedad o depresión (las cuales como ya vimos están muy relacionadas también). Puede deberse a que las personas con baja resiliencia tienden a afrontar peor los problemas, viéndose más afectados por ellos (lo que las deprime) y viéndose superados por ellos (lo que les provoca ansiedad).

Resumiendo, un bajo nivel de resiliencia supone un factor de riesgo de cara a poder padecer ansiedad y depresión debido a la relación que existe entre las variables, relación que se ve patente en el mapa del AC de la Figura 7.3.

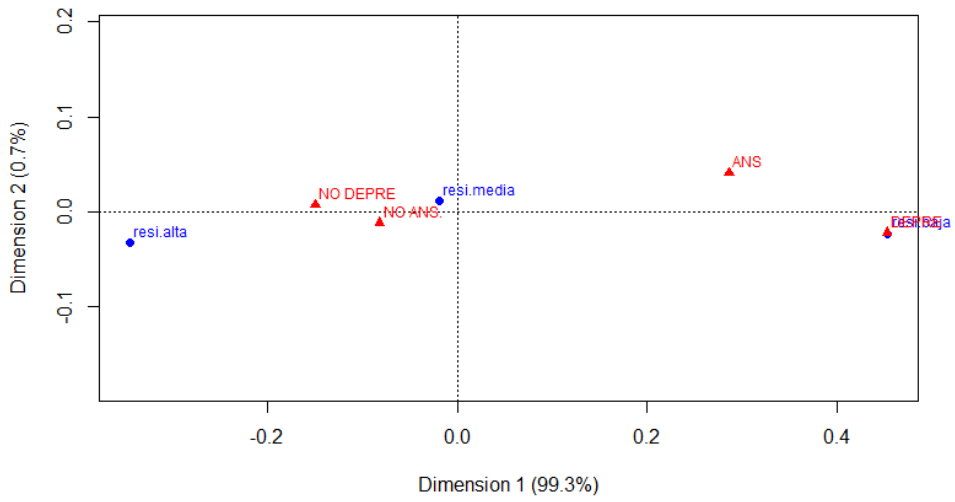

Figura 7.3: Mapa simétrico del AC asociado a la tabla del Cuadro 7.8

7.3. AC de la edad sobre el resto de variables

Ahora vamos a estudiar el efecto de la edad sobre las demás variables de interés de nuestro estudio. Comenzamos dividiendo la escala continua de edad (que va de 18 a 90 años) en una escala con tres niveles. Realizamos la división igual que con la resiliencia, tomando como pacientes en edad joven a aquellos que se encuentran entre la media más-menos la desviación típica de la variable edad. Así, dividimos el intervalo $[18, 90]$ en $[18, 48'61 - 16'95]$, $[48'61-16'95, 48'61+16'95]$, $[48'61+16'95, 90] = [18, 31'65]$, $[31'65, 65'56]$, $[65'56, 90]$, donde el primer intervalo corresponde a los jóvenes, el segundo a la edad media y el tercero a la edad anciana. Construimos en base a estos nuevos intervalos la matriz binaria (representada en el Cuadro 7.9) que codifica a qué grupo de edad pertenece cada paciente y realizamos el AC con esta matriz.

	Recuento
Joven	152
Edad media	506
Anciano	162

Cuadro 7.9: Tabla con los recuentos para cada grupo de edad

Con la edad ya codificada, construimos la tabla de contingencia que cruza la edad con el nivel de resiliencia (Cuadro 7.10). Sobre esta tabla calculamos el estadístico ji-cuadrado, $\chi^2 = 4{,}06$ y calculamos la probabilidad que deja en la distribución ji-cuadrado con $2 * 2 = 4$ grados de libertad, lo que nos da el p-valor 0,4, por lo que aceptamos la hipótesis de homogeneidad de la tabla y concluimos que los tres grupos de edad tienen unos niveles de resiliencia homogéneos.

	resi.baja	resi.media	resi.alta
Joven	14	114	24
Edad.media	72	362	72
Anciano	27	114	21

Cuadro 7.10: Tabla de contingencia para el cruce de la edad con el nivel de resiliencia

El cruce de la edad con la ansiedad y la depresión produce la tabla del Cuadro 7.11. Para este cruce, el estadístico ji-cuadrado es $\chi^2 = 12$, que deja un p-valor en la distribución ji-cuadrado con $3 * 2 = 6$ grados de libertad de 0,06. Rechazamos a un nivel del 5 % la hipótesis nula de homogeneidad, además al tratarse de una tabla de tan baja dimensión no merece la pena realizar el AC.

	NO ANS.	ANS	NO DEPRE	DEPRE
Joven	115	37	123	29
Edad media	391	115	386	120
Anciano	132	30	107	55

Cuadro 7.11: Tabla de contingencia para el cruce de la edad con tener o no ansiedad o depresión

El último cruce que haremos con la edad podría incluirse en la siguiente sección.

