Definición general del análisis de correspondencia y su interpretación

Definición general: Técnica descriptiva que resume gráficamente la información contenida en una tabla de contingencia.

Interpretación:

1. Una manera de representar la relación entre dos variables categóricas en un espacio de dimensión menor, análogo a componentes principales, pero definiendo la distancia entre los puntos de manera coherente con la interpretación de los datos y en lugar de usar la distancia Euclidiana usamos la **distancia ji-cuadrada**
    - Desde este enfoque, el análisis de correspondencia es el equivalente de componentes principales para datos cualitativos.
2. Es un procedimiento objetivo de asignar valores numéricos a variables cualitativas, lo cual tiene mas relación con MDS.

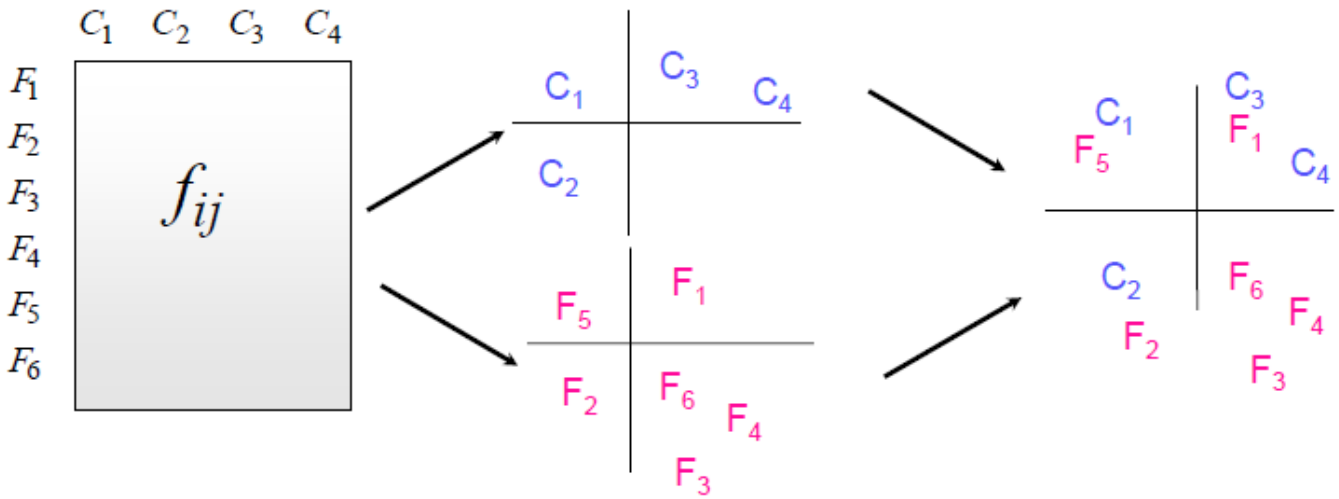
Tabla de contingencia

Fue desarrollado por Benzecri (1973) para descubrir y entender la estructura y el modelo de los datos en tablas de contingencias

Tabla de contingencia con *I* filas y *J* columnas

X2	*1*	.....	*j*	.....	*J*	*Total*
X1						
1 *2* . . *i* . *I*	x_{11} x_{21} . . . . x_{I1}		x_{1j} x_{ij}		x_{1J} x_{IJ}	$x_{1.}$ $x_{I.}$
Total	$x._1$		$x._j$		$x._J$	$x..$

Análisis de Correspondencias



Objetivo: Representar las filas y las columnas de una tabla de contingencia de dos vías como puntos en un espacio vectorial de baja dimensión, de forma que los correspondientes espacios se puedan superponer para obtener una representación conjunta

Análisis de correspondencia

- Si *puntos filas* están muy cercanos indica que las filas tienen un perfil similar a través de las columnas (su distribución está condicionada a las columnas)
- Si *puntos columna* están muy cercanos indica que las columnas tienen un perfil similar a través de las filas (su distribución está condicionada a las filas)
- Puntos filas que son muy cercanos a los puntos columnas representan combinaciones que ocurren mas frecuentemente de las que podría esperarse de un modelo de independencia.
- Esto es, un modelo en el cual las I categorías de la variable $\boldsymbol{X}_1$ (filas) están incorrelacionadas con las J categorías de la variable $\boldsymbol{X}_2$ (columnas).

Análisis de correspondencia

Los resultados de un análisis de correspondencia incluyen obtener la "mejor" representación de los datos en dos dimensiones, a través de las coordenadas de los puntos graficados y una medida llamada *inercia* que cuantifica la cantidad de información retenida en cada dimensión.

