

Capítulo 7

ANÁLISIS DE CORRESPONDENCIAS

7.1 INTRODUCCIÓN

El análisis de correspondencias es una técnica descriptiva para representar tablas de contingencia, es decir, tablas donde recogemos las frecuencias de aparición de dos o más variables cualitativas en un conjunto de elementos. Constituye el equivalente de componentes principales y coordenadas principales para variables cualitativas. La información de partida ahora es una matriz de dimensiones $I \times J$, que representa las frecuencias absolutas observadas de dos variables cualitativas en n elementos. La primera variable se representa por filas, y suponemos que toma I valores posibles, y la segunda se representa por columnas, y toma J valores posibles. Por ejemplo, la tabla 7.1 presenta la clasificación de $n = 5387$ escolares escoceses por el color de sus ojos, que tiene cuatro categorías posibles y $I = 4$, y el color de su cabello, que tiene cinco categorías posibles y $J = 5$. Esta tabla tiene interés histórico ya que fué utilizada por Fisher en 1940 para ilustrar un método de análisis de tablas de contingencia que está muy relacionado con el que aquí presentamos.

En general, una tabla de contingencia es un conjunto de números positivos dispuestos en una matriz, donde el número en cada casilla representa la frecuencia absoluta observada para esa combinación de las dos variables.

Una manera de llegar a una tabla de contingencia $I \times J$ es definir I variables binarias para

		Color	del	pelo		
C. ojos	rubio	pelirrojo	castaño	oscuro	negro	total
claros	688	116	584	188	4	1580
azules	326	38	241	110	3	718
castaños	343	84	909	412	26	1774
oscuros	98	48	403	618	85	1315
total	1455	286	2137	1391	118	5387

Tabla 7.1: Tabla de Contingencia del color de los ojos y el color del pelo de escolares escoceses. Recogida por Fisher en 1940

las categorías de las filas y J para las de las columnas y disponer estas variables en matrices \mathbf{X}_a para las filas y \mathbf{X}_b para las columnas. Por ejemplo, la matriz \mathbf{X}_a para la variable color de los ojos contendrá 4 variables en columnas correspondientes a las 4 categorías consideradas para indicar el color de ojos, y en cada fila sólo una columna tomará el valor uno, la que corresponda al color de ojos de la persona. La matriz tendrá 5387 filas correspondientes a las personas incluidas en la muestra. Por tanto, la matriz \mathbf{X}_a de dimensiones 5387×4 será de la forma:

$$\mathbf{X}_a = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

donde hemos tomado las categorías para el color de ojos en el mismo orden que aparecen en las filas de la tabla 7.1. Por ejemplo, el primer dato corresponde a una persona de ojos claros, ya que tiene un uno en la primera columna. El segundo dato tiene un uno en la cuarta categoría, que corresponde a ojos oscuros. Finalmente, el último elemento de la matriz corresponde a una persona de ojos azules. De la misma forma, la matriz \mathbf{X}_b tendrá dimensiones 5387×5 y las columnas indicarán el color del cabello de cada persona. Observemos que estas matrices \mathbf{X} de variables binarias tienen tantas columnas como categorías y sus variables son linealmente dependientes, ya que siempre la suma de los valores de una fila es uno, al ser las categorías excluyentes y exhaustivas. Al realizar el producto $\mathbf{X}_a \mathbf{X}_b$ sumaremos todas las personas que tienen cada par de características y se obtiene la tabla de contingencia.

El análisis de correspondencias es un procedimiento para resumir la información contenida en una tabla de contingencia. Puede interpretarse de dos formas equivalentes. La primera, como una manera de representar las variables en un espacio de dimensión menor, de forma análoga a componentes principales, pero definiendo la distancia entre los puntos de manera coherente con la interpretación de los datos y en lugar de utilizar la distancia euclídea utilizamos la distancia ji-cuadrado. Desde este enfoque, el análisis de correspondencias es el equivalente de componentes principales para datos cualitativos. La segunda interpretación está más próxima al escalado multidimensional: es un procedimiento objetivo de asignar valores numéricos a variables cualitativas. Vamos a analizar estos dos aspectos.

7.2 BÚSQUEDA DE LA MEJOR PROYECCIÓN

En adelante trabajaremos con la matriz \mathbf{F} de frecuencias relativas obtenida dividiendo cada casilla por n , el total de elementos observados. Llamaremos f_{ij} a las frecuencias relativas que verifican

$$\sum_{i=1}^I \sum_{j=1}^J f_{ij} = 1$$

La matriz \mathbf{F} puede considerarse por filas o por columnas. Cualquier análisis lógico de esta matriz debe de ser equivalente al aplicado a su transpuesta, ya que la elección de la variable

	Sobre.	Not.	Aprob.	Sus.	Total
Zona A	0,03	0,06	0,15	0,06	0,3
Zona B	0,07	0,14	0,35	0,14	0,7
Total	0,1	0,2	0,5	0,2	1

Tabla 7.2: Clasificación de estudiantes por zona geográfica y calificación obtenida

que se coloca en filas, en lugar de en columnas, es arbitraria, y no debe influir en el análisis. Vamos a presentar primero el análisis por filas de esta matriz, que será simétrico al análisis por columnas, que estudiaremos a continuación.

7.2.1 Proyección de las Filas

Vamos a analizar la matriz de frecuencias relativas, \mathbf{F} , por filas. Entonces las I filas pueden tomarse como I puntos en el espacio \mathcal{R}^J . Vamos a buscar una representación de estos I puntos en un espacio de dimensión menor que nos permita apreciar sus distancias relativas. El objetivo es el mismo que con componentes principales, pero ahora tendremos en cuenta las peculiaridades de este tipo de datos. Estas peculiaridades provienen de que la frecuencia relativa de cada fila es distinta, lo que implica que:

(1) Todas las filas (puntos en \mathcal{R}^J) no tienen el mismo peso, ya que algunas contienen más datos que otras. Al representar el conjunto de las filas (puntos) debemos dar más peso a aquellas filas que contienen más datos.

(2) La distancia euclídea entre puntos no es una buena medida de su proximidad y debemos modificar esta distancia, como veremos a continuación.

Comenzando con el primer punto, cada fila de la matriz \mathbf{F} tiene una frecuencia relativa $f_{i.} = \sum_{j=1}^J f_{ij}$, y el conjunto de estas frecuencias relativas se calcula con:

$$\mathbf{f} = \mathbf{F}'\mathbf{1}$$

debemos dar a cada fila un peso proporcional a su frecuencia relativa y los términos del vector \mathbf{f} pueden directamente considerarse como pesos, ya que son números positivos que suman uno.

Con relación a la medida de distancia a utilizar entre las filas, observemos que la distancia euclídea no es una buena medida de las diferencias reales entre las estructuras de las filas. Por ejemplo, supongamos la tabla 7.2 donde se presentan las frecuencias relativas de estudiantes clasificados por su procedencia geográfica, (A ó B) y sus calificaciones. Aunque las frecuencias relativas de las dos filas son muy distintas, las dos filas tienen exactamente la misma estructura relativa: simplemente, hay más del doble de estudiantes de la zona B que de la A, pero la distribución de calificaciones es idéntica en ambas zonas. Si calculamos la distancia euclídea entre las zonas obtendremos un valor alto, que no refleja una estructura distinta de las filas sino sólo que tienen distinta frecuencia relativa. Supongamos que dividimos cada casilla por la frecuencia relativa de la fila, $f_{i.}$. Con esto se obtiene la tabla 7.3 donde los números que aparecen en las filas representan la frecuencia relativa de la variable columna condicionada a la variable fila. Ahora las dos filas son idénticas, y esto es coherente con una distancia euclídea cero entre ambas.

	Sobre.	Not.	Aprob.	Sus.	Total
Zona A	0,1	0,2	0,5	0,2	1
Zona B	0,1	0,2	0,5	0,2	1

Tabla 7.3: Clasificación de estudiantes por zona geográfica y calificación obtenida

	Color del cabello					
C. ojos	rubio	pelirrojo	castaño	oscuro	negro	total
claros	0.435	0.073	0.369	0.119	0.003	1
azules	0.454	0.053	0.336	0.153	0.004	1
castaños	0.193	0.047	0.512	0.232	0.015	1
oscuros	0.075	0.037	0.307	0.518	0.065	1

Tabla 7.4: Tabla de frecuencias relativas del color del cabello condicionada al color de los ojos para los escolares escoceses

Para analizar que medida de distancia debemos utilizar, llamaremos \mathbf{R} a la matriz de frecuencias relativas condicionadas al total de la fila, que se obtiene con:

$$\mathbf{R} = \mathbf{D}_f^{-1} \mathbf{F} \quad (7.1)$$

donde \mathbf{D}_f es una matriz diagonal $I \times I$ con los términos del vector \mathbf{f} , f_i , frecuencias relativas de las filas, en la diagonal principal. Esta operación transforma la matriz original de frecuencias relativas, \mathbf{F} , en otra matriz cuyas casillas por filas suman uno. Cada fila de esta matriz representa la distribución de la variable en columnas condicionada al atributo que representa la fila. Por ejemplo, la tabla 7.4 presenta las frecuencias relativas condicionadas para la tabla 7.1. En este caso $I = 4$, $J = 5$. Esta tabla permite apreciar mejor la asociación entre las características estudiadas.

