Análisis Factorial de Correspondencias

El Análisis Factorial de Correspondencias fue ideado por el estadístico francés Benzecri en 1973. Sin embargo, su incorporación a los programas estadísticos estándares fue bastante más tardía. Fusionando el Análisis de Proximidades con el Análisis de Componentes Principales, el Análisis Factorial de Correspondencias permite profundizar en el análisis de la similaridad de modalidades, de la asociación entre atributos y de las relaciones de atracción y repulsión de sus modalidades en tablas de contingencia construidas generalmente sobre variables de tipo cualitativo.

Con el fin de introducir la técnica, consideremos el siguiente ejemplo. Imaginemos que estamos analizando las preferencias de los comensales de un mesón sobre los componentes de una menú que consta de un primer plato, un segundo plato, un postre y una bebida a elegir de entre las siguientes opciones:

Primer Plato:	Ensalada, Sopa, Macarrones.
Segundo Plato:	Carne, Pescado.
Postre:	Flan, Helado, Fruta.
Bebida:	Agua, Vino, Cerveza.

Intuitivamente sería sencillo estar de acuerdo con que los clientes no eligen platos, postres y bebidas aleatoriamente sino que lo hacen, en general, en función de preferencias que pueden ser comunes en grupos amplios de personas. Así, es frecuente ver que los amantes de las comidas abundantes pueden tender a formar sus menús con platos como los macarrones y la carne acompañados de vino o cerveza; mientras que los más preocupados por su línea pueden tender más comúnmente a comer más ligeramente formando sus menús con platos como ensalada y pescado acompañados probablemente por agua.

Haciendo uso de los conocimientos adquiridos al estudiar la dependencia entre estas variables cualitativas, diríamos que existe una cierta asociación entre platos, postres y bebidas que se manifiesta en atracciones o repulsiones de sus modalidades según las conductas o comportamientos de los individuos observados. La pregunta es pues ¿cómo describir, cuantificar y explicar de forma simple esta atracción o repulsión entre las modalidades de estas variables?

El análisis de la intensidad de las atracciones y repulsiones entre las modalidades que pueden presentar las características cualitativas puede realizarse a partir del análisis de las frecuencias conjuntas observadas y recogidas en tablas de contingencias. Ello no es muy dificultoso cuando las tablas son bidimensionales. Sin embargo, a medida que necesitamos aumentar la dimensión de la tabla, lo que es frecuente en el campo del análisis multivariante, el problema se complica extraordinariamente. Los modelos logarítmico lineales para tablas de contingencia pueden ayudarnos a ello, como ya se ha visto anteriormente. Sin embargo, el Análisis Factorial de Correspondencias nos permita realizar una más profunda aproximación cuantitativa al problema expuesto, cuya naturaleza es de origen

eminentemente cualitativa, permitiéndonos no sólo la descripción y cuantificación de las relaciones entre modalidades, sino también su representación factorial, a partir de la que podremos tratar de explicar el porqué de tales relaciones.

De forma simplificada, el Análisis Factorial de Correspondencias trata de resolver el problema de explicar de forma simple como se atraen o repelen las modalidades de las variables cualitativas observadas. Así, en el ejemplo, trataría de explicar cómo la bebida se puede asociar con un tipo de primer plato concreto, o con algún tipo de segundo plato, o con una combinación de ambos; por ejemplo, trataría de ver si hay algún tipo de atracción entre el vino y la carne como segundo plato; o si la tiene con los macarrones como primer plato; o si podríamos concluir que macarrones-carne-vino constituyen una elección generalizada en cierto tipo de individuos para los que estas modalidades de primer plato, segundo plato y bebida se atraen con una cierta intensidad.

Además del análisis de atracción-repulsión entre modalidades de atributos (variables cualitativas) diferentes, la técnica del Análisis Factorial de Correspondencias también permite realizar estudios de proximidad (similaridad/disimilaridad) entre las modalidades de una misma variables; es decir, permite evaluar el parecidos de las distribuciones de casos o individuos que presentan sendas modalidades comparadas de una misma variable, proporcionándonos la información necesaria para evaluar la homogeneidad o sustituibilidad de las mismas.

Todo ello lo realiza mediante la proyección de las modalidades sobre un espacio métrico en el que aplicará el Análisis de Componentes principales para facilitar una interpretación causal simple de los comportamientos de similitud-atracción de aquéllas a través de las primeras componentes principales o factores.