Análisis de correspondencia. Desarrollo algebraico

- Sea $\boldsymbol{X}$, una tabla a dos vías $I \times J$, cuyos elementos x_{ij} representan las frecuencias de aparición (o conteos) de la i-esima categoría de $\boldsymbol{X}_1$ y la j-esima categoría de $\boldsymbol{X}_2$
- Si n es el total de frecuencias en la tabla de datos $\boldsymbol{X}$, primero se construye una matriz $\mathbf{F} = \{f_{ij}\}$ de frecuencias relativas, dividiendo cada elemento de $\boldsymbol{X}$ por n. Entonces

$$f_{ij} = \frac{x_{ij}}{n},\ i = 1,2,\ldots,I \quad j = 1,2,\ldots,J, \quad \boldsymbol{F} = \frac{1}{n}\boldsymbol{X}$$

La matriz $\boldsymbol{F}$ es llamada la matriz de correspondencia y satisface

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

Análisis de correspondencia. Desarrollo algebraico

- El análisis de la matriz $\boldsymbol{F}$ se puede hacer por filas o columnas, cualquier análisis de esta matriz debe ser equivalente al aplicado a su transpuesta.
- Se presentará el análisis por filas de la matriz $\boldsymbol{F}$ que será simétrico al análisis por columnas.

Proyección de las filas de la matriz $\boldsymbol{F}$

- Las I filas de $\boldsymbol{F}$ pueden tomarse como I puntos en el espacio $\mathbb{R}^J$
- Se busca una representación de estos I puntos en un espacio de dimensión menor que permita apreciar sus distancias relativas
- El objetivo es el mismo que en componentes principales pero tomando en cuenta la peculiaridad de los datos

Análisis de correspondencia. Desarrollo algebraico

La peculiaridad de los datos proviene del hecho de que la frecuencia relativa de cada fila es distinta, lo que implica que:

1. Todas las filas (puntos en $\mathbb{R}^j$) no tienen el mismo peso, ya que algunas contienen mas datos que otras. Al representar el conjunto de las filas (puntos) debemos dar más peso aquellas filas que contienen mas datos.
2. La distancia euclidiana entre los puntos no es una buena medida de su proximidad, entonces debe modificarse.

Denotamos por r_i a la **frecuencia relativa** o **masa** de la fila i de $\boldsymbol{F}$, como

$$r_i = \sum_{j=1}^{J} f_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}, i = 1, 2, \ldots, I, \qquad \underset{(I\times 1)}{\boldsymbol{r}} = \underset{(I\times J(J\times 1)}{\boldsymbol{F}} \underset{}{\mathbf{1}}$$

Análisis de correspondencia. Desarrollo algebraico

- Debemos dar a cada fila un peso proporcional a su **frecuencia relativa o masa**.
- Los elementos del vector $\boldsymbol{r}$ se toman como sus pesos, ya que son números positivos que suman 1
- Denotamos por $\boldsymbol{R}$ a la matriz de frecuencias relativas condicionadas al total de la fila, llamada *matriz de perfiles*, que se obtiene como

$$\boldsymbol{R} = \boldsymbol{D}_r^{-1}\boldsymbol{F}, \text{ donde } \boldsymbol{D}_r = diag(r_1, r_2, \ldots, r_I)$$

  Entonces se transforma la matriz original $\boldsymbol{F}$ en otra matriz cuyas casillas por fila suman uno.
- Cada fila de $\boldsymbol{R}$, $\boldsymbol{R}_i$ puede considerarse un punto o un vector en el espacio $\mathbb{R}^J$. Como la suma de los componentes de $\boldsymbol{R}_i$ es uno, todos los puntos están en un espacio de dimensión $\boldsymbol{J-1}$.

Análisis de correspondencia. Desarrollo algebraico

- La idea es proyectar estos puntos en un espacio de dimensión menor de tal forma que las filas que tengan la misma estructura estén lo más próximas y las que tengan una estructura muy diferente estén alejadas.
- Por tanto debemos definir una medida de distancia entre dos filas $\boldsymbol{R}_a$ y $\boldsymbol{R}_b$. Una posibilidad es utilizar la distancia euclidiana
- La distancia euclidiana, tiene el inconveniente de tratar igual a todos los componentes de estos vectores.
- Sin embargo en este caso, el atributo j tiene diferente peso en cada fila, y por tanto la distancia euclidiana no será adecuada
- Para obtener comparaciones razonables debemos tener en cuenta la frecuencia relativa de aparición del atributo que estudiamos

Análisis de correspondencia. Desarrollo algebraico

- 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

$$(R_{aj} - R_{bj})^2 = (f_{aj}/r_a - f_{bj}/r_b)^2$$

se considera la suma de los términos

$$(R_{aj} - R_{bj})^2/c_j$$

donde

$$c_j = \sum_{i=1}^{I} f_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}, j = 1, 2, \ldots, J, \qquad \underset{(Jx1)}{\boldsymbol{c}} = \underset{(JxI)}{\boldsymbol{F}'} \underset{(Ix1)}{\mathbf{1}}$$

c_j representa la *frecuencia relativa o masa* de la columna j de $\boldsymbol{F}$