Llamaremos \mathbf{r}'_i a la fila i de la matriz \mathbf{R} de frecuencias relativas condicionadas por filas, que puede considerarse un punto (o un vector) en el espacio \mathbb{R}^J . Como la suma de los componentes de \mathbf{r}'_i es uno, todos los puntos están en un espacio de dimensión $J-1$. Queremos proyectar estos puntos en un espacio de dimensión menor de manera que las filas que tengan la misma estructura estén próximas, y las que tengan una estructura muy diferente, alejadas. Para ello, debemos definir una medida de distancia entre dos filas $\mathbf{r}_a, \mathbf{r}_b$. Una posibilidad es utilizar la distancia euclídea, pero esta distancia tiene el inconveniente de tratar igual a todos los componentes de estos vectores. Por ejemplo, en la tabla 7.1 las personas de cabello rubio tienen una diferencia en frecuencia relativa entre los ojos azules y claros de $0,454-0,435=0,019$, y las personas de cabello negro tienen una diferencia en frecuencia relativa entre los ojos castaños y azules de $0,015 - 0,004=0,011$. Hay una diferencia mayor en el primer caso que en el segundo y, sin embargo, intuitivamente vemos que la segunda diferencia es mayor que la primera. La razón es que en el primer caso el cambio relativo es pequeño, del orden del 4% ($0,019/0,454$), mientras que en el segundo caso el cambio relativo es muy grande: las personas de cabello negro tienen ojos castaños casi cuatro veces más frecuentemente ($0,015/0,004=3,75$ veces) que ojos azules. Como los componentes representan frecuencias relativas, no parece adecuado que una diferencia de 0,01 se considere igual en un atributo

de alta frecuencia (por ejemplo, pasar de 0,60 a 0,61) que en un atributo de baja frecuencia (por ejemplo, pasar de 0,0001 a 0,0101).

Para obtener comparaciones razonables entre estas frecuencias relativas tenemos que tener en cuenta la frecuencia relativa de aparición del atributo que estudiamos. En atributos raros, pequeñas diferencias absolutas pueden ser grandes diferencias relativas, mientras que en atributos con gran frecuencia, la misma diferencia será poco importante. Una manera intuitiva de construir las comparaciones es ponderar las diferencias en frecuencia relativa entre dos atributos inversamente proporcional a la frecuencia de este atributo. Es decir, en lugar de sumar los términos $(r_{aj} - r_{bj})^2 = (f_{aj}/f_{a.} - f_{bj}/f_{b.})^2$ que miden la diferencia que las filas a y b tienen en la columna j sumaremos los términos $(r_{aj} - r_{bj})^2 / f_{.j}$ donde $f_{.j} = \sum_{i=1}^I f_{ij}$ es la frecuencia relativa de la columna j . La expresión de la distancia entre dos filas, \mathbf{r}_a y \mathbf{r}_b de \mathbf{R} vendrá dada en esta métrica por

$$D^2(\mathbf{r}_a, \mathbf{r}_b) = \sum_{j=1}^J \left(\frac{f_{aj}}{f_{a.}} - \frac{f_{bj}}{f_{b.}} \right)^2 \frac{1}{f_{.j}} = \sum_{j=1}^J \frac{(r_{aj} - r_{bj})^2}{f_{.j}} \quad (7.2)$$

que puede escribirse matricialmente como

$$D^2(\mathbf{r}_a, \mathbf{r}_b) = (\mathbf{r}_a - \mathbf{r}_b)' \mathbf{D}_c^{-1} (\mathbf{r}_a - \mathbf{r}_b) \quad (7.3)$$

donde \mathbf{D}_c es una matriz diagonal con términos $f_{.j}$. A la distancia (7.2) ó (7.3) se la conoce como distancia χ^2 , y se analizará con más detalle en la sección siguiente.

Observemos que esta distancia equivale a la distancia euclídea entre los vectores transformados $\mathbf{y}_i = \mathbf{D}_c^{-1/2} \mathbf{r}_i$. Podemos pues simplificar el problema definiendo una matriz de datos transformada, sobre la que tiene sentido considerar la distancia euclídea entre filas. Llamando:

$$\mathbf{Y} = \mathbf{R} \mathbf{D}_c^{-1/2} = \mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1/2} \quad (7.4)$$

obtenemos una matriz \mathbf{Y} que contiene términos del tipo

$$y_{ij} = \left\{ \frac{f_{ij}}{f_{i.} f_{.j}^{1/2}} \right\} \quad (7.5)$$

que ya no suman uno ni por filas ni por columnas. Las casillas de esta matriz representan las frecuencias relativas condicionadas por filas, $f_{ij}/f_{i.}$, pero estandarizadas por su variabilidad, que depende de la raíz cuadrada de la frecuencia relativa de la columna. De esta manera las casillas son directamente comparables entre sí. La tabla 7.5 indica este matriz resultado de estandarizar las frecuencias relativas de la tabla 7.1 dividiendo cada casilla por la raíz cuadrada de la frecuencia relativa de la columna correspondiente, que se obtiene de la tabla 7.1. Por ejemplo, el primer elemento de la tabla 7.5 se obtiene como $0.435/\sqrt{(1455/5387)} =$

.837	.316	.587	.235	.015
.873	.228	.536	.301	.029
.374	.205	.815	.455	.095
.147	.161	.484	1.022	.440

Tabla 7.5: Matriz estandarizada por fila y por variabilidad del color de los ojos y el color del pelo de escolares

0.0114. En esta tabla la estructura de las columnas es similar a la de la tabla 7.1 de frecuencias relativas, ya que hemos dividido todas las casillas de cada columna por la misma cantidad.

Podríamos tratar a esta matriz como una matriz de datos estándar, con observaciones en filas y variables en columnas, y preguntarnos como proyectarla de manera que se preserven las distancias relativas entre las filas, es decir, las filas con estructura similar aparezcan próximas en la proyección. Esto implica encontrar una dirección \mathbf{a} de norma unidad,

$$\mathbf{a}'\mathbf{a} = 1 \quad (7.6)$$

tal que el vector de puntos proyectados sobre esta dirección,

$$\mathbf{y}_p(\mathbf{a}) = \mathbf{Y} \mathbf{a} \quad (7.7)$$

tenga variabilidad máxima. El vector \mathbf{a} se encontrará maximizando $\mathbf{y}_p(\mathbf{a})'\mathbf{y}_p(\mathbf{a}) = \mathbf{a}'\mathbf{Y}'\mathbf{Y} \mathbf{a}$ con la condición (7.6), y este problema se ha resuelto en el capítulo 5 al estudiar componentes principales: el vector \mathbf{a} es un vector propio de la matriz $\mathbf{Y}'\mathbf{Y}$. Sin embargo, este tratamiento de la matriz \mathbf{Y} como una matriz de variables continuas no es del todo correcto porque las filas tienen una distinta frecuencia relativa, $f_{i.}$, y por tanto deben tener distinto peso. Aquellas filas con mayor frecuencia relativa deben de tener más peso en la representación que aquellas otras con frecuencia relativa muy baja, de manera que las filas con gran número de individuos estén bien representadas, aunque esto sea a costa de representar peor las filas con pocos elementos. En consecuencia, daremos a cada fila un peso proporcional al número de datos que contiene. Esto puede hacerse maximizando la suma de cuadrados ponderada.

$$m = \mathbf{a}'\mathbf{Y}'\mathbf{D}_f \mathbf{Y} \mathbf{a} \quad (7.8)$$

sujeto a (7.6), que equivale a

$$m = \mathbf{a}'\mathbf{D}_c^{-1/2}\mathbf{F}'\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2} \mathbf{a}. \quad (7.9)$$

Alternativamente, podemos construir una matriz de datos \mathbf{Z} definida por

$$\mathbf{Z} = \mathbf{D}_f^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2} \quad (7.10)$$

cuyos componentes son

$$z_{ij} = \left\{ \frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}} \right\}$$

y que estandariza las frecuencias relativas en cada casilla por el producto de las raíces cuadradas de las frecuencias relativas totales de la fila y la columna, y escribir el problema de encontrar el vector \mathbf{a} como el problema de maximizar $m = \mathbf{a}' \mathbf{Z}' \mathbf{Z} \mathbf{a}$ sujeto a la restricción (7.6). Este es el problema resuelto en componentes principales, cuya solución es

$$\mathbf{D}_c^{-1/2} \mathbf{F}' \mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1/2} \mathbf{a} = \lambda \mathbf{a} \quad (7.11)$$

y \mathbf{a} debe ser un vector propio de la matriz $\mathbf{Z}' \mathbf{Z}$ donde \mathbf{Z} está dado por (7.9) y λ su valor propio.