Podemos resumir, pues, los objetivos de esta técnica como sigue:

- Descubrir las relaciones de atracción-repulsión existentes entre las distintas modalidades de diferentes variables cualitativas enfrentadas en una tabla de contingencia.
- Descubrir las relaciones de proximidad existentes entre las distintas modalidades de una misma variable cualitativa.
- Visualizar y Caracterizar de forma simple las relaciones anteriores en un espacio de dimensión lo más reducida posible.

Cuando el Análisis Factorial de Correspondencias se aplica al estudio de las modalidades de sólo 2 variables enfrentadas en una Tabla de Contingencia (de dimensión 2), recibe el nombre de Análisis Factorial de Correspondencias Simple. Cuando se aplica al estudio de las modalidades de más de dos variables, se conoce como Análisis Factorial de Correspondencias Múltiple.

Análisis Factorial de Correspondencias Simple

El Análisis de Correspondencias simple, como ya hemos dicho, parte de una Tabla de Contingencia de dimensión 2 en la que se enfrentan las dos variables cualitativas cuyas modalidades se van a estudiar, y que notaremos de forma genérica como sigue:

		Atributo B (modalidades)				
		B_1	B_2	...	B_q	Total
Atributo A (modalidades)	A_1	f_{11}	f_{12}	...	f_{1q}	$f_{1\cdot}$
	A_2	f_{21}	f_{22}	...	f_{2q}	$f_{2\cdot}$
	. . .	. . .	. . .	. . .	. . .	. . .
	A_p	f_{p1}	f_{p2}	...	f_{pq}	$f_{p\cdot}$
	Total	$f_{\cdot 1}$	$f_{\cdot 2}$	...	$f_{\cdot a}$	*1*

siendo f_{ij}, como de costumbre, las frecuencias relativas observadas conjuntamente para la pareja de modalidades (A_i,B_j). Lógicamente, la suma de todas esas frecuencias relativas es la unidad. Análogamente, la notación $f_{i\cdot}$ y $f_{\cdot j}$ se refiere a las correspondientes frecuencias relativas de las modalidades A_i y B_j de las distribuciones marginales de los respectivos atributos A y B.

Para simplificar algunas pruebas que realizaremos en este capítulo, realizaremos las siguientes notaciones matriciales:

$$F = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1q} \\ f_{21} & f_{22} & \cdots & f_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ f_{p1} & f_{p2} & \cdots & f_{pq} \end{pmatrix}_{p\cdot q} \qquad D_p = \begin{pmatrix} f_{1\cdot} & 0 & \cdots & 0 \\ 0 & f_{2\cdot} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f_{p\cdot} \end{pmatrix}_{p\cdot p} \qquad D_q = \begin{pmatrix} f_{\cdot 1} & 0 & \cdots & 0 \\ 0 & f_{\cdot 2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f_{\cdot q} \end{pmatrix}_{q\cdot q}$$

Las distribuciones de frecuencias relativas de las distribuciones condicionadas (fila) $B|A_i$, para i=1,2,...,p, serían:

$$\left(\frac{f_{i1}}{f_{i\cdot}}, \frac{f_{i2}}{f_{i\cdot}}, ..., \frac{f_{iq}}{f_{i\cdot}} \right), \quad i=1,....,p$$

Nótese que éstas distribuciones anteriores están informando de las distribuciones de las modalidades B_1, B_2, ..., B_q (comportamiento del atributo B) restringidas a aquellos casos en los que se presentan respectivamente las modalidades A_1 o A_2 o ,.....,A_p para el atributo A. Por tanto pueden verse como vectores que caracterizan respectivamente las p modalidades del atributo A en un cierto espacio de dimensión q.

Si dos de estas distribuciones condicionadas filas, por ejemplo $B|A_i$ y $B|A_{i'}$, son iguales (o muy parecidas), entonces la distribución de la variables B para individuos que presentaran $A=A_i$ ó $A=A_{i'}$ no variaría (o variaría poco); es decir, no dependería (o dependería poco) de la modalidad observada para la variable A y, por tanto, estas modalidades A_i y $A_{i'}$ que consideramos están induciendo el mismo efecto en la distribución de B. En este sentido estas modalidades A_i y $A_{i'}$ son similares (o parecidas) y podríamos decir que tienen un efecto similar sobre B.