Análisis de correspondencia. Desarrollo algebraico

La expresión de la distancia entre dos filas $\boldsymbol{R}_a$ y $\boldsymbol{R}_b$ de $\boldsymbol{R}$ está dada por

$$\boldsymbol{D}^2(\boldsymbol{R}_a, \boldsymbol{R}_b) = \sum_{j=1}^{J} \left(\frac{f_{aj}}{r_a} - \frac{f_{bj}}{r_b} \right)^2 \frac{1}{c_j} = \sum_{j=1}^{J} \frac{(R_{aj} - R_{bj})^2}{c_j}$$

que en forma matricial se puede escribir como

$$\boldsymbol{D}^2(\boldsymbol{R}_a, \boldsymbol{R}_b) = (\boldsymbol{R}_a - \boldsymbol{R}_b)' \boldsymbol{D}_c^{-1} (\boldsymbol{R}_a - \boldsymbol{R}_b)$$

donde

$$\boldsymbol{D}_c = diag(c_1, c_2, \ldots, c_j)$$

- La distancia $\boldsymbol{D}^2(\boldsymbol{R}_a, \boldsymbol{R}_b)$ es la distancia χ^2 y equivale a la distancia euclidiana entre los vectores transformados

$$\boldsymbol{y}_i = \boldsymbol{D}_c^{-1/2} \boldsymbol{R}_i$$

Análisis de correspondencia. Desarrollo algebraico

- Por lo anterior, el problema se simplifica definiendo una matriz de datos transformados, sobre la que tiene sentido considerar la distancia euclidiana entre filas.
- Definiendo

$$\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{D}_c^{-1/2} = \boldsymbol{D}_r^{-1}\boldsymbol{F}\boldsymbol{D}_c^{-1/2}$$

Obtenemos una matriz $\boldsymbol{Y}$ cuyas entradas tienen la forma

$$y_{ij} = \left\{ \frac{f_{ij}}{r_i c_j^{1/2}} \right\}$$

Los elementos de $\boldsymbol{Y}$ ya no suman **uno**, ni por renglón ni por columna

- Representan las frecuencias relativas condicionadas por fila, pero estandarizadas por su variabilidad, que depende de la raíz cuadrada de la frecuencia relativa de la columna

Análisis de correspondencia. Desarrollo algebraico

- De esta manera, las entradas de $\boldsymbol{Y}$ son directamente comparables entre si.
- La matriz de datos transformada $\boldsymbol{Y}$ se puede tratar como una matriz de datos estándar, donde las filas representan a las observaciones y las columnas a las variables.
- La pregunta es 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.
- Lo anterior implica encontrar un vector $\boldsymbol{a}$ de norma la unidad, esto es ,$\boldsymbol{a'a} = \mathbf{1}$ tal que el vector de puntos proyectados sobre $\boldsymbol{a}$,

$$\boldsymbol{y}_p(\boldsymbol{a}) = \boldsymbol{Ya}$$

Tenga variabilidad máxima

Análisis de correspondencia. Desarrollo algebraico

- El vector $\boldsymbol{a}$ se encuentra maximizando

$$Var(\boldsymbol{y}_p(\boldsymbol{a})) = \boldsymbol{y}_p(\boldsymbol{a})' \boldsymbol{y}_p(\boldsymbol{a}) = \boldsymbol{a}' \boldsymbol{Y}' \boldsymbol{Y} \boldsymbol{a}$$

  Con la condición de que $\boldsymbol{a}'\boldsymbol{a} = 1$.
- El problema se convierte en un problema de componentes principales: el vector $\boldsymbol{a}$ es vector propio de la matriz $\boldsymbol{Y}'\boldsymbol{Y}$.
- Sin embargo, este tratamiento de la matriz $\boldsymbol{Y}$ como una matriz de variables continuas no es del todo correcto.
- En la matriz $\boldsymbol{Y}$, las filas tienen una frecuencia relativa distinta r_i y por tanto deben tener distinto peso.
- Las filas con mayor frecuencia relativa deben tener mas peso en la representación que aquellas con frecuencia relativa muy baja

Análisis de correspondencia. Desarrollo algebraico

- De esta manera, las filas con gran número de individuos deben estar bien representadas, aunque esto sea a costa de representar peor las filas con pocos elementos.
- En consecuencia, se dará a cada fila de $\boldsymbol{Y}$ un peso proporcional al número de datos que contiene, es decir $\boldsymbol{D_r^{1/2}Y}$
- Lo anterior implica maximizar ahora la suma de cuadrados ponderada:

$$\boldsymbol{m = a' Y' D_r Y a} \qquad \boldsymbol{a' a = 1}$$

O equivalentemente

$$\boldsymbol{m = a' D_c^{-1/2} F' D_r^{-1} F D_c^{-1/2} a}$$

Análisis de correspondencia. Desarrollo algebraico

- Alternativamente, podemos construir una matriz de datos $\boldsymbol{Z}$ definida por

$$\boldsymbol{Z} = \boldsymbol{D_r}^{-1/2} \boldsymbol{F} \boldsymbol{D_c}^{-1/2}$$

cuyos componentes son

$$z_{ij} = \left\{ \frac{f_{ij}}{\sqrt{r_i c_j}} \right\}$$

Lo cual estandariza las frecuencias relativas en cada casilla mediante el producto raíces cuadradas de las frecuencias relativas totales de la fila y la columna.