Vamos a comprobar que la matriz $\mathbf{Z}' \mathbf{Z}$ tiene como mayor valor propio siempre el 1 y como vector propio $\mathbf{D}_c^{-1/2}$. Multiplicando por la izquierda en (7.11) por $\mathbf{D}_c^{-1/2}$ se obtiene:

$$\mathbf{D}_c^{-1} \mathbf{F}' \mathbf{D}_f^{-1} \mathbf{F} (\mathbf{D}_c^{-1/2} \mathbf{a}) = \lambda (\mathbf{D}_c^{-1/2} \mathbf{a})$$

Las matrices $\mathbf{D}_f^{-1} \mathbf{F}$ y $\mathbf{F} \mathbf{D}_c^{-1}$ representan matrices de frecuencias relativas por filas y por columnas y su suma por filas y columnas respectivamente es uno. Por tanto $\mathbf{D}_f^{-1} \mathbf{F} \mathbf{1} = \mathbf{1}$ y $\mathbf{D}_c^{-1} \mathbf{F}' \mathbf{1} = \mathbf{1}$, que implica que la matriz $\mathbf{D}_c^{-1} \mathbf{F}' \mathbf{D}_f^{-1} \mathbf{F}$ tiene un valor propio 1 unido a un vector propio $\mathbf{1}$. En consecuencia, haciendo $(\mathbf{D}_c^{-1/2} \mathbf{a}) = \mathbf{1}$ concluimos que la matriz $\mathbf{Z}' \mathbf{Z}$ tiene un valor propio igual a uno con vector propio $\mathbf{D}_c^{-1/2}$.

Olvidando esta solución trivial, que no da información sobre la estructura de las filas, tomaremos el valor propio mayor menor que la unidad y su vector propio asociado \mathbf{a} . Entonces, proyectando la matriz \mathbf{Y} sobre la dirección \mathbf{a} encontrada:

$$\mathbf{y}_f(\mathbf{a}) = \mathbf{Y} \mathbf{a} = \mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1/2} \mathbf{a} \quad (7.12)$$

y el vector $\mathbf{y}_f(\mathbf{a})$ es la mejor representación de las filas de la tabla de contingencia en una dimensión. Análogamente, si extraemos el vector propio ligado al siguiente mayor valor propio obtenemos una segunda coordenada y podemos representar las filas en un espacio de dimensión dos. Las coordenadas de la representación de cada fila vendrán dadas por las filas de la matriz

$$\mathbf{C}_f = \mathbf{Y} \mathbf{A}_2 = \mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1/2} \mathbf{A}_2$$

donde $\mathbf{A}_2 = [\mathbf{a}_1 \mathbf{a}_2]$ contiene en columnas los dos vectores propios $\mathbf{Z}' \mathbf{Z}$. La matriz \mathbf{C}_f es $I \times 2$ y las dos coordenadas de cada fila proporcionan la mejor representación de las filas de la matriz \mathbf{F} en un espacio de dos dimensiones. El procedimiento se extiende sin dificultad para representaciones en más dimensiones, calculando vectores propios adicionales de la matriz $\mathbf{Z}' \mathbf{Z}$.

En resumen el procedimiento que hemos presentado para buscar una buena representación de las filas de la tabla de contingencia es:

- (1) Caracterizar las filas por sus frecuencias relativas condicionadas, y considerarlas como puntos en el espacio.
- (2) Definir la distancia entre los puntos por la distancia χ^2 , que tiene en cuenta que cada coordenada de las filas tiene distinta precisión.

(3) Proyectar los puntos sobre las direcciones de máxima variabilidad, teniendo en cuenta que cada fila tiene un peso distinto e igual a su frecuencia relativa.

El procedimiento operativo para obtener la mejor representación bidimensional de las filas de la tabla de contingencia es:

- (1) Calcular la matriz $\mathbf{Z}'\mathbf{Z}$ y obtener sus vectores y valores propios.
- (2) Tomar los dos vectores propios, $\mathbf{a}_1, \mathbf{a}_2$, ligados a los mayores valores propios menores que la unidad de esta matriz.
- (3) Calcular las proyecciones $\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{a}_i$, $i = 1, 2$, y representarlas gráficamente en un espacio bidimensional.

Ejemplo 7.1 *Aplicaremos este análisis a la matriz de la tabla 7.1. La matriz de frecuencias relativas estandarizada por filas, \mathbf{R} , se presenta en la tabla 7.4.*

La variable transformada, \mathbf{Y} , se calcula como

$$\mathbf{Y} = \mathbf{R} \mathbf{D}_c^{-1/2} = \mathbf{R} \left(\frac{1}{5387} \begin{bmatrix} 1455 & & & & \\ & 286 & & & \\ & & 2137 & & \\ & & & 1391 & \\ & & & & 118 \end{bmatrix} \right)^{-1/2}$$

dando lugar a

$$\mathbf{Y} = \begin{bmatrix} .837 & .316 & .587 & .235 & .015 \\ .873 & .228 & .536 & .301 & .029 \\ .374 & .205 & .815 & .455 & .095 \\ .147 & .161 & .484 & 1.022 & .440 \end{bmatrix}$$

Esta matriz puede interpretarse como una matriz de datos donde por filas tenemos observaciones y por columnas variables. Para obtener la mejor representación de las filas en un espacio de dimensión dos, vamos a obtener los vectores propios de la matriz $\mathbf{Y}\mathbf{D}_f\mathbf{Y}$. Los tres primeros valores y vectores propios de esta matriz se presentan en la tabla siguiente por filas:

valor propio	vector				propio
1	-0.5197	-0.2304	-0.6298	-0.5081	-0.1480
0.1992	-0.6334	-0.1204	-0.0593	0.6702	0.3629
0.0301	-0.5209	-0.0641	0.7564	-0.3045	-0.2444

Los otros dos valores propios de esta matriz son 0,0009 0,0000. La proyección de los puntos sobre el espacio definido por los valores propios .1992 y .0301 se presenta en la figura 7.1

El eje de abscisas contiene la primera dimensión que explica el .1992/(.1992+.0301+.0009)=.8653. Vemos que se separan claramente los ojos claros y azules frente a castaños y oscuros. La primera dimensión es pues claro frente a oscuro. La segunda dimensión separa las características puras, ojos claros o azules y negros, frente a la mezclada, castaños.

Ejemplo 7.2 *En un estudio de mercado 4 evaluadores han indicado que características consideran importantes en un tipo de producto. El resultado es la matriz \mathbf{F} donde en columnas se representan los evaluadores y en filas los productos.*

Figura 7.1: Proyección de las filas de la matriz de los colores de ojos y pelo sobre el mejor espacio de dimensión 2.

$$F = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline c_1 & 0 & 0 & 1 & 0 \\ c_2 & 1 & 1 & 0 & 0 \\ c_3 & 0 & 1 & 0 & 1 \\ c_4 & 0 & 0 & 0 & 1 \\ c_5 & 0 & 1 & 0 & 0 \\ c_6 & 1 & 1 & 1 & 0 \end{array}$$

Esta matriz es una tabla de contingencia muy simple donde las frecuencias posibles son cero o uno. La matriz \mathbf{Z} es

$$\mathbf{Z} = \begin{bmatrix} 0 & 0 & .707 & 0 \\ .5 & .35 & 0 & 0 \\ 0 & .35 & 0 & .50 \\ 0 & 0 & 0 & .707 \\ 0 & .5 & 0 & 0 \\ .408 & .289 & .408 & 0 \end{bmatrix}$$

y los valores propios de $\mathbf{Z}'\mathbf{Z}$ son (1, 0.75, 0.50, 0.17). El vector propio asociado al mayor valor propio menor que uno es $v = (0.27, 0, 0.53, -0.80)$. La proyección de las filas de \mathbf{Y} sobre las dos direcciones principales conduce a la figura 7.2

Se observa que las características más próximas son la 2 y la 5. Las elecciones de los evaluadores parecen ser debidas a dos dimensiones. La primera explica el $0.75/(0.75+0.50+0.17)=52.83\%$ de la variabilidad y la segunda el 35%. La primera dimensión tiene en cuenta las similitudes

Figura 7.2: Proyección de las características de los productos

aparentes por las elecciones de las personas: las características c3 y c4 son elegidas por la misma persona y por nadie más, por lo que estas características aparecen juntas en un extremo. En el lado opuesto aparecen la c1 y c6, que son elegidas por la misma persona, y las c2 y c5 que son elegidas por personas que también eligen la c6. En la segunda dimensión las características extremas son las c1 y c2.