Vamos a cruzar los tres grupos de edad con las enfermedades que hemos agrupado anteriormente. Este cruce produce la tabla de contingencia del Cuadro 7.12. Para esta tabla rechazamos la hipótesis de homogeneidad al dejar el estadístico $\chi^2 = 117{,}66$ un p-valor de $1{,}11e-16$, lo que supone una prueba significativa. Por tanto, existirán diferencias en los distintos grupo de edad respecto a las enfermedades padecidas. En la Figura 7.4 construimos el mapa asimétrico, representando las columnas en coordenadas principales y las filas como vértices para poder tomar los tres grupos de edad como referencia de cara a realizar conclusiones.

	Cardio	Arterial	Pulmon	Cuerpo	Piel	Hormona	Ins.Ren	Hepat.	Depre	Cancer
Joven	2	4	11	1	11	9	0	4	10	1
Edad media	103	117	32	12	23	65	1	26	73	11
Anciano	115	89	17	17	6	56	12	9	30	19

Cuadro 7.12: Tabla de contingencia para el cruce de la edad con las enfermedades

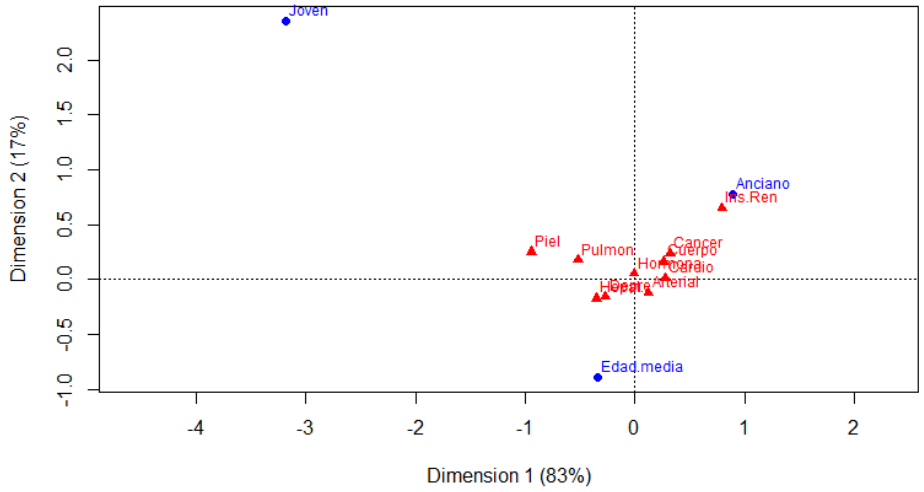

Figura 7.4: Mapa asimétrico del AC asociado a la tabla del Cuadro 7.12

Del mapa de la Figura 7.4 comenzamos destacando que los perfiles se encuentran representados perfectamente al considerar una tabla de dimensión 2, por tanto las conclusiones que hagamos serán fiables. Vemos una clara ordenación en el eje horizontal, situándose de izquierda a derecha los tres grupos de edad ordenados de menor a mayor. Como para los grupos de edad hemos representado los vértices, interpretaremos que una enfermedad

situada a la derecha es más habitual en ancianos que otras que se sitúe más a la izquierda. En este eje tenemos explicado el 83 % de la inercia, por lo que bastará estudiar este eje. Concluimos entonces que las enfermedades cardíacas, arteriales, relacionadas con el cuerpo (reuma y osteoporosis), la diabetes, el hipertiroidismo, la insuficiencia renal y el cáncer se asocian con un grupo de edad anciano, y la hepatopatía, las enfermedades pulmonares y las relacionadas con la piel con los jóvenes. Por tanto, por ejemplo, la edad es un factor de riesgo ante la posibilidad de padecer insuficiencia renal. El hecho de que el vértice de los jóvenes se encuentre tan alejado del de la edad media y anciana indica que en general la edad es un factor de riesgo ante cualquier enfermedad, pues todos los perfiles de las enfermedades se encuentran más próximos a los vértices de las edades mayores.

7.4. AC sobre las enfermedades

El primer contraste que realizamos en esta sección es el que hacemos sobre la tabla que cruza las enfermedades y los fármacos. A priori esperamos que se asocien los fármacos con las enfermedades para los que se recetan, pero hasta comprobarlo estadísticamente no lo podremos asegurar. La tabla de contingencia para este cruce se muestra en el Cuadro 7.13. Esta tabla tiene una inercia de 0,14 y $n = 2742$, por lo que el estadístico de contraste será $\chi^2 = 0{,}14 * 2742 = 383{,}88$. En la distribución ji-cuadrado con $9 * 12 = 108$ grados de libertad, este estadístico deja un p-valor del orden de 0, por lo que rechazamos la hipótesis de homogeneidad de la tabla. Existirán entonces diferencias entre las enfermedades respecto a la medicación que toman los pacientes que las padecen.