Análisis de correspondencia. Desarrollo algebraico

- Considerando esta matriz $\boldsymbol{Z}$, el problema de encontrar el vector $\boldsymbol{a}$ se transforma al problema de maximizar

$$\boldsymbol{m} = \boldsymbol{a}' \boldsymbol{Z}' \boldsymbol{Z} \boldsymbol{a} \qquad \boldsymbol{a}' \boldsymbol{a} = 1$$

• Lo cual es el problema de componentes principales cuya solución está dada por los vectores y valores propios de $\boldsymbol{Z}'\boldsymbol{Z}$:

$$\boldsymbol{D}_c^{-1/2} \boldsymbol{F}' \boldsymbol{D}_r^{-1} \boldsymbol{F} \boldsymbol{D}_c^{-1/2} \boldsymbol{a} = \lambda \boldsymbol{a}$$

donde $\boldsymbol{a}$ debe ser un vector propio de la matriz $\boldsymbol{Z}'\boldsymbol{Z}$ y λ su valor propio asociado.

Análisis de correspondencia. Desarrollo algebraico

Vamos a comprobar que la matriz $\boldsymbol{Z'Z}$ siempre tiene como valor propio máximo el 1 y como vector propio $\boldsymbol{D_c^{1/2}1}$.
Multiplicando $\boldsymbol{D_c^{-1/2}F'D_r^{-1}FD_c^{-1/2}a} = \lambda \boldsymbol{a}$ por $\boldsymbol{D_c^{-1/2}}$ se obtiene

$$\boldsymbol{D_c^{-1}F'D_r^{-1}FD_c^{-1/2}a} = \lambda(\boldsymbol{D_c^{-1/2}a})$$

Las matrices

$$\boldsymbol{D_r^{-1}F} \quad \text{y} \quad \boldsymbol{FD_c^{-1}}$$

representan las matrices de perfiles por filas y por columnas y su suma por filas y columnas correspondiente es **uno**.
Por tanto

$$\boldsymbol{D_r^{-1}F}\underset{J\text{x}1}{\mathbf{1}} = \underset{I\text{x}1}{\mathbf{1}} \text{ y } \boldsymbol{D_c^{-1}F'}\underset{I\text{x}1}{\mathbf{1}} = \underset{J\text{x}1}{\mathbf{1}}$$

que implica que

$$\boldsymbol{D_c^{-1}F'D_r^{-1}F}\mathbf{1} = \mathbf{1}$$

Por tanto $\boldsymbol{D_c^{-1}F'D_r^{-1}F}$ tiene un valor propio de 1 asociado a un vector propio de $\mathbf{1}$

Análisis de correspondencia. Desarrollo algebraico

- Si esta expresión

$$\boldsymbol{D}_c^{-1}\boldsymbol{F}'\boldsymbol{D}_r^{-1}\boldsymbol{F}1 = \mathbf{1}$$

la multiplicamos por $\boldsymbol{D}_c^{1/2}$, obtenemos

$$\boldsymbol{D}_c^{1/2}\boldsymbol{D}_c^{-1}\boldsymbol{F}'\boldsymbol{D}_r^{-1}\boldsymbol{F}\mathbf{1} = \boldsymbol{D}_c^{1/2}\mathbf{1}$$
$$\Rightarrow \boldsymbol{D}_c^{-1/2}\boldsymbol{F}'\boldsymbol{D}_r^{-1}\boldsymbol{F}\boldsymbol{D}_c^{-1/2}(\boldsymbol{D}_c^{1/2}\mathbf{1}) = (\boldsymbol{D}_c^{1/2}\mathbf{1})$$

Es decir, $\boldsymbol{Z'Z}$ tiene un valor propio de 1 con vector propio $\boldsymbol{D}_c^{1/2}\mathbf{1}$.

- Olvidando esta solucion trivial que no proporciona información sobre la estructura de las filas, tomamos el mayor valor propio menor que 1 y su vector propio asociado $\boldsymbol{a}$.
- Entonces proyectando la matriz $\boldsymbol{Y}$ sobre la dirección encontrada dada por el vector $\boldsymbol{a}$, obtenemos el vector proyectado:

$$\boldsymbol{y}_r(\boldsymbol{a}) = \boldsymbol{Y}\boldsymbol{a} = \boldsymbol{D}_r^{-1}\boldsymbol{F}\boldsymbol{D}_c^{-1/2}\boldsymbol{a}$$

el vector $\boldsymbol{y_r(a)}$ es la mejor representación de las filas de la tabla de contingencia en un espacio de **una** dimension.