7.2.2 Proyección de las columnas

Podemos aplicar a las columnas de la matriz \mathbf{F} un análisis equivalente al de las filas. Las columnas serán ahora puntos en \mathbb{R}^I . Llamando

$$\mathbf{c} = \mathbf{F}'\mathbf{1}$$

al vector de frecuencias relativas de las columnas y \mathbf{D}_c a la matriz diagonal que contiene estas frecuencias relativas en la diagonal principal, de acuerdo con la sección anterior la mejor representación de los J puntos (columnas) en un espacio de dimensión menor, con la métrica χ^2 conducirá, por simetría, a estudiar la matriz $\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1/2}$. Observemos que, si ahora consideramos la matriz \mathbf{F}' y volvemos al problema de representarla por filas (que es equivalente a representar \mathbf{F} por columnas), el problema es idéntico al que hemos resuelto en la sección anterior. Ahora la matriz que contiene las frecuencias relativas de las filas \mathbf{F}' es \mathbf{D}_c y la que contiene la de las columnas es \mathbf{D}_f . Intercambiando el papel de estas matrices, las direcciones de proyección son los vectores propios de la matriz

$$\mathbf{Z}\mathbf{Z}' = \mathbf{D}_f^{-1/2}\mathbf{F}\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1/2} \quad (7.13)$$

donde \mathbf{Z} es la matriz $I \times J$ definida por (7.10). Como $\mathbf{Z}'\mathbf{Z}$ y $\mathbf{Z}\mathbf{Z}'$ tienen los mismos valores propios no nulos, esa matriz tendrá también un valor propio unidad ligado al vector propio $\mathbf{1}$. Esta solución trivial no se considera. Llamando \mathbf{b} al vector propio ligado al mayor valor

propio distinto de la unidad de $\mathbf{Z}\mathbf{Z}'$, la mejor representación de las columnas de la matriz en un espacio de dimensión uno vendrá dada por

$$\mathbf{y}_c(\mathbf{b}) = \mathbf{Y}'\mathbf{b} = \mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1/2}\mathbf{b} \quad (7.14)$$

y, análogamente, la mejor representación en dimensión dos de las columnas de la matriz vendrá dada por las coordenadas definidas por las filas de la matriz

$$\mathbf{C}_c = \mathbf{Y}'\mathbf{B}_2 = \mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1/2}\mathbf{B}_2$$

donde $\mathbf{B}_2 = [\mathbf{b}_1\mathbf{b}_2]$ contiene en columnas los dos vectores propios ligados a los valores propios mayores de $\mathbf{Z}\mathbf{Z}'$ y menores que la unidad. La matriz \mathbf{C}_c es $J \times 2$ y cada fila es la mejor representación de las columnas de la matriz \mathbf{F} en un espacio de dos dimensiones.

7.2.3 Análisis Conjunto

Dada la simetría del problema conviene representar conjuntamente las filas y las columnas de la matriz. Observemos que las matrices $\mathbf{Z}'\mathbf{Z}$ y $\mathbf{Z}\mathbf{Z}'$ tienen los mismos valores propios no nulos y que los vectores propios de ambas matrices que corresponden al mismo valor propio están relacionados. En efecto, si \mathbf{a}_i es un vector propio de $\mathbf{Z}'\mathbf{Z}$ ligado al valor propio λ_i :

$$\mathbf{Z}'\mathbf{Z}\mathbf{a}_i = \lambda_i\mathbf{a}_i$$

entonces, multiplicando por \mathbf{Z}

$$\mathbf{Z}\mathbf{Z}'(\mathbf{Z}\mathbf{a}_i) = \lambda_i(\mathbf{Z}\mathbf{a}_i)$$

y obtenemos que $\mathbf{b}_i = \mathbf{Z}\mathbf{a}_i$ es un vector propio de $\mathbf{Z}\mathbf{Z}'$ ligado al valor propio λ_i . Una manera rápida de obtener estos vectores propios es calcular directamente los vectores propios de la matriz de dimensión más pequeña, $\mathbf{Z}'\mathbf{Z}$ o $\mathbf{Z}\mathbf{Z}'$, y obtener los otros vectores propios como $\mathbf{Z}\mathbf{a}_i$ o $\mathbf{Z}'\mathbf{b}_i$. Alternativamente podemos utilizar la descomposición en valores singulares de la matriz \mathbf{Z} o \mathbf{Z}' , estudiada al introducir los biplots en el capítulo anterior. Esta descomposición aplicada a \mathbf{Z} es

$$\mathbf{Z} = \mathbf{B}_r\mathbf{D}_r\mathbf{A}_r' = \sum_{i=1}^r \lambda_i^{1/2}\mathbf{b}_i\mathbf{a}_i'$$

donde \mathbf{B}_r contiene en columnas los vectores propios de $\mathbf{Z}\mathbf{Z}'$, \mathbf{A}_r los de $\mathbf{Z}'\mathbf{Z}$ y \mathbf{D}_r es digonal y contiene los valores singulares, $\lambda_i^{1/2}$, o raíces de los valores propios no nulos y $r = \min(I, J)$. Entonces la representación de las filas se obtiene con (7.12) y la de las columnas con (7.14). La representación de la matriz \mathbf{Z} con h dimensiones (habitualmente $h = 2$) implica aproximar esta matriz mediante $\hat{\mathbf{Z}}_h = \mathbf{B}_h\mathbf{D}_h\mathbf{A}_h'$. Esto es equivalente, por (7.10), a una aproximación a la tabla de contingencia observada mediante:

$$\hat{\mathbf{F}}_h = \mathbf{D}_f^{1/2}\hat{\mathbf{Z}}_h\mathbf{D}_c^{1/2}, \quad (7.15)$$

y una forma de juzgar la aproximación que estamos utilizando es reconstruir la tabla de contingencia con esta expresión.

Si deseamos eliminar el valor propio unidad desde el principio, dado que no aparta información de interés, podemos reemplazar la matriz \mathbf{F} por $\mathbf{F} - \widehat{\mathbf{F}}_e$, donde $\widehat{\mathbf{F}}_e$ es la matriz de frecuencias esperadas que viene dada por

$$\widehat{\mathbf{F}}_e = \frac{1}{n} \mathbf{r} \mathbf{c}'.$$

Puede comprobarse que la matriz $\mathbf{F} - \widehat{\mathbf{F}}_e$ tiene rango $r - 1$, y ya no tiene el valor propio igual a la unidad.

La proporción de variabilidad explicada por cada dimensión se calcula como en componentes principales descartando el valor propio igual a uno y tomando la proporción que representa cada valor propio con relación al resto.

En resumen, el análisis de correspondencias de una tabla de contingencia de dimensiones $I \times J$ se realiza en los pasos siguientes

(1) Se calcula la tabla de frecuencias relativas, \mathbf{F} .

(1) Se calcula la tabla estandarizada \mathbf{Z} , de frecuencias relativas las mismas dimensiones de la tabla original, $I \times J$, dividiendo cada celda de \mathbf{F} por la raíz de los totales de su fila y columna, $z_{ij} = \{f_{ij} / \sqrt{f_{i.} f_{.j}}\}$.

(2) Se calculan los h (normalmente $h = 2$) vectores propios ligados a valores propios mayores, pero distintos de la unidad, de la matriz de menor dimensión de las $\mathbf{Z}\mathbf{Z}'$ y $\mathbf{Z}'\mathbf{Z}$. Si obtenemos los vectores propios \mathbf{a}_i de $\mathbf{Z}'\mathbf{Z}$, los \mathbf{b}_i de $\mathbf{Z}\mathbf{Z}'$ se obtienen por $\mathbf{b}_i = \mathbf{Z}\mathbf{a}_i$. Análogamente si se obtienen los \mathbf{b}_i de $\mathbf{Z}\mathbf{Z}'$ $\mathbf{a}_i = \mathbf{Z}'\mathbf{b}_i$. Las I filas de la matriz se presentarán como I puntos en \mathbb{R}^h y las coordenadas de cada fila vienen dadas por

$$\mathbf{C}_f = \mathbf{D}_f^{-1/2} \mathbf{Z} \mathbf{A}_2$$

donde \mathbf{A}_2 tiene en columnas los dos vectores propios de $\mathbf{Z}'\mathbf{Z}$. Las J columnas se representarán como J puntos en \mathbb{R}^h y las coordenadas de cada columna son

$$\mathbf{C}_c = \mathbf{D}_c^{-1/2} \mathbf{Z}' \mathbf{B}_2$$

Ejemplo 7.3 *Vamos a representar conjuntamente las filas y las columnas de la matriz de los colores. La figura 7.3 presenta esta representación. Se observa que el gráfico describe de manera clara la relación entre ambas variables. La dimensión principal gradúa la tonalidad de claro a oscuro y la segunda separa los castaños de los casos más extremos.*

Es importante calcular conjuntamente los vectores propios para evitar problemas de signos, ya sea calculando los vectores propios de una matriz y obteniendo los otros como producto por la matriz \mathbf{Z} o bien a través de la descomposición en valores singulares. La razón es que si \mathbf{v} es un vector propio también lo es $-\mathbf{v}$ y al calcular separadamente las coordenadas y superponerlas podemos obtener un resultado como el que se presenta en la figura 7.4. En esta figura se han calculado separadamente las dos representaciones y luego se han superpuesto. El lector puede comprobar que si cambiamos de signo las coordenadas del eje de ordenadas se obtiene la representación de la figura (7.3). Estos problemas de signos se evitan calculado los vectores conjuntamente.