	Benzo	Antidepre	MetGluc	ado1	Antiinf	Insul2	hipSed	Estatinas	cortic	betabloq	diureticos	Glucosamin	antipsic
Cardio	49	46	324	48	32	9	40	101	3	37	77	3	2
Arterial	42	34	292	41	27	11	35	124	2	25	46	3	5
Pulmon	13	10	34	2	3	0	11	10	2	2	7	0	0
Cuerpo	12	9	23	3	5	1	7	6	2	2	6	1	0
Piel	7	5	15	1	2	0	5	4	1	2	2	0	0
Hormona	22	32	224	64	15	11	28	53	0	15	29	1	2
Ins.Ren	4	1	39	6	0	3	1	8	0	3	9	0	0
Hepat.	10	7	23	4	6	0	6	10	1	1	4	0	0
Depre	53	76	86	13	17	3	46	21	2	8	15	2	3
Cancer	4	8	36	8	4	1	5	9	0	2	4	1	0

Cuadro 7.13: Tabla de contingencia para el cruce de las enfermedades y fármacos

Como estamos ante una tabla no homogénea, representamos el mapa del AC para esta tabla y estudiamos el resultado para encontrar cuales son las asociaciones entre variables

que rompen la homogeneidad. Este mapa se presenta en la Figura 7.5. Mostramos el mapa simétrico ya que debido al alto número de variables consideradas en esta tabla, en un mapa asimétrico se dificultaría la interpretación e incluso distinguir entre las etiquetas de los perfiles. En el mapa de la Figura 7.5 observamos como se cumple lo que parecía razonable desde un principio. Los antidiabéticos y la insulina se hallan cerca de las enfermedades de diabetes e hipertiroidismo, los antidepresivos cerca de los antidepresivos, los medicamentos para el corazón (estatinas, betabloqueantes) cerca de las dolencias cardíacas y las enfermedades del cuerpo, pulmones y piel cerca de los antiinflamatorios y con una tendencia hacia los corticoides. Esta cercanía como ya comentamos a lo largo del trabajo se interpreta como asociación entre los perfiles, y al tener un 84,6 % de inercia explicada en el mapa, tomamos las conclusiones anteriores con seguridad.

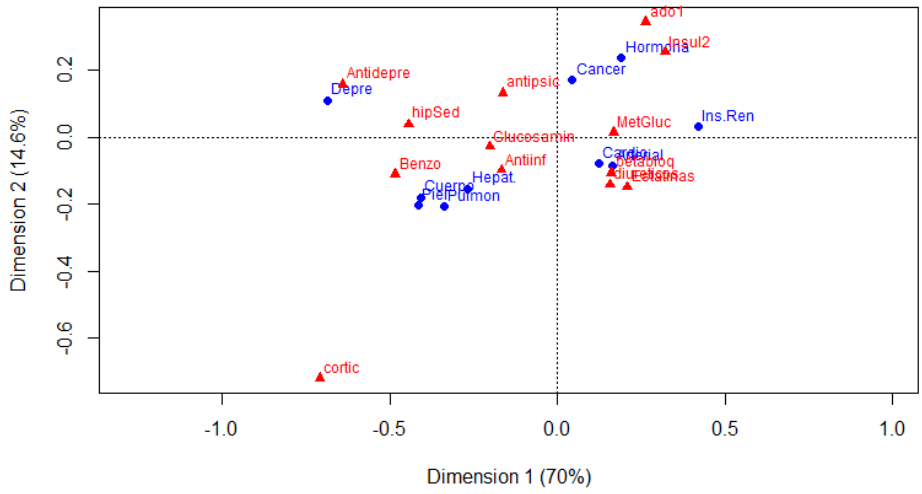

Figura 7.5: Mapa simétrico del AC asociado a la tabla del Cuadro 7.13

Los siguientes dos cruces constituyen la parte más interesante desde el punto de vista médico. La resiliencia es una medida que todavía no está muy estudiada y establecer la relación de ésta con alguna enfermedad sería un descubrimiento. Sería un avance para poder explicar la resiliencia mejor, su mejor entendimiento y poder estudiar con mayor profundidad su comportamiento. Lo mismo ocurre con la ansiedad y la depresión, que a pesar de ser enfermedades más estudiadas tampoco existen muchos trabajos probando su relación con otras enfermedades. El investigador insistió en la relevancia de estos dos cruces y de la importancia que pueden tener en futuras investigaciones y estudios.

Mostramos en el Cuadro 7.14 el cruce de las enfermedades con el padecer o no ansiedad y depresión. El estadístico de contraste para esta tabla es $\chi^2 = 76{,}16$. En la distribución ji-cuadrado con $3 * 9 = 27$ grados de libertad este estadístico deja un p-valor de $1{,}42e-6$, por lo que tenemos pruebas significativas de que la tabla no es homogénea. Esto nos indica que existirá relación entre padecer o no ansiedad y depresión y ciertas enfermedades.