Análisis de correspondencia. Desarrollo algebraico

- De igual manera, si extraemos el vector propio asociado al siguiente valor propio mayor de $\boldsymbol{Z'Z}$, obtenemos una segunda coordenada y podemos representar las filas en un espacio de **dos** dimensiones.
- Las coordenadas de la representación de cada fila están dadas por las filas de la matriz

$$\boldsymbol{C}_r = \boldsymbol{YA}_2 = \boldsymbol{D}_r^{-1}\boldsymbol{F}\boldsymbol{D}_c^{-1/2}\boldsymbol{A}_2$$

  Donde $\boldsymbol{A}_2 = \left[\boldsymbol{a}_1 \vdots \boldsymbol{a}_2\right]$ contiene en columnas los dos vectores propios de $\boldsymbol{Z}'\boldsymbol{Z}$
- La matriz $\boldsymbol{C}_r$ es de tamaño $I \times 2$ y las dos coordenadas de cada fila proporcionan la mejor representación de las filas de la matriz $\boldsymbol{F}$ en un espacio de dos dimensiones

Análisis de correspondencia. Desarrollo algebraico

- El procedimiento se extiende sin dificultad para obtener representaciones en más dimensiones, calculando vectores propios adicionales de la matriz $\boldsymbol{Z'Z}$.
- 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 (Obtención de $\boldsymbol{R}$)
  2. Definir la distancia entre los puntos mediante la distancia χ^2, que tiene en cuenta que cada coordenada de las filas tiene distinta precisión (Obtención de $\boldsymbol{Y}$)
  3. Proyectar los puntos de $\boldsymbol{Y}$ sobre las direcciones de máxima variabilidad, teniendo en cuenta que cada fila tiene un peso distinto e igual a sus frecuencias relativas. (Obtención de $\boldsymbol{Z}$)

Análisis de correspondencia. Desarrollo algebraico

- El procedimiento operativo para obtener la mejor representación bidimensional de las filas de la tabla de contingencia es:
  1. Calcular la matriz $\boldsymbol{Z'Z}$ y obtener sus valores y vectores propios.
  2. Tomar los dos vectores propios $\boldsymbol{a}_1$ y $\boldsymbol{a}_2$, asociados a los valores propios mayores (menores a 1) de esta matriz.
  3. Calcular las proyecciones $\boldsymbol{D_r^{-1}FD_c^{-1/2}a_i}$, $i = 1,2$ y representarlas gráficamente en un espacio bidimensional.

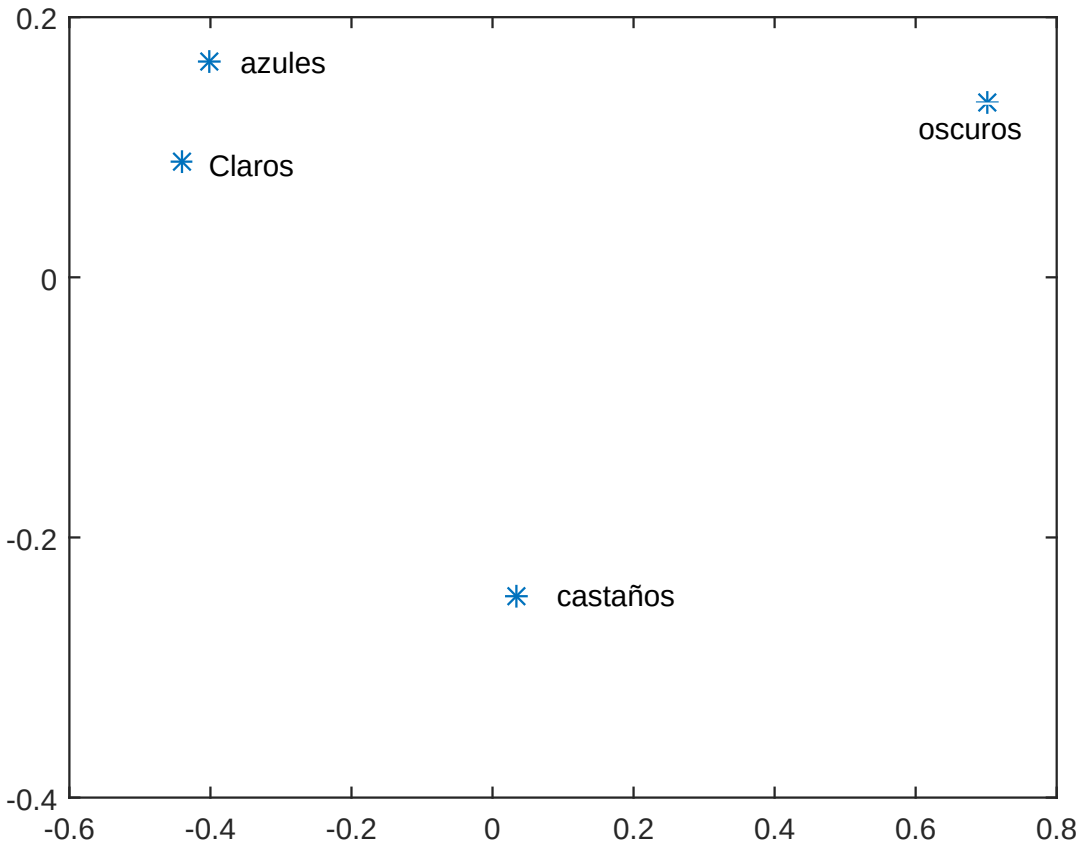
Ejemplo: Análisis de correspondencia por fila