Figura 7.3: Representación de los colores de ojos y cabello para los escolares escoceses.

Figura 7.4:

7.3 LA DISTANCIA JI-CUADRADO

El contraste de independencia entre las variables fila y columna en una tabla de contingencia $I \times J$ se realiza con la estadístico

$$X^2 = \sum \frac{(\text{fr. observadas} - \text{fr. esperadas})^2}{\text{fr. esperadas}}$$

que, en la hipótesis de independencia, sigue una distribución χ^2 con $(I - 1) \times (J - 1)$ grados de libertad. De acuerdo con la notación anterior, la frecuencia esperada en cada celda de la fila i , suponiendo independencia de filas y columnas, se obtendrá repartiendo el total de la fila, $nf_{i.}$, proporcionalmente a la frecuencia relativa de cada columna, $f_{.j}$. Por ejemplo, la frecuencia esperada de la primera casilla de la tabla 5.1 se obtendrá multiplicando el número total de elementos de la fila, 1580, por la proporción de personas rubias sobre el total, 1455/5387. Por tanto, el estadístico X^2 para contrastar la independencia puede escribirse:

$$X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}} \quad (7.16)$$

donde $f_{i.} = \sum_{j=1}^J f_{ij}$ es la frecuencia relativa de la fila i y $f_{.j} = \sum_{i=1}^I f_{ij}$ la de columna j . Como

$$\frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}} = \frac{nf_{i.}}{f_{.j}} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}^2}$$

la expresión del estadístico X^2 puede también escribirse como :

$$X^2 = n \sum_{i=1}^I f_{i.} \sum_{j=1}^J \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 \frac{1}{f_{.j}}. \quad (7.17)$$

En esta representación la distribución condicionada de las frecuencias relativas de cada fila, $\left\{ \frac{f_{ij}}{f_{i.}} \right\}$, se compara con la distribución media de las filas $\{f_{.j}\}$, y cada coordenada se pondera inversamente a la frecuencia relativa que existe en esa columna. Se suman luego todas las filas, pero dando a cada fila un peso tanto mayor cuanto mayor es su frecuencia, $nf_{i.}$.

Vamos a ver que esta representación es equivalente a calcular las distancias entre los vectores de la matriz de frecuencias relativas por filas, \mathbf{R} , definida en (7.1) si medimos la distancia con la métrica χ^2 . Consideremos los vectores \mathbf{r}'_i , filas de la matriz \mathbf{R} . La media de estos vectores es

$$\bar{\mathbf{r}} = \frac{\sum_{i=1}^I w_i \mathbf{r}_i}{\sum_{i=1}^I w_i}$$

donde los w_i son coeficientes de ponderación. La media aritmética se obtiene con $w_i = 1$, dando a todas las filas el mismo peso. Sin embargo, en este caso esta ponderación no es

conveniente, porque debemos dar más peso a las filas que contengan más datos. Podemos ponderar por la frecuencia relativa de cada fila, $w_i = f_{i.}$, y entonces $\sum w_i = \sum f_{i.} = 1$. Como las frecuencias relativas de las filas vienen dadas por el vector columna $\mathbf{D}_f \mathbf{1}$, tenemos que

$$\bar{\mathbf{r}} = \mathbf{R}' \mathbf{D}_f \mathbf{1}$$

y utilizando (7.1)

$$\bar{\mathbf{r}} = \mathbf{F}' \mathbf{D}_f^{-1} \mathbf{D}_f \mathbf{1} = \mathbf{F}' \mathbf{1} = \mathbf{c}$$

y el valor medio de las filas viene dado por el vector cuyos componentes son las frecuencias relativas de las columnas. La distancia de cualquier vector de fila, \mathbf{r}_i , a su media, \mathbf{c} , con la métrica χ^2 será

$$(\mathbf{r}_i - \mathbf{c})' \mathbf{D}_c^{-1} (\mathbf{r}_i - \mathbf{c})$$

donde la matriz \mathbf{D}_c^{-1} se obtuvo en (7.3) para construir la distancia χ^2 . La suma de todas estas distancias, ponderadas por su importancia, que se conoce como inercia total de la tabla, es

$$I_T = \sum_{i=1}^I f_{i.} (\mathbf{r}_i - \mathbf{c})' \mathbf{D}_c^{-1} (\mathbf{r}_i - \mathbf{c})$$

y esta expresión puede escribirse como

$$I_T = \sum_{i=1}^I f_{i.} \sum_{j=1}^J \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 / f_{.j}$$

y si comparamos con (7.17) vemos que la inercia total es igual a X^2/n .

Se demuestra que la inercia total es la suma de los valores propios de la matriz $\mathbf{Z}'\mathbf{Z}$ eliminado el uno. Por tanto, el análisis de las filas (o de las columnas ya que el problema es simétrico) puede verse como una descomposición de los componentes del estadístico X^2 en sus fuentes de variación.

La distancia χ^2 tiene una propiedad importante que se conoce como el principio de equivalencia distribucional. Esta propiedad es que si dos filas tienen la misma estructura relativa, $f_{ij}/f_{i.}$ y las unimos en una nueva fila única, las distancias entre las restantes filas permanecen invariables. Esta misma propiedad por simetría se aplica a las columnas. Esta propiedad es importante, porque asegura una cierta invarianza del procedimiento ante agregaciones o desagregaciones irrelevantes de las categorías. Para demostrarlo, consideremos la distancia χ^2 entre las filas a y b

$$\sum_{j=1}^J \left(\frac{f_{aj}}{f_{a.}} - \frac{f_{bj}}{f_{b.}} \right)^2 \frac{1}{f_{.j}}$$

es claro que esta distancia no se modifica si unimos dos filas en una, ya que esta unión no va a afectar a las frecuencias $f_{ij}/f_{i.}$ ni tampoco a $f_{.j}$. Vamos a comprobar que si unimos dos filas con la misma estructura la distancia de la nueva fila al resto es la misma que las

de las filas originales. En efecto, supongamos que para las filas 1 y 2, se verifica que para $j = 1, \dots, J$

$$\frac{f_{1j}}{f_{1.}} = \frac{f_{2j}}{f_{2.}} = g_j$$

entonces, si unimos estas dos filas en una nueva fila, se obtiene que, para la nueva fila

$$\frac{f_{1j} + f_{2j}}{f_{1.} + f_{2.}} = g_j$$

y su distancia a cualquier otra fila permanecerá invariable.

Esta propiedad garantiza que no perdemos nada al agregar categoría homogéneas ni podemos ganar nada por desagregar una categoría homogénea.

Ejemplo 7.4 *Se han contabilizado los pesos y las alturas de 100 estudiantes universitarios y se han formado 4 categorías tanto para el peso como para la altura. Para el peso, las categorías se denotan P1, de 51 a 60 k., P2, de 61 a 70 k., P3, de 71 a 80 k. y P4, de 81 a 90 k. Para la altura se denotan A1, de 151 a 160 cm., A2, de 161 a 170 cm., A3, de 171 a 180 cm. y A4, de 181 a 190 cm. La siguiente tabla de contingencia muestra las frecuencias de cada grupo:*

Peso/Altura	A1	A2	A3	A4
P1	15	8	3	0
P2	10	15	7	2
P3	2	7	17	3
P4	0	2	3	6

Realizar proyecciones por filas, por columnas y conjunta de filas y columnas. Comprobar como las proyecciones por filas y por columnas separan claramente las categorías, pero que la proyección conjunta asocia claramente cada categoría de un peso con la de una altura.