	NO ANS.	ANS	NO DEPRE	DEPRE
Cardio	179	41	155	65
Arterial	162	48	143	67
Pulmon	49	11	44	16
Cuerpo	18	12	15	15
Piel	29	11	29	11
Hormona	105	25	86	44
Ins.Ren	11	2	9	4
Hepat.	29	10	27	12
Depre	63	50	48	65
Cancer	27	4	24	7

Cuadro 7.14: Tabla de contingencia para los cruces de enfermedades y ansiedad y depresión

En la Figura 7.6 mostramos el mapa asimétrico del AC para el cruce de las enfermedades y la ansiedad y depresión y en la Figura 7.7 el simétrico para poder observar mejor los perfiles. Observamos que el mapa simétrico se divide en la primera dimensión en dos "lados", uno asociado a no padecer ansiedad o depresión y otro con padecerlas. Este eje explica el 97,1 % de la inercia de la tabla, por lo que bastaría considerar este eje. Como nota, vemos que en la segunda dimensión se produce la distinción entre ansiedad y depresión una vez se está en el lado izquierdo o derecho del mapa. Este segundo eje no presenta casi dispersión ni mucha varianza explicada debido a la relación existente entre ansiedad y depresión que ya demostramos antes. Por tanto, solo vamos a interpretar el primer eje principal. La alta inercia que explica este eje hace que no tengamos problemas obviando el otro, y más una vez demostrada la relación entre ansiedad y depresión.

Concluimos que la depresión y las enfermedades que afectan al cuerpo como la reuma y la osteoporosis son enfermedades asociadas a padecer ansiedad y depresión. Parece lógico para la depresión, pero para la variable cuerpo probablemente sea consecuencia de que el dolor constante que provocan estas dolencias cause depresión y ansiedad en los pacientes.

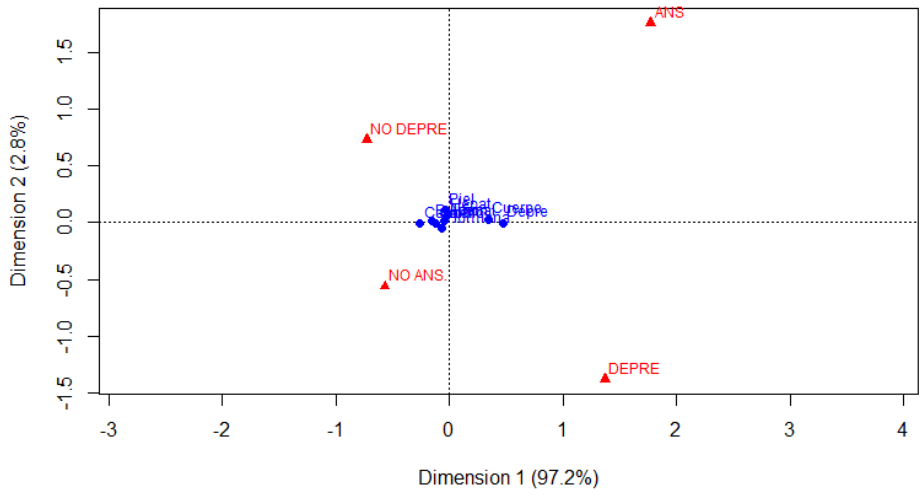

Figura 7.6: Mapa asimétrico del AC asociado a la tabla del Cuadro 7.14

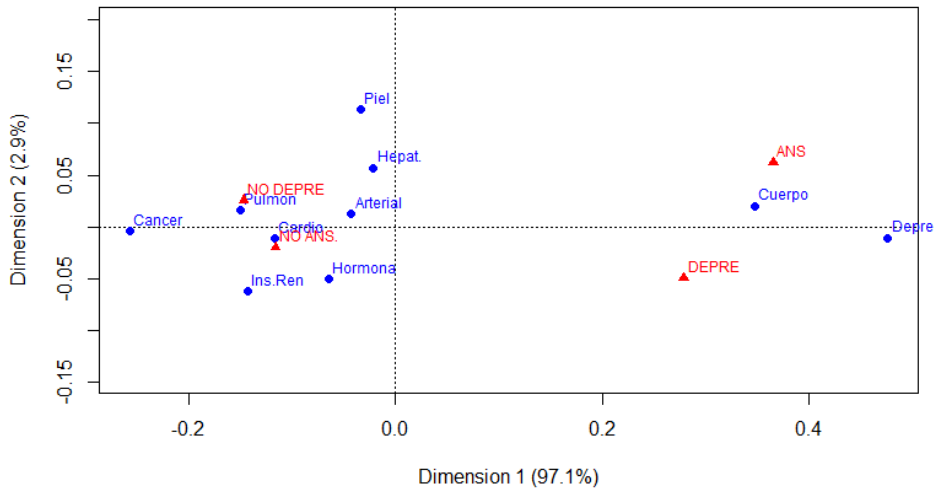

Figura 7.7: Mapa simétrico del AC asociado a la tabla del Cuadro 7.14

El último cruce que consideraremos en este estudio es el de las enfermedades con los niveles de resiliencia. Recalcamos que es el más interesante y novedoso. En el Cuadro 7.15 mostramos la tabla de contingencia para este cruce. El valor del estadístico para el contraste de la hipótesis de homogeneidad de esta tabla es $\chi^2 = 29{,}75$. Este valor deja en la distribución ji-cuadrado con $9 * 2$ grados de libertad un p-valor de 0,04, por lo que para los niveles de significación habituales rechazamos la hipótesis de homogeneidad de la tabla. Esto ya de por sí es relevante al comprobar estadísticamente que las enfermedades afectan a los niveles de resiliencia de una persona y viceversa.