Aplicar un analisis de correspondencia por fila a la siguiente tabla de contingencia

	Color de cabello					
Color de 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	681	85	**1315**
total	**1455**	**286**	**2137**	**1391**	**118**	**5387**

- La tabla presenta la clasificacion de $n = 5387$ estudiantes escoceses por el color de sus ojos, que tiene 4 categorías posibles ($I = 4$), y por el color de su cabello, que tiene 5 categorías posibles ($J = 5$).
- La tabla tiene interés histórico debido a que fue utilizada por Fisher (1940) para ilustrar un método de análisis de tablas de contingencias relacionado al análisis de correspondencia

Ejemplo: Análisis de correspondencia por fila



Proyección de las columnas de F

- El análisis anterior para las filas de $\boldsymbol{F}$ se puede aplicar también para las columnas. Las J columnas de $\boldsymbol{F}$ son puntos en $\mathbb{R}^I$
- Se busca una representación de estos J puntos en un espacio de dimensión menor que I que permita apreciar sus distancias relativas.
- Para lo cual construimos las frecuencias relativas por columna

$$c_j = \sum_{i=1}^{I} f_{ij} = \sum_{i=1}^{n} \frac{x_{ij}}{n}, j = 1, 2, \ldots, J \qquad \underset{(J\text{x}1)}{\boldsymbol{c}} = \underset{(J\text{x}I)}{\boldsymbol{F}}{}' \underset{I\text{x}1}{\mathbf{1}}$$

y

$$\boldsymbol{D}_c = diag(c_1, c_2, \ldots, c_J)$$

- La mejor representación de los J puntos (columnas) en un espacio de dimensión menor, considerando la distancia χ^2 conducirá, por simetría a estudiar la matriz equivalente

$$\boldsymbol{Y} = \boldsymbol{D}_c^{-1} \boldsymbol{F}' \boldsymbol{D}_r^{-1/2}$$

Proyección de las columnas de F

- Ahora la matriz que contiene las frecuencias relativas de las filas de $\boldsymbol{F}$' es $\boldsymbol{D_c}$ y la matriz que contiene las frecuencias relativas de las columnas de $\boldsymbol{F}$' es $\boldsymbol{D_r}$.
- Intercambiando el papel de las matrices, las direcciones de proyección son los vectores propios de la matriz

$$\boldsymbol{ZZ' = D_r^{-1/2} F D_c^{-1} F' D_r^{-1/2}}$$

- $\boldsymbol{Z'Z}$ y $\boldsymbol{ZZ'}$ tienen los mismos valores propios no nulos, entonces $\boldsymbol{ZZ'}$ también tendrá un valor propio de 1 asociado a un vector propio que es una solución trivial.
- Denotando por $\boldsymbol{b}$ al vector propio asociado al mayor valor propio distinto de 1 de $\boldsymbol{ZZ'}$, entonces la mejor representación de las columnas de la matriz en un espacio de **una dimensión** está dada por

$$\boldsymbol{y_c(b) = Y' b = D_c^{-1} F' D_r^{-1/2} b}$$

Proyección de las columnas de F

- Análogamente, la mejor representación en **dos** dimensiones de las columnas de la matriz $\boldsymbol{Y}$ están dadas por las coordenadas definidas por las filas de la matriz

$$\boldsymbol{C_c = Y' B_2 = D_c^{-1} F' D_r^{-1/2} B_2}$$

donde $\boldsymbol{B_2 = \left[b_1 \vdots b_2 \right]}$ contiene en columnas los dos vectores propios asociados a los valores propios mayores de $\boldsymbol{ZZ'}$ menores a 1.