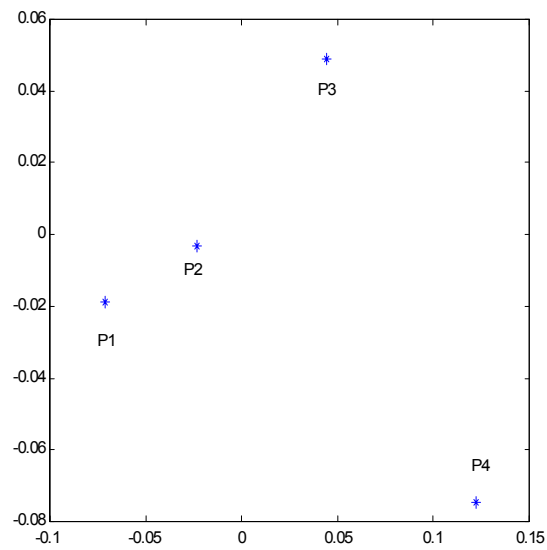
Para la proyección por filas, la variable Y queda:

$$Y = RD_c^{-\frac{1}{2}} = \begin{bmatrix} 0.1110 & 0.0544 & 0.0211 & 0 \\ 0.0566 & 0.0780 & 0.0376 & 0.0177 \\ 0.0133 & 0.0427 & 0.1070 & 0.0312 \\ 0 & 0.0321 & 0.0498 & 0.1645 \end{bmatrix}$$

Los tres valores propios y vectores propios diferentes de uno de esta matriz son:

valor propio	vector propio			
0.3717	-0.6260	-0.1713	0.3673	0.6662
0.1401	-0.2974	-0.0064	0.6890	-0.6610
0.0261	0.4997	-0.8066	0.3007	0.0964

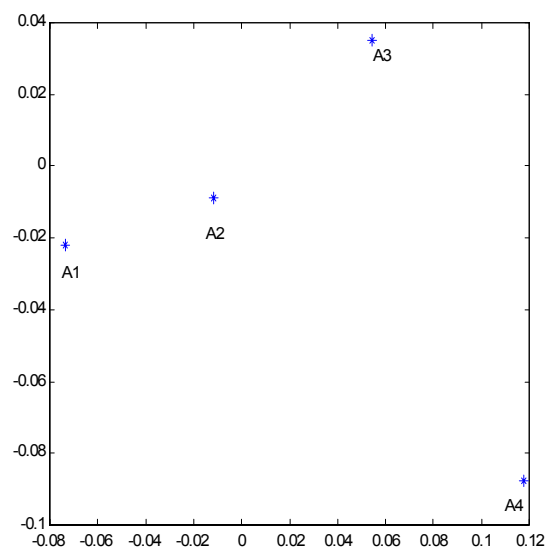
La proyección por filas es:



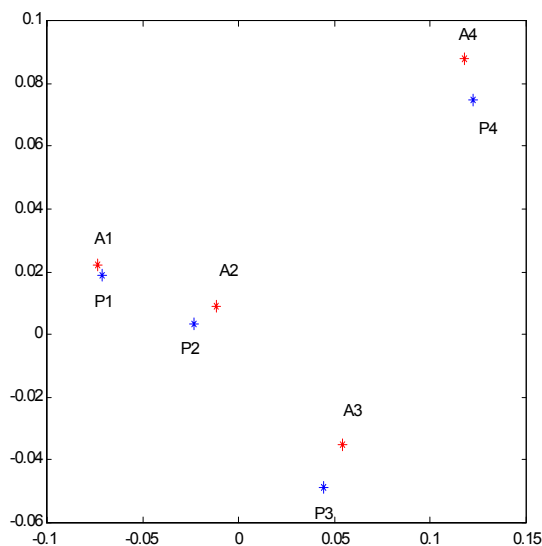
Para las columnas, los tres valores propios y vectores propios diferentes de uno de esta matriz son:

valor propio	vector propio			
0.3717	-0.5945	-0.2216	0.3929	0.6656
0.1401	-0.2568	-0.0492	0.7034	-0.6609
0.0261	0.5662	-0.7801	0.2466	0.1005

La proyección por columnas es:



El resultado de la proyección conjunta es el siguiente donde alturas y pesos quedan asociados:



Ejemplo 7.5 Del conjunto de datos MUNDODES, se ha tomado la esperanza de vida de hombres y de mujeres. Se han formado 4 categorías tanto para la mujer como para el hombre. Se denotan por $M1$ y $H1$, a las esperanzas entre menos de 41 a 50 años, $M2$ y $H2$, de 51 a 60 años, $M3$ y $H3$, de 61 a 70, y $M4$ y $H4$, para entre 71 a más de 80. La siguiente tabla de contingencia muestra las frecuencias de cada grupo:

Mujer/Hombre	H1	H2	H3	H4
M1	10	0	0	0
M2	7	12	0	0
M3	0	5	15	0
M4	0	0	23	19

Realizar proyecciones por filas, por columnas y conjunta de filas y columnas. Comprobar que en la proyección por filas las categorías están claramente separadas y que en el caso del hombre, las dos últimas categorías están muy cercanas. Comprobar en la proyección conjunta la cercanía de las categorías $H3$ con $M3$ y $M4$.

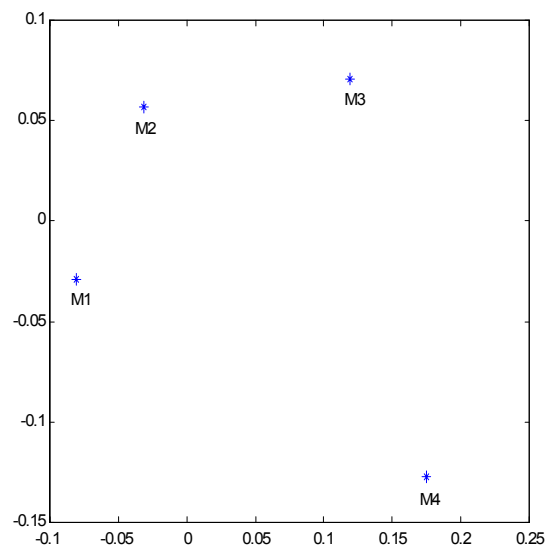
Para la proyección por filas, la variable Y queda:

$$Y = RD_c^{-\frac{1}{2}} = \begin{bmatrix} 0.2425 & 0 & 0 & 0 \\ 0.0894 & 0.1532 & 0 & 0 \\ 0 & 0.0606 & 0.1217 & 0 \\ 0 & 0 & 0.0888 & 0.1038 \end{bmatrix}$$

Los tres valores propios y vectores propios diferentes de uno de esta matriz son:

valor propio	vector propio			
0.8678	0.7221	0.3551	-0.4343	-0.4048
0.3585	-0.5249	0.7699	0.0856	-0.3528
0.1129	-0.1274	0.3072	-0.6217	0.7091

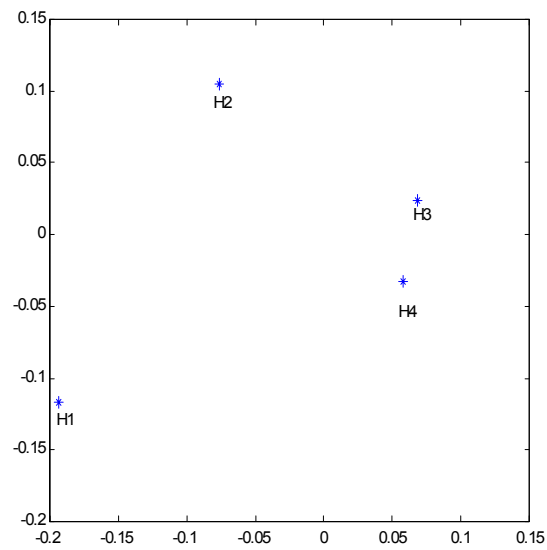
La proyección por filas es:



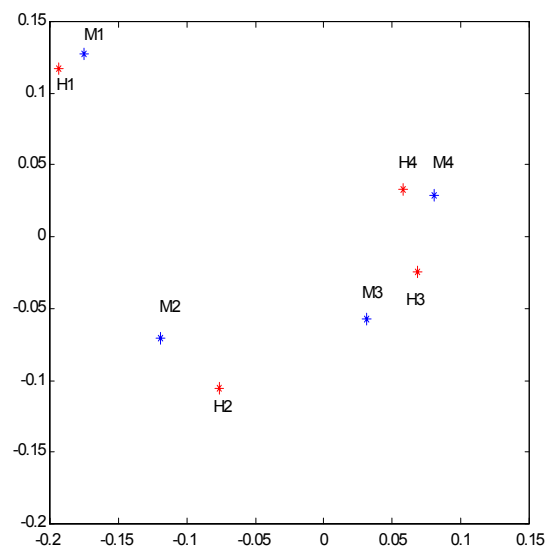
Para las columnas, los tres valores propios y vectores propios diferentes de uno de esta matriz son:

valor propio	vector propio			
0.8678	-0.5945	-0.5564	0.1503	0.5606
0.3585	-0.6723	0.5172	0.4265	-0.3141
0.1129	-0.2908	0.4628	-0.7588	0.3543