	resi.baja	resi.media	resi.alta
Cardio	39	157	24
Arterial	45	141	24
Pulmon	11	34	15
Cuerpo	8	20	2
Piel	6	26	8
Hormona	30	87	13
Ins.Ren	2	9	2
Hepat.	7	31	1
Depre	31	74	8
Cancer	2	24	5

Cuadro 7.15: Tabla de contingencia para el cruce de las enfermedades y resiliencia

Una vez comprobada la no homogeneidad de la tabla del Cuadro 7.15 representamos el mapa asimétrico (Figura 7.8) y simétrico (Figura 7.9) para este cruce. De nuevo, el alto número de variables y la baja inercia hacen que nos sea difícil interpretar los perfiles, de ahí que incluyamos el mapa simétrico también para estudiar las asociaciones entre las variables.

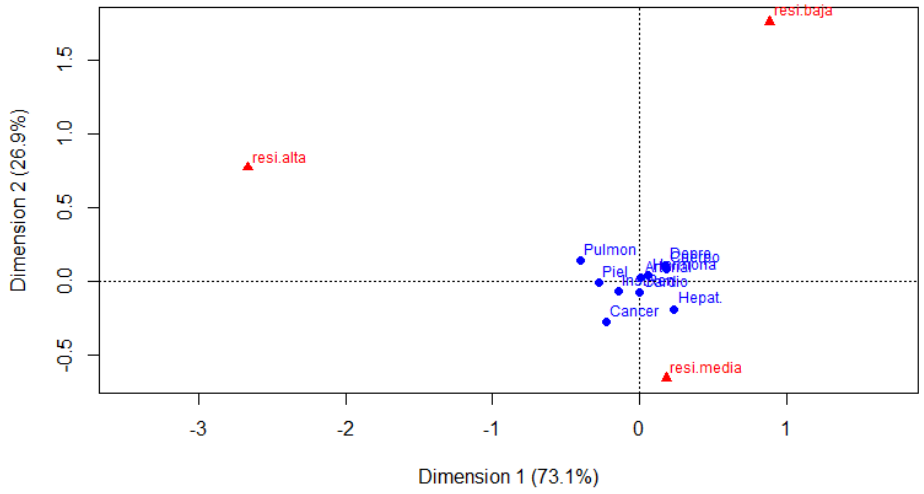

Figura 7.8: Mapa asimétrico del AC asociado a la tabla del Cuadro 7.15

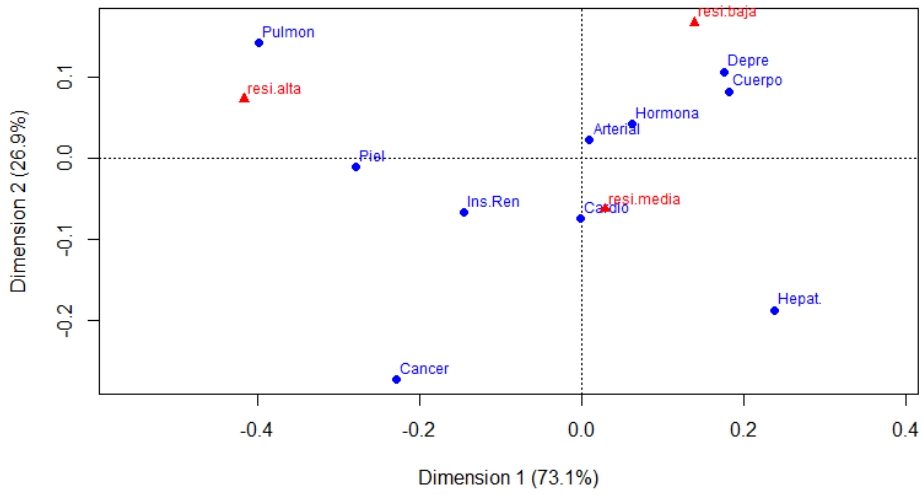

Figura 7.9: Mapa simétrico del AC asociado a la tabla del Cuadro 7.15

Estos mapas representan el 100 % de la inercia, por lo que los perfiles están perfectamente representados y podremos establecer conclusiones fiables. Vemos como en el primer eje principal se ordenan los niveles de resiliencia de izquierda a derecha de mayor a menor.

Podemos tomar el origen como eje divisor, encontrando a la derecha del mapa el nivel alto de resiliencia y a la izquierda los niveles medio y bajo. El eje horizontal explica el 73,1 % de la inercia, por lo que bastará con considerar este eje. Así, una enfermedad cuyo perfil se encuentre a la izquierda se asociará con un nivel alto de resiliencia, es decir, los pacientes con esta enfermedad tienden a tener alta resiliencia. Análogamente con el lado derecho.