- La matriz $\boldsymbol{C}_c$ es de tamaño $J \times 2$ y cada fila es la mejor representación de las columnas de la matriz $\boldsymbol{F}$ en un espacio de **dos dimensiones**

Análisis conjunto

- Dada la simetría del problema conviene representar conjuntamente las filas y las columnas de la matriz $\boldsymbol{F}$.
- Las matrices $\boldsymbol{Z'Z}$ y $\boldsymbol{ZZ'}$ tienen los mismos valores propios no nulos
- Además, los vectores propios de ambas matrices que corresponden al mismo valor propio están relacionados.
- Sea $\boldsymbol{a}_i$ un vector propio de $\boldsymbol{Z'Z}$ asociado al valor propio λ_i :

$$\boldsymbol{Z'Za_i} = \lambda_i \boldsymbol{a}_i$$

- Multiplicando por $\boldsymbol{Z}$

$$\boldsymbol{ZZ'(Za_i)} = \lambda_i(\boldsymbol{Za_i})$$

- Entonces $\boldsymbol{b_i} = \boldsymbol{Za_i}$ es un vector propio de $\boldsymbol{ZZ'}$ asociado a λ_i

Análisis conjunto

Una forma rápida para obtener los vectores propios es calcularlos directamente de la matriz de dimensión más pequeña, $\boldsymbol{Z'Z}$ ó $\boldsymbol{ZZ'}$ y obtener los otros vectores propios como $\boldsymbol{Za_i}$ ó $\boldsymbol{Z'b_i}$.
Alternativamente, se puede utilizar la descomposición en valores singulares de la matriz $\boldsymbol{Z}$ ó $\boldsymbol{Z'}$. La descomposición aplicada a $\boldsymbol{Z}$ es

$$\boldsymbol{Z} = \boldsymbol{B}_s \boldsymbol{\Lambda}_s \boldsymbol{A}_s' = \sum_{i=1}^{s} \lambda_i^{1/2} \boldsymbol{b_i a_i}'$$

donde $\boldsymbol{B}_s$ contiene en columnas los vectores propios de $\boldsymbol{ZZ'}$, $\boldsymbol{A}_s$ los vectores propios de $\boldsymbol{Z'Z}$ y $\boldsymbol{\Lambda}_s$ es diagonal y contiene los valores singulares, $\lambda_i^{1/2}$, ó raíces de los valores propios no nulos de $\boldsymbol{ZZ'}$ ó $\boldsymbol{Z'Z}$ y donde $s = min(I, J)$.

Análisis conjunto

- Entonces, la representación de las filas en el espacio conjunto se obtiene mediante

$$y_r(a) = Ya = D_r^{-1} F D_c^{-1/2} a$$

- Y la representación de las columnas en el mismo espacio se obtiene como

$$y_c(b) = Y'b = D_c^{-1} F' D_r^{-1/2} b$$

- La representación de la matriz Z con h dimensiones (normalmente h=2) implica aproximar esta matriz mediante.

$$\hat{Z}_h = B_h \Lambda_h A'_h$$

- Esto es equivalente a obtener una aproximación a la tabla de contingencia observada F mediante la expresión

$$\hat{F}_h = D_r^{1/2} \hat{Z}_h D_c^{1/2}$$

Análisis conjunto

- Si se desea eliminar el valor propio **uno** desde el principio, dado que no aporta información de interés, se puede reemplazar la matriz $\boldsymbol{F}$ por $\boldsymbol{F} - \hat{\boldsymbol{F}}_{\boldsymbol{e}}$, donde $\hat{\boldsymbol{F}}_e$ es la matriz de frecuencias esperadas que viene dada por

$$\hat{\boldsymbol{F}}_{\boldsymbol{e}} = \frac{1}{n}\boldsymbol{r}\boldsymbol{c}^{'}$$

- Se puede probar que la matriz $\boldsymbol{F} - \hat{\boldsymbol{F}}_{\boldsymbol{e}}$ 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 respecto a la suma de todos los valores propios distintos de 1.

Análisis conjunto

El análisis de correspondencia de una tabla de contingencia de dimensiones $I \times J$ se realiza en los siguientes pasos:

1. Se calcula la tabla de frecuencias relativas $\boldsymbol{F}$.
2. Se calcula la tabla estandarizada $\boldsymbol{Z}$ de frecuencias relativas que tiene las mismas dimensiones de la tabla original, $I \times J$, y que se obtiene dividiendo cada celda de $\boldsymbol{F}$ por la raíz de los totales de su fila y columna

$$z_{ij} = \left\{ \frac{f_{ij}}{\sqrt{r_i c_j}} \right\}$$

3. Se calculan los h vectores propios (normalmente h=2) asociados a los valores propios mayores (distintos de 1), de la matriz de menor dimension de $\boldsymbol{ZZ'}$ y $\boldsymbol{Z'Z}$.

Análisis conjunto

- Si obtenemos los vectores propios $\boldsymbol{a}_i$ de $\boldsymbol{Z'Z}$, los $\boldsymbol{b}_i$ de $\boldsymbol{ZZ'}$ se obtienen por $\boldsymbol{b}_i = \boldsymbol{Z}\boldsymbol{a}_i$.
- Análogamente si se obtienen los $\boldsymbol{b}_i$ de $\boldsymbol{ZZ'}$, $\boldsymbol{a}_i = \boldsymbol{Z'}\boldsymbol{b}_i$.
- Las I filas de la matriz se representarán como I puntos en $\mathbb{R}^h$ y las coordenadas de cada fila vienen dadas por

$$\boldsymbol{C_r} = \boldsymbol{D_r}^{-1/2}\boldsymbol{Z}\boldsymbol{A_2}$$

donde $\boldsymbol{A_2}$ tiene en columnas los dos vectores propios de $\boldsymbol{Z'Z}$. Las J columnas se representarán como J puntos en $\mathbb{R}^h$ y las coordenadas de cada columna son

$$\boldsymbol{C_c} = \boldsymbol{D_c}^{-1/2}\boldsymbol{Z}'\boldsymbol{B_2}$$

donde $\boldsymbol{B_2}$ tiene en columnas los dos vectores propios de $\boldsymbol{ZZ'}$.