La proyección por columnas es:



El resultado de la proyección conjunta es:



7.4 ASIGNACIÓN DE PUNTUACIONES

El análisis de correspondencias puede aplicarse también para resolver el siguiente problema. Supongamos que se desea asignar valores numéricos $y_c(1), \dots, y_c(J)$ a las columnas de una matriz \mathbf{F} de observaciones, o, en otros términos, convertir la variable en columnas en una

variable numérica. Por ejemplo, en la tabla (7.3) el color del cabello puede considerarse una variable continua y es interesante cuantificar las clases de color definidas. Una asignación de valores numéricos a las columnas de la tabla inducirá automáticamente unos valores numéricos para las categorías de la variable en filas. En efecto, podemos asociar a la fila i el promedio de la variable y_c en esa fila, dado por:

$$y_i = \frac{\sum_{j=1}^J f_{ij} y_c(j)}{\sum_{j=1}^J f_{ij}} = \sum_{j=1}^J r_{ij} y_c(j) \quad (7.18)$$

donde $r_{ij} = f_{ij}/f_i$ es la frecuencia relativa condicionada a la fila. El vector de valores así obtenido para todas las filas será un vector $I \times 1$ dado por:

$$\mathbf{y}_f = \mathbf{R} \mathbf{y}_c = \mathbf{D}_r^{-1} \mathbf{F} \mathbf{y}_c \quad (7.19)$$

Análogamente, dadas unas puntuaciones \mathbf{y}_f para las filas, las puntuaciones de las columnas pueden estimarse igualmente por sus valores medios en cada columna, obteniendo el vector $J \times 1$:

$$\mathbf{y}_c = \mathbf{D}_c^{-1} \mathbf{F}' \mathbf{y}_f \quad (7.20)$$

Escribiendo conjuntamente (7.19) y (7.20) resultan las ecuaciones:

$$\mathbf{y}_f = \mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1} \mathbf{F}' \mathbf{y}_f \quad (7.21)$$

$$\mathbf{y}_c = \mathbf{D}_c^{-1} \mathbf{F}' \mathbf{D}_f^{-1} \mathbf{F} \mathbf{y}_c \quad (7.22)$$

que indican que las puntuaciones \mathbf{y}_f y \mathbf{y}_c se obtienen como vectores propios de estas matrices. Observemos que estas puntuaciones admiten una solución trivial tomando $\mathbf{y}_c = (1, \dots, 1)'_J$, $\mathbf{y}_f = (1, \dots, 1)'_I$. En efecto, las matrices $\mathbf{D}_c^{-1} \mathbf{F}'$ y $\mathbf{D}_f^{-1} \mathbf{F}$ suman uno por filas, ya que son de frecuencias relativas. Esta solución equivale en (7.21) y (7.22) al valor propio 1 de la correspondiente matriz. Para encontrar una solución no trivial al problema, vamos a exigir que ambas ecuaciones se cumplan aproximadamente introduciendo un coeficiente de proporcionalidad, $\lambda < 1$, pero que queremos sea tan próximo a uno como sea posible. Multiplicando (7.19) por $\mathbf{D}_f^{1/2}$ y (7.20) por $\mathbf{D}_c^{1/2}$ e introduciendo este coeficiente de proporcionalidad tenemos que

$$\lambda(\mathbf{D}_f^{1/2} \mathbf{y}_f) = \mathbf{D}_f^{-1/2} \mathbf{F} \mathbf{D}_c^{-1/2} (\mathbf{D}_c^{1/2} \mathbf{y}_c) \quad (7.23)$$

$$\lambda(\mathbf{D}_c^{1/2} \mathbf{y}_c) = \mathbf{D}_c^{-1/2} \mathbf{F}' \mathbf{D}_f^{-1/2} (\mathbf{D}_f^{1/2} \mathbf{y}_f) \quad (7.24)$$

Para resolver estas ecuaciones, llamemos $\mathbf{b} = \mathbf{D}_f^{1/2} \mathbf{y}_f$, $\mathbf{a} = \mathbf{D}_c^{1/2} \mathbf{y}_c$ y $\mathbf{Z} = \mathbf{D}_f^{-1/2} \mathbf{F} \mathbf{D}_c^{-1/2}$. Sustituyendo estas definiciones en (7.23) y (7.24), obtenemos $\lambda \mathbf{b} = \mathbf{Z} \mathbf{a}$ y $\lambda \mathbf{a} = \mathbf{Z}' \mathbf{b}$ y sustituyendo una de estas ecuaciones en la otra se obtiene

$$\lambda^2 \mathbf{b} = \mathbf{Z} \mathbf{Z}' \mathbf{b} \quad (7.25)$$

$$\lambda^2 \mathbf{a} = \mathbf{Z}' \mathbf{Z} \mathbf{a} \quad (7.26)$$

Estas ecuaciones muestran que \mathbf{b} y \mathbf{a} son vectores propios ligados al valor propio λ^2 de las matrices $\mathbf{Z} \mathbf{Z}'$ y $\mathbf{Z}' \mathbf{Z}$. Los vectores de puntuaciones se obtendrán después a partir de la definición de $\mathbf{b} = \mathbf{D}_f^{1/2} \mathbf{y}_f$, con lo que resulta:

$$\mathbf{y}_f = \mathbf{D}_f^{-1/2} \mathbf{b} \quad (7.27)$$

y como $\mathbf{a} = \mathbf{D}_c^{1/2} \mathbf{y}_c$,

$$\mathbf{y}_c = \mathbf{D}_c^{-1/2} \mathbf{a} \quad (7.28)$$

Las matrices $\mathbf{Z} \mathbf{Z}'$ o $\mathbf{Z}' \mathbf{Z}$ siempre admite el valor propio 1 ligado a un vector propio $(1, \dots, 1)'$. Tomando como \mathbf{a} y \mathbf{b} los vectores propios ligados al segundo mayor valor propio, $\lambda < 1$, de estas matrices obtenemos las puntuaciones óptimas de filas y columnas.

Podemos obtener una representación gráfica de las filas y columnas de la matriz de la forma siguiente: si sustituimos las puntuaciones \mathbf{y}_c dadas por (7.28), que se denominan a veces "factores" asociados a las columnas, en la ecuación (7.19) y escribimos

$$\mathbf{y}_f(\mathbf{a}) = \mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1/2} \mathbf{a}$$

obtenemos las proyecciones de las filas encontradas en (7.12). Análogamente, sustituyendo los "factores" \mathbf{y}_f asociados a las filas en (7.20) y escribiendo

$$\mathbf{y}_c(\mathbf{b}) = \mathbf{D}_c^{-1} \mathbf{F}' \mathbf{D}_f^{-1/2} \mathbf{b}$$

encontramos las proyecciones de las columnas de (7.14).

Concluimos que el problema de asignar puntuaciones de una forma consistente a las filas y a las columnas de una tabla de contingencia, es equivalente al problema de encontrar una representación óptima en una dimensión de las filas y las columnas de la matriz. En otros términos, el análisis de correspondencia proporciona en la primera coordenada de las filas y columnas una forma consistente de asignar puntuaciones numéricas a las filas y a las columnas de la tabla de contingencia.

Ejemplo 7.6 La tabla adjunta indica las puntuaciones alta (A), media (M) y baja (B) obtenidas por 4 profesores P_1, \dots, P_4 , que han sido evaluados por un total de 49 estudiantes. ¿Qué puntuaciones habría que asignar a las categorías alta, media y baja? ¿y a los profesores?