Concluimos entonces que las enfermedades del pulmón, piel y el cáncer y la insuficiencia renal son enfermedades en las que los pacientes que los padecen tienden a tener altos niveles de resiliencia. Para el caso del cáncer puede deberse al hecho de que al enfrentarse a una enfermedad tan dura el paciente no tiene otra alternativa más que ser fuerte, es decir tener una mayor resiliencia. Igualmente con las enfermedades pulmonares y la insuficiencia renal, las cuales suelen ser persistentes y difíciles de sobrellevar. Comprobamos así una de las hipótesis de la resiliencia que aún no han sido testadas, que es que aumenta cuando se pone a una persona ante una enfermedad grave, en particular cuando el paciente se enfrenta al cáncer o a una enfermedad pulmonar principalmente. Llama más la atención las enfermedades de la piel y su asociación con la alta resiliencia, para la cual tendrá que indagar más nuestro investigador. Respecto a los niveles bajos de resiliencia, se asocian a estos la depresión y las enfermedades del cuerpo.

7.5. Perímetros convexos

En el capítulo 6 introducimos el concepto de perímetros convexos en la sección del bootstraping. Recordamos, consiste en construir nuevas tablas simuladas a partir de las frecuencias de la tabla original y representar los perfiles de todas las tablas que hayamos construido en el mismo mapa. Una vez representados los perfiles de todas las tablas, representamos **perímetros convexos** que contienen todos los perfiles de cada fila. En esta sección realizaremos las simulaciones y la construcción de los perímetros convexos con R. Recortaremos los perímetros considerando el 95 % de los perfiles más internos de cada perímetro convexo para así crear regiones con este nivel de confianza.

Vamos a realizar para cada contraste 500 simulaciones y representar los perfiles de todas ellas. Ya explicamos en el tema 6 la interpretación de estas construcciones. Si dos perímetros no se cortan tenemos seguridad de que los perfiles asociados a cada perímetro son significativamente distintos respecto a las variables consideradas.

Consideramos tres cruces para la construcción de los perímetros convexos. El primero es el cruce de las enfermedades con la edad. La tabla a partir de la cual construimos las 500 simulaciones es la del Cuadro 7.12. Una vez representados los perfiles de las simulaciones, trazamos una linea delimitando el 95 % de los perfiles más interiores y obtenemos la Figura 7.10.

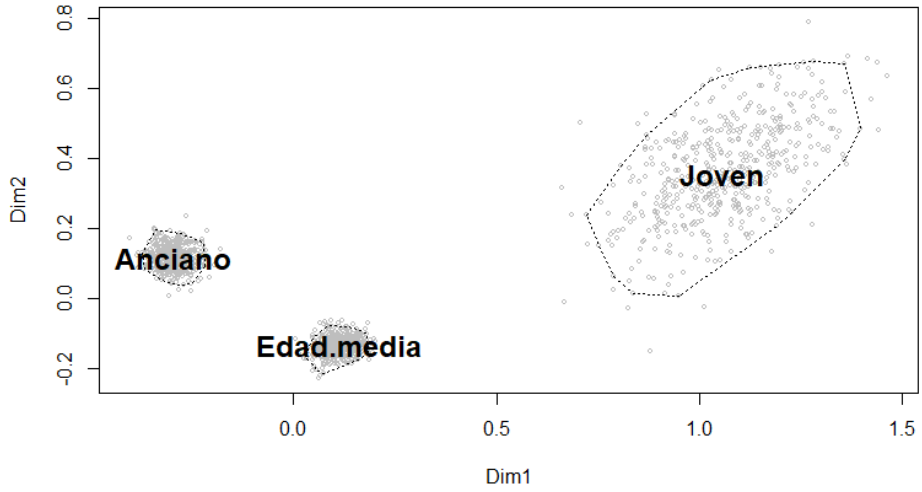

Figura 7.10: Perímetros convexos para el cruce de las enfermedades y la edad

Vemos muy claramente como los perfiles asociados a cada grupo de edad son sustancialmente distintos respecto a las enfermedades dos a dos. Es decir, existen diferencias significativas entre las frecuencias de las enfermedades consideradas entre el grupo de edad joven y anciano, entre el joven y la edad media y entre la edad media y la anciana. Esto concuerda con el AC que realizamos en el mapa de la Figura 7.4, donde además vimos qué enfermedades provocan estas diferencias entre los grupos de edad.

El otro cruce para el que construimos los perímetros convexos de las enfermedades y el padecer o no ansiedad y depresión. A partir de la tabla del Cuadro 7.14 construimos las simulaciones y las representamos como perfiles en el mapa. Mostramos los perímetros asociados a este cruce en la Figura 7.11.

La interpretación para la Figura 7.11 es la misma que ya concluimos antes a partir del

mapa del AC para este cruce (Figura 7.6 y Figura 7.7). No podemos establecer la relación entre ansiedad y depresión por el solapamiento de sus perímetros convexos pues podrían no existir en otra dimensión no considerada. Lo que podemos hacer es concluir que tenemos pruebas significativas de las diferencias de los perfiles asociados con padecer ansiedad y depresión y los que no. Así, respecto a padecer ansiedad o depresión, las frecuencias de las enfermedades consideradas son diferentes, por lo que dependiendo de si el paciente tiene ansiedad y depresión o no tenderá a tener unas enfermedades u otras.