La distribución Ji cuadrada

- La prueba de independencia entre las variables fila y columna en una tabla de contingencia $I \times J$ se realiza con el estadístico:

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

  Que bajo la hipótesis de independencia sigue, asintóticamente, una distribución χ^2 con $(I-1)\times(J-1)$ grados de libertad
- El estadístico χ^2 para probar la independencia se puede escribir como

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(nf_{ij} - nr_i c_j)^2}{nr_i c_j}$$

  O también como

$$\chi^2 = n \sum_{i=1}^{I} r_i \sum_{j=1}^{J} \left(\frac{f_{ij}}{r_i} - c_j\right)^2 \frac{1}{c_j}$$

La distribución Ji cuadrada

- Se puede probar que esta expresión es equivalente a calcular las distancias entre los vectores de la matriz de frecuencias relativas por filas, $\boldsymbol{R}$ definida como $\boldsymbol{R} = \boldsymbol{D}_r^{-1}\boldsymbol{F}$ si medimos la distancia con la métrica χ^2
- También se puede probar que el valor medio o centroide de las filas de $\boldsymbol{R}$ está dado por el vector cuyos componentes son las frecuencias relativas de las columnas, es decir por el vector $\boldsymbol{c}$.
- De igual forma, el centroide de las columnas de $\boldsymbol{R}$ está dado por el vector $\boldsymbol{r}$ de frecuencias relativas de las filas

La distribución Ji cuadrada

- La distancia de cualquier vector fila $\boldsymbol{R_i}$ a su media $\boldsymbol{c}$, con la métrica χ^2 será

$$(\boldsymbol{R_i} - \boldsymbol{c})' \boldsymbol{D_c^{-1}} (\boldsymbol{R_i} - \boldsymbol{c})$$

  donde la matriz $\boldsymbol{D_c^{-1}}$ se obtiene como en $\boldsymbol{D^2(R_a, R_b)}$.
- La suma de todas estas distancias, ponderadas por su importancia se conoce como **inercia total de la tabla** y se define como

$$I_T = \sum_{i=1}^{I} r_i (\boldsymbol{R_i} - \boldsymbol{c})' \boldsymbol{D_c^{-1}} (\boldsymbol{R_i} - \boldsymbol{c})$$

  Lo cual se puede escribir como

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

La distribución Ji cuadrada

- Si se compara con

$$\chi^2 = n \sum_{i=1}^{I} r_i \sum_{j=1}^{J} \left(\frac{f_{ij}}{r_i} - c_j \right)^2 \frac{1}{c_j}$$

  la inercia total es igual a $\frac{\chi^2}{n}$
- También se puede demostrar que la **inercia total** es la suma de los valores propios al cuadrado de la matriz $\boldsymbol{Z'Z}$ eliminando el **uno**, es decir

$$I_T = \sum_{i=1}^{I} r_i (\boldsymbol{R}_i - \boldsymbol{c})' \boldsymbol{D}_c^{-1} (\boldsymbol{R}_i - \boldsymbol{c}) = \sum_{k=1}^{J-1} \lambda_k^2$$

- Por tanto, el análisis de las filas (o de las columnas debido a que el problema es simétrico) se puede ver como una descomposición de los componentes del estadístico χ^2 en sus fuentes de variación.

La distribución Ji cuadrada

- Esto implica también que existe una inercia para cada dimensión de la representación, definida por los valores propios λ_i
- Debido a que la inercia es una medida de la variación total de los datos de la tabla, ¿como se interpretaría un valor grande para la proporción

$$(\lambda_1^2 + \lambda_2^2) / \sum_{k=1}^{J-1} \lambda_k^2$$

?

- Geométricamente se dice que la asociación en los datos es bien representada por los puntos en un plano, y esta aproximación representa casi toda la variación en los datos más allá de lo que se podría representar con una solución de rango 1.
- Algebraicamente se puede decir que la aproximación

$$\boldsymbol{Z} \approx \boldsymbol{B}_2 \boldsymbol{\Lambda}_2 \boldsymbol{A}'_2 = \lambda_1^{1/2} \boldsymbol{b}_1 \boldsymbol{a}'_1 + \lambda_2^{1/2} \boldsymbol{b}_2 \boldsymbol{a}'_2$$

es **muy buena**