	A	M	B	
P_1	2	6	2	10
P_2	4	4	4	12
P_3	1	10	4	15
P_4	7	5	0	12
	14	25	10	49

Entonces la matriz $\mathbf{Z} = \mathbf{D}_f^{-1/2} \mathbf{F} \mathbf{D}_c^{-1/2}$ es

$$\mathbf{Z} = \begin{bmatrix} .169 & .380 & .200 \\ .309 & .230 & .365 \\ .069 & .516 & .327 \\ .540 & .288 & 0 \end{bmatrix}$$

Vamos a obtener la descomposición en valores singulares de esta matriz. Es :

$$\mathbf{Z} = \begin{bmatrix} .452 & .166 & -.249 \\ .495 & -.004 & .869 \\ .553 & .581 & -.317 \\ .495 & -.797 & -.288 \end{bmatrix} \begin{bmatrix} 1 & & \\ & .45 & \\ & & .22 \end{bmatrix} \begin{bmatrix} .534 & -.816 & .221 \\ .714 & .296 & -.634 \\ .452 & .497 & .741 \end{bmatrix}$$

que conduce a las variables

$$\mathbf{y} = \mathbf{D}_f^{-1/2} \mathbf{b}_i = \begin{bmatrix} .143 & .052 & -.079 \\ .143 & -.001 & .251 \\ .143 & .150 & -.082 \\ .143 & -.230 & -.083 \end{bmatrix}$$

$$\mathbf{z} = \mathbf{D}_c^{-1/2} \mathbf{a} = \begin{bmatrix} .143 & -.218 & .059 \\ .143 & .059 & -.127 \\ .143 & .157 & .234 \end{bmatrix}$$

La mejor puntuación -en el sentido de la máxima discriminación- corresponde a (multiplicando por -1 el segundo vector propio para que los números más altos correspondan a puntuaciones altas y favorecer la interpretación) 218, -059, -157 y a los profesores (multiplicando por -1 el segundo vector propio, para ser consistentes con el cambio anterior) 230 -150 001 -052. Si queremos trasladar estas puntuaciones a una escala entre cero y diez, escribiremos

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \times 10$$

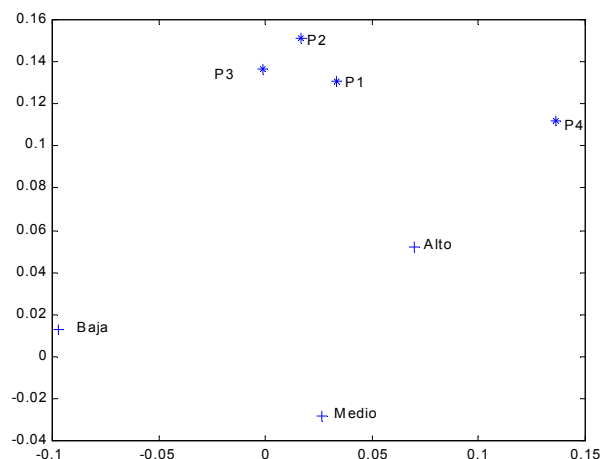


Figura 7.5: Proyección de los profesores y de las puntuaciones

y las puntuaciones se convierten en 10, 7.4 y 0 y. Las evaluaciones de los profesores al pasarlas a la escala de cero a diez se convierten en 10, 0, 3.98, 2.57. La figura 7.5 presenta la proyección de los profesores y de las categorías sobre el plano de mejor representación.

Ejemplo 7.7 La tabla de contingencia siguiente indica las puntuaciones, muy buena (MB), buena (B), regular (R) o mala (M) obtenidas por las 5 películas nominadas a los Oscars a la mejor película del año 2001 que han sido evaluadas por un total de 100 críticos de cine de todo el mundo. ¿Que puntuaciones habría que asignar a las categorías? ¿y a las películas?

<i>Películas/Puntuación</i>	<i>M</i>	<i>R</i>	<i>B</i>	<i>MB</i>	
<i>P1</i>	1	7	2	10	20
<i>P2</i>	0	3	2	15	20
<i>P3</i>	2	7	2	9	20
<i>P4</i>	0	1	3	16	20
<i>P5</i>	1	3	3	13	20
	4	21	12	63	100

La matriz P es:

$$\begin{bmatrix} 0.1118 & 0.3416 & 0.1291 & 0.2817 \\ 0 & 0.1464 & 0.1291 & 0.4226 \\ 0.2236 & 0.3416 & 0.1291 & 0.2535 \\ 0 & 0.0488 & 0.1936 & 0.4507 \\ 0.1118 & 0.1464 & 0.1936 & 0.3662 \end{bmatrix}$$

Las variables que se obtienen son:

$$y = Dr^{-\frac{1}{2}}b = \begin{bmatrix} 0.1000 & -0.0934 & -0.1124 & 0.1365 & 0.0000 \\ 0.1000 & 0.0721 & -0.1208 & -0.1234 & 0.0707 \\ 0.1000 & -0.1356 & 0.0707 & -0.1078 & -0.0707 \\ 0.1000 & 0.1304 & 0.0334 & 0.0435 & -0.1414 \\ 0.1000 & 0.0266 & 0.1291 & 0.0512 & 0.1414 \end{bmatrix}$$

$$z = Dc^{-\frac{1}{2}}a = \begin{bmatrix} 0.1000 & -0.2382 & 0.3739 & -0.2085 \\ 0.1000 & -0.1580 & -0.1053 & 0.0396 \\ 0.1000 & 0.0369 & 0.1282 & 0.2357 \\ 0.1000 & 0.0608 & -0.0130 & -0.0448 \end{bmatrix}$$

La mejor puntuación para las categorías corresponde a -0.2382, -0.1580, 0.0369 y 0.0608. Para las películas (multiplicando por -1 el segundo vector propio) a -0.0934, 0.0721, -0.1356, 0.1304 y 0.0266. Si trasladamos todas las puntuaciones entre cero y diez, obtenemos para las categorías los valores 0, 2.6823, 9.2007 y 10. Para las cinco películas tenemos 1.5864, 7.8082, 0, 10 y 6.0977. La proyección conjunta muestra como la película más cercana a la puntuación muy buena (MB) es P4:

7.5 Lecturas complementarias

El análisis de correspondencias puede extenderse para estudiar tablas de cualquier dimensión con el nombre de análisis de correspondencias múltiple. En este enfoque se utiliza la descomposición en valores singulares para aproximar simultáneamente todas las tablas bidimensionales que pueden obtenerse de una tabla multidimensional. Una buena introducción

desde el punto de vista de componentes principales con la métrica ji-cuadrado se encuentra en Gower y Hand (1995). Presentaciones de esta técnica como extensión del análisis de correspondencias presentado en este capítulo se encuentran en Greenacre (1984) y Lebart et al (1984). La literatura sobre análisis de correspondencias está sobre todo en francés, véase Lebart et al (1997) y Saporta (1990). En español Cuadras (1990) y Escofier y Pages (1990). Jackson (1991) contiene una sucinta descripción del método con bastantes referencias históricas y actuales. Lebart, Salem y Bécue (2000) presenta interesantes aplicaciones del análisis de correspondencias para el estudio de textos.

Ejercicios 7

7.1 Demostrar que la traza de las matrices $\mathbf{Z}'\mathbf{Z}$ y $\mathbf{Z}\mathbf{Z}'$ es la misma.

7.2 Demostrar que el centro de los vectores \mathbf{r}_i de las filas, donde cada fila tiene un peso \mathbf{f} es el vector \mathbf{c} de las frecuencias relativas de las columnas (calcule $\bar{\mathbf{r}} = \sum f_i \mathbf{r}_i = \mathbf{R}'\mathbf{D}_f \mathbf{1}$)

7.3 Demostrar que dada una matriz de datos \mathbf{X} donde cada fila tiene un peso \mathbf{W} la operación que convierte a esta matriz en otra de media cero es $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{W})\mathbf{X}$.

7.4 Demostrar que la suma de las distancias de Mahalanobis ponderadas de las filas es igual a la de las columnas, donde la suma de las filas es $\sum f_i (\mathbf{r}_i - \mathbf{c})' \mathbf{D}_c^{-1} (\mathbf{r}_i - \mathbf{c})$.

7.5 Supongamos que estudiamos dos características en los elementos de un conjunto que pueden darse en los niveles alto, medio y bajo en ambos casos. Si las frecuencias relativas con las que aparecen estos niveles son las mismas para las dos características, indicar la expresión de la representación de las filas y columnas en el plano bidimensional.

7.6 En el ejemplo 7.5 ¿qué podemos decir de la puntuación óptima para cuantificar las filas y columnas?

7.7 Indicar cómo afecta a la representación de filas y columnas que la tabla de contingencias sea simétrica, es decir, $f_{ij} = f_{ji}$.

7.8 Justificar que la variable $\frac{(f_{ij} - r_i c_j / n)}{\sqrt{r_i c_j / n}}$ es aproximadamente una variable normal estándar.

7.9 Demostrar que si definimos la matriz \mathbf{X} con elemento genérico $x_{ij} = (f_{ij} - f_{i.} f_{.j}) / \sqrt{f_{i.} f_{.j}}$ la matriz $\mathbf{X}'\mathbf{X}$ tiene los mismos valores propios que la $\mathbf{Z}'\mathbf{Z}$, donde $z_{ij} = f_{ij} / \sqrt{f_{i.} f_{.j}}$ salvo el valor propio 1 que aparece en $\mathbf{Z}'\mathbf{Z}$, y no en $\mathbf{X}'\mathbf{X}$.