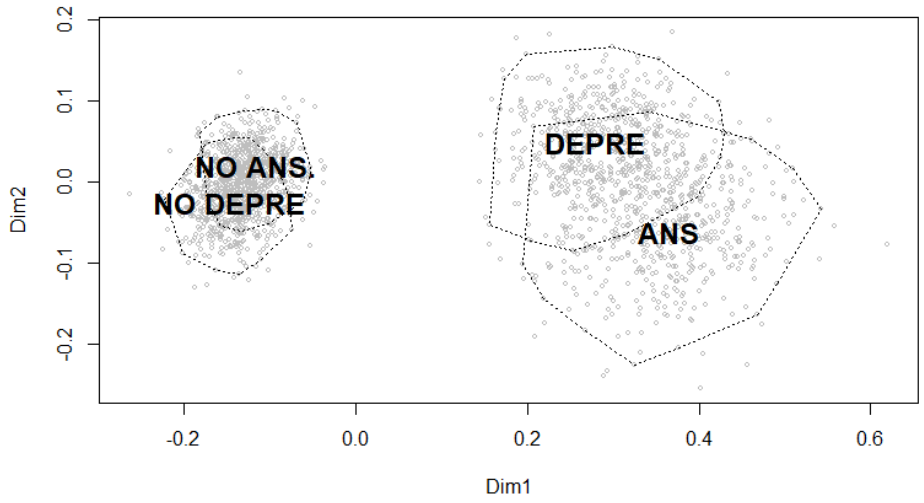

Figura 7.11: Perímetros convexos para el cruce de las enfermedades y la ansiedad y depresión

De nuevo el último cruce es el más interesante desde el punto de vista médico. Cruzamos ahora las enfermedades con los niveles de resiliencia y construimos las simulaciones a partir de la tabla del Cuadro 7.15. Para este cruce obtenemos los perímetros convexos de la Figura 7.12. Volvemos a observar una clara diferenciación en las frecuencias (perfiles) que tienen las enfermedades consideradas respecto al nivel de resiliencia de los pacientes. Por tanto, un nivel de resiliencia u otro serán determinantes a la hora de padecer una enfermedad u otra. A partir del AC realizado anteriormente representado en los mapas de la Figura 7.8 y 7.9 comprobamos qué enfermedades se asocian con cada nivel de resiliencia.

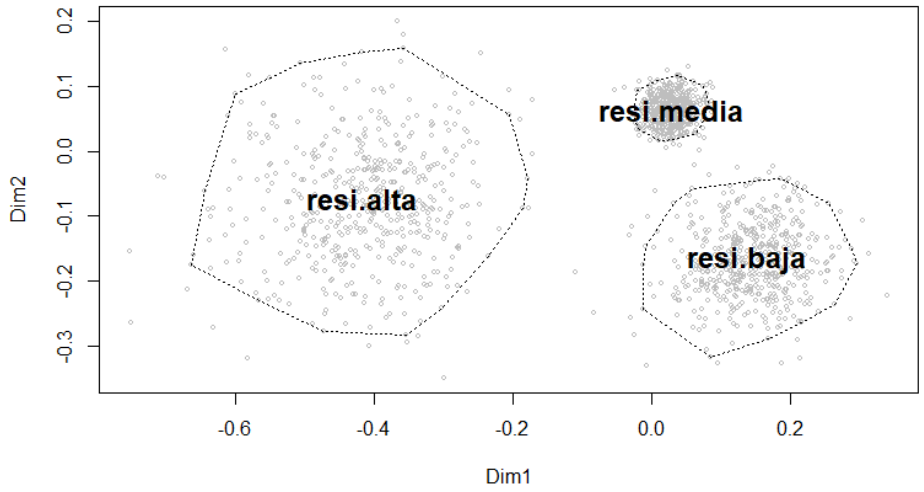

Figura 7.12: Perímetros convexos para el cruce de las enfermedades y los niveles de resiliencia

Glosario

Recogemos todos los conceptos y notaciones tratados a lo largo del trabajo por orden de introducción:

- **tabla de contingencia**: Tabla que recoge las frecuencias o recuentos de los cruces de las variables consideradas.
- n_{ij}: Frecuencia observada para la celda (i,j) de la tabla de contingencia.
- $n_{i\bullet} = \sum_{j=1}^{m} n_{ij}$, marginal de la fila i-ésima. (2.1)
- $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$, marginal de la columna j-ésima. (2.2)
- $n_{\bullet\bullet} = n = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$, total de frecuencias de la tabla de contingencia. (2.3)
- **perfil fila**: $a_i = (n_{i1}, n_{i2}, ..., n_{im})^T / n_{i\bullet} = (n_{i1}/n_{i\bullet}, n_{i2}/n_{i\bullet}, \cdots, n_{im}/n_{i\bullet})^T$, vector de las frecuencias de una fila divididas por su total. (2.4)
- **vértice**: p.e. $(1, 0, ..., 0)^T$, perfil extremo que concentra toda la frecuencia en una sola coordenada o componenete.
- **masa**: $r_i = \frac{n_{i\bullet}}{n_{\bullet\bullet}}$, marginal de la fila dividido por el total de la tabla.
- **perfil fila medio**: $c = \left(\frac{\sum_{i=1}^{k} n_{i1}}{n_{\bullet\bullet}}, \frac{\sum_{i=1}^{k} n_{i2}}{n_{\bullet\bullet}}, \cdots, \frac{\sum_{i=1}^{k} n_{im}}{n_{\bullet\bullet}} \right)^T$ (2.5)
- **Distancia** χ^2: $\sqrt{\frac{(\frac{n_{i1}}{n_{i\bullet}} - \frac{n_{j1}}{n_{j\bullet}})^2}{c_1} + \frac{(\frac{n_{i2}}{n_{i\bullet}} - \frac{n_{j2}}{n_{j\bullet}})^2}{c_2} + \cdots + \frac{(\frac{n_{im}}{n_{i\bullet}} - \frac{n_{jm}}{n_{j\bullet}})^2}{c_m}}$ (3.1)
- $e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n_{\bullet\bullet}}$, valor esperado celda (i,j). (3.2)
- $\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$, estadístico ji-cuadrado.

- **Inercia**: $\chi^2/n = \sum_i r_i \cdot \|a_i - c\|_c^2 = \sum_i r_i \sum_j \left(\frac{p_{ij}}{r_i} - c_j\right)^2 / c_j$ (3.5)
- $\|a_i - c\|_c = \sqrt{\sum_j (a_{ij} - c_j)^2/c_j}$, distancia χ^2 entre el i-ésimo perfil fila a_i y el perfil fila medio c.
- **DVS**: Descomposición en valores singulares.
- **Escalado óptimo**: Escala de las categorías que maximiza la inercia de la tabla.
- **Eje pricipal**: Eje del mapa del AC.
- **Inercia principal**: Inercia explicada por un eje principal.
- **Coordenada principal**: Coordenada de un perfil en un eje principal.
- **Coordenada estándar**: Coordenada de un vértice en un eje principal.
- **Mapa asimétrico**: Mapa que representa las filas en coordenadas principales y las columnas en coordenadas estándares o viceversa.
- **Mapa simétrico**: Mapa que representa las filas y las columnas en coordenadas principales .
- **Inercia de filas**: $r_i \cdot \|a_i - c\|_c^2$, contribución de cada fila a la inercia.
- $\left(\frac{f_{ik}}{d_i}\right)^2$, contribución del eje k a la inercia del i-ésimo perfil.
- **Biplot**: representación del AC para cada perfil fila y columna conjuntamente a partir del producto escalar de dos vectores.
- **Fórmula de reconstitución**: $p_{ij} = r_i c_j \left(1 + \sum_{k=1}^{K} \sqrt{\lambda_k}\phi_{ik}\gamma_{jk}\right)$ (4,14)
- **Biplot estándar**: Biplot de los perfiles escalados.
- **Doblado de los datos**: Expresión de los datos en relación a su diferencia con el valor más bajo y más alto de la escala.
- $f_{ik} = \sum_j \left(\frac{p_{ij}}{r_i}\right) \gamma_{jk}$, k-ésima coordenada principal de la fila i. (4.17)
- γ_{jk}: k-ésima coordenada estándar de la j-ésima columna.
- $g_{ik} = \sum_i \left(\frac{p_{ij}}{c_j}\right) \phi_{ik}$, k-ésima coordenada principal de la columna j. (4.18)
- ϕ_{ik}: k-ésima coordenada estándar de la i-ésima fila.
- **ACM**: Análisis de correspondencias múltiple.

- **Matriz binaria** (Z): Matriz con un 1 en la componente (i,j) si el individuo i pertenece a la categoría j y 0 si no.
- **Matriz de Burt**: $B = Z^T Z$
- **ACCo**: Análisis de correspondencias conjunto.
- **ACC**: Análisis de correspondencias canónico.
- **Bootstrapping**: Técnica que considera los datos que tenemos como si fuesen la población total para crear nuevas muestras tomadas de esta.
- **Inercia intergrupos**: Medida de la variabilidad de la tabla una vez agrupados los datos originales de alguna forma.
- **Inercia intragrupos**: Diferencia entre la inercia total y la intergrupos.

Bibliografía

La práctica del análisis de correspondencias, Michael Greenacre, Barcelona, 2008.

Análisis de correspondencias simples y múltiples, Santiago de la Fuente Fernández, Universidad Autónoma de Madrid (UAM), 2011.

The biplot graphic display of matrices with application to principal component analisis, K.R. Grabiel, 1971.

Análisis de Correspondencias, Salvador Figueras, M., 2003

An introduction to the bootstrap. Efron, B. & Tibshirani, R. J., 1993.

Análisis de correspondencias. Luis Joaristi Olariaga & Luis Lizasoain Hernández, 2000.