En definitiva, estamos representando las distintas modalidades del atributo A como puntos de un espacio *q*-dimensional caracterizados por las coordenadas que inducen las respectivas distribuciones de frecuencias relativas condicionadas del atributo B a cada modalidad de A.

Análogamente, las distribuciones de frecuencias relativas de las distribuciones condicionadas (columna) $A|B_j$, para j=1,2,...,q, serían:

$$\left(\frac{f_{1j}}{f_{\cdot j}}, \frac{f_{2j}}{f_{\cdot j}}, ..., \frac{f_{pj}}{f_{\cdot j}} \right), \quad j = 1,....,q$$

y están informando de las distribuciones de las modalidades A_1, A_2, ..., A_p (comportamiento del atributo A) restringidas a aquellos casos en los que se presentan respectivamente las modalidades B_1 o B_2 o ,.....,B_q para el atributo B. Por tanto pueden verse como vectores que caracterizan respectivamente las q modalidades del atributo B en un cierto espacio de dimensión p.

Y si dos de estas distribuciones condicionadas columnas, por ejemplo $A|B_j$ y $A|B_{j'}$, son iguales (o muy parecidas), entonces la distribución de la variables A para individuos que presentaran $B=B_j$ ó $B=B_{j'}$ no variaría (o variaría poco); es decir, no dependería (o dependería poco) de la modalidad observada para la variable B y, por tanto, estas modalidades B_j y $B_{j'}$ que consideramos están induciendo el mismo efecto en la distribución de A. En este sentido estas modalidades B_j y $B_{j'}$ son similares (o parecidas) y podríamos decir que tienen un efecto similar sobre A.

Estamos ahora, pues, representando las distintas modalidades del atributo B como puntos de un espacio *p*-dimensional caracterizados por las coordenadas que inducen las respectivas distribuciones de frecuencias relativas condicionadas del atributo A a cada modalidad de B.

Matricialmente, podemos expresar la matrices de todas las distribuciones condicionadas filas y columnas, respectivamente, como:

$$\begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} & \frac{f_{12}}{f_{1\cdot}} & \cdots & \frac{f_{1q}}{f_{1\cdot}} \\ \frac{f_{21}}{f_{2\cdot}} & \frac{f_{22}}{f_{2\cdot}} & \cdots & \frac{f_{2q}}{f_{2\cdot}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p\cdot}} & \frac{f_{p2}}{f_{p\cdot}} & \cdots & \frac{f_{pq}}{f_{p\cdot}} \end{pmatrix}_{p\cdot q} = D_p^{-1} F \qquad \begin{pmatrix} \frac{f_{11}}{f_{\cdot 1}} & \frac{f_{12}}{f_{\cdot 2}} & \cdots & \frac{f_{1q}}{f_{\cdot q}} \\ \frac{f_{21}}{f_{\cdot 1}} & \frac{f_{22}}{f_{\cdot 2}} & \cdots & \frac{f_{2q}}{f_{\cdot q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{\cdot 1}} & \frac{f_{p2}}{f_{\cdot 2}} & \cdots & \frac{f_{pq}}{f_{\cdot q}} \end{pmatrix}_{p\cdot q} = F D_q^{-1}$$

Y si llamamos: $1_p = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}_p'$ $\qquad$ $1_q = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}_q'$

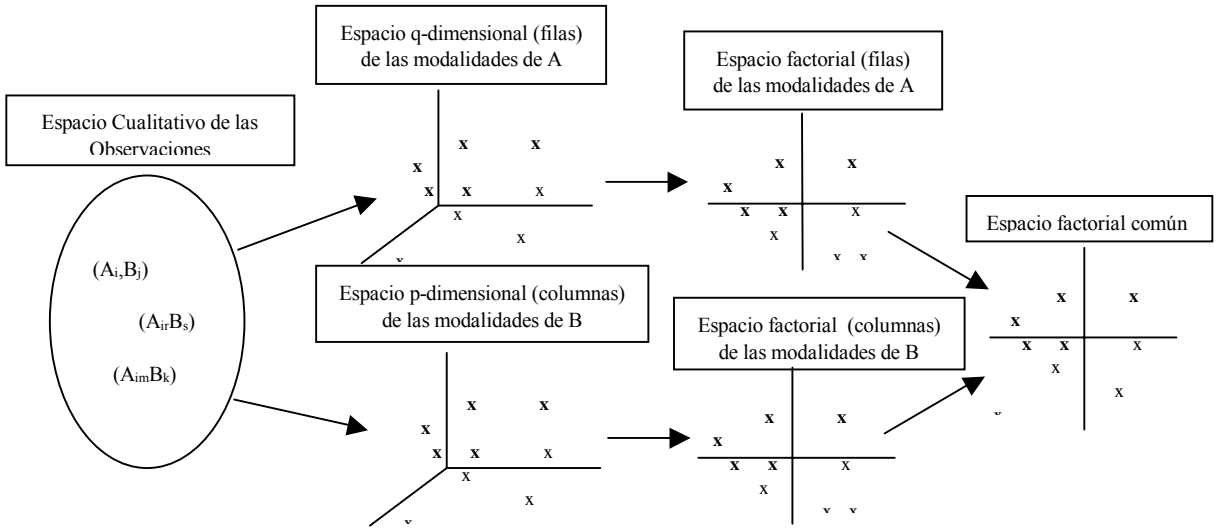
entonces, se verifica que:

$$F 1_q = D_p 1_p \quad , \quad 1_p' F = 1_q' D_q \quad \text{y} \quad 1_p' D_p 1_p = 1_p' F 1_q = 1_q' D_q 1_q = 1$$

Tanto en un caso como en otro, tiene sentido analizar las nubes de puntos así construidas desde la óptica de las proximidades, los conglomerados, los factores, etc. ya que están ubicadas en subespacios métricos de $\Re^p$ o $\Re^q$ respectivamente ser las coordenadas de los puntos frecuencias; es decir, números reales acotados en [0,1].

Estas son las ideas que subyacen en el planteamiento del Análisis Factorial de Correspondencias de Benzecri. El análisis en este sentido de la similaridad (proximidad) existente entre las distintas modalidades de un atributo (variable cualitativa) representadas por las respectivas distribuciones de frecuencias condicionadas antes indicadas, en el sentido de si condicionan o no de la misma manera la distribución de modalidades del otro atributo, nos permitirá analizar la homogeneidad de los mismos en dos espacios diferentes, uno de dimensión *q* y otro de dimensión *p;* para lo que introduce y emplea su llamada *distancia de Benzecri*. El análisis de la asociación entre modalidades de los dos diferentes atributos será consecuencia de que podremos conectar estos dos espacios y, en consecuencia, proyectarlos en un espacio común donde la proximidad será interpretada como atracción y el alejamiento como repulsión. Finalmente, para simplificar la representación de estos resultados y facilitar su interpretación, aplicaremos el Análisis de Componentes Principales sobre estos espacios.

Gráficamente, podemos representar el proceso que sigue y que iremos reproduciendo a lo largo de este tema, en el siguiente esquema:



Análisis Factorial de Correspondencias Simple: Espacio de las Filas.

Para medir el grado de proximidad entre las modalidades de A en el espacio q-dimensional, Benzecri propone la siguiente medida de disimilaridad conocida como distancia de Benzecri:

$$d^2(i,i') = \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

Como puede verse, esta distancia propuesta por Benzecri tiene una forma típica de estadístico χ^2, suma de diferencias al cuadrado, pero ponderando inversamente cada término por la frecuencia marginal de la modalidad correspondiente al sumando.

Efectivamente, cumple con las propiedades exigibles a una distancia ya que es, además, un ejemplo claro de disimilaridad euclidizable.

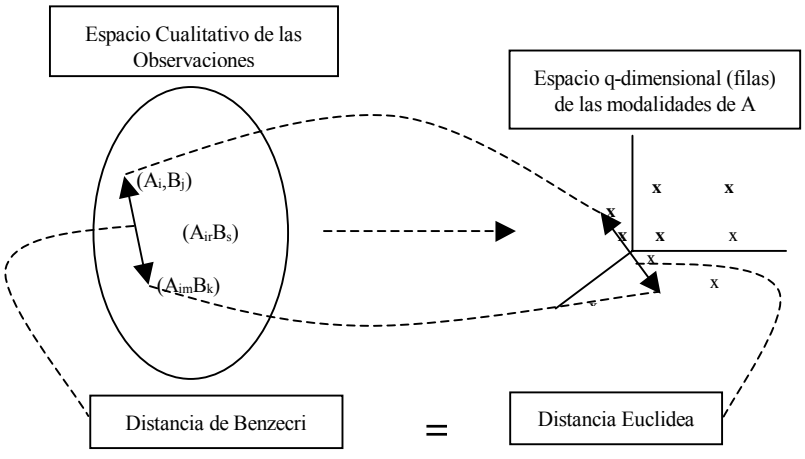
Obsérvese que si proyectamos cada modalidad A_i del atributo A en un punto del espacio euclideo *q*-dimensional de coordenadas:

$$\left(\frac{f_{i1}}{f_{i\cdot}\sqrt{f_{\cdot 1}}},...,\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}},...,\frac{f_{iq}}{f_{i\cdot}\sqrt{f_{\cdot q}}} \right) \quad i=1,2,...,p$$

el conjunto de modalidades de A quedaría representado en dicho espacio euclídeo por la nube de los *p* puntos proyectados de forma análoga, siendo la distancia euclidea entre cada dos de esos puntos justamente la distancia de Benzecri calculada entre las modalidades correspondientes:

$$d^2(i,i') = \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 = \sum_{j=1}^{q} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \frac{f_{i'j}}{f_{i'\cdot}\sqrt{f_{\cdot j}}} \right)^2$$

Así, en el **nuevo espacio**, d^2 sería el cuadrado de la distancia euclidea. Es decir, estamos transformando el espacio cualitativo inicial de las modalidades, en otro espacio q-dimensional de tipo euclideo en el que la distancia de Benzecri entre cada dos modalidades A_i y $A_{i'}$, $d(i,i')$,coincide con la distancia euclidea entre los correspondientes puntos proyectados.



La matriz de coordenadas de las p modalidades filas (de A) en el espacio q-dimensionalproyectado será:

$$X = (X_1, X_2, ..., X_q) = \left(\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \right) \right)_{\substack{i=1,2,...,p \\ j=1,2,...,q}} = D_p^{-1} F D_q^{-1/2}$$

Propiedades del Espacio de las filas:

a) La nube de puntos está contenida en el hiperplano de dimensión q-1:

$$\sum_{j=1}^{q} \sqrt{f_{\cdot j}} \cdot X_j = 1 \quad \text{, o matricialmente} \quad X D_q^{1/2} 1_q = 1_p$$

Para demostrarlo, basta sustituir X por su expresión matricial

$$X D_q^{1/2} 1_q = D_p^{-1} F D_q^{-1/2} D_q^{1/2} 1_q = D_p^{-1} F 1_q = D_p^{-1} D_p 1_p = 1_p$$

b) El centroide de la nube de puntos es:

$$\bar{x} = (\sqrt{f_{\cdot 1}}, ..., \sqrt{f_{\cdot j}}, ..., \sqrt{f_{\cdot q}})' \Rightarrow \bar{x}'\bar{x} = 1 \quad \text{, o matricialmente} \quad \bar{x} = D_q^{1/2} 1_q$$

Para verlo, tengamos en cuenta que cada modalidad fila *i-ésima* (punto de la nube) se encuentra ponderada por una frecuencia marginal $f_{i\cdot}$; por lo que el centroide de la nube de puntos X puede calcularse como:

$$\bar{x} = X' D_p 1_p = \left(D_p^{-1} F D_q^{-1/2} \right)' D_p 1_p = D_q^{-1/2} F' D_p^{-1} D_p 1_p = D_q^{-1/2} F' 1_p = D_q^{-1/2} D_q 1_q = D_q^{1/2} 1_q$$

que, obviamente está en el hiperplano de la propiedad a), ya que se verifica que:

$$\bar{x}'\bar{x} = (D_q^{1/2} 1_q)' D_q^{1/2} 1_q = 1_q' D_q 1_q = 1_p' F 1_q = 1$$

c) La matriz de datos centradas será entonces:

$$X_c = \left(\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) \right)_{\substack{i=1,2,...,p \\ j=1,2,...,q}} = X - 1_p \bar{x}' = X - 1_p 1_q' D_q^{1/2}$$

d) La matriz de Varianzas y Covarianzas de la nube de puntos, $S=((S_{lk}))$, siendo S_{lk} la $cov(X_i,X_j)$, tiene por elementos:

$$S_{lk} = \sum_{i=1}^{p} \frac{f_{il} f_{ik}}{f_{i\cdot}\sqrt{f_{\cdot l}}\sqrt{f_{\cdot k}}} - \sqrt{f_{\cdot l}} \cdot \sqrt{f_{\cdot k}} \text{ , o matricialmente } S = X_c' D_p X_c = X' D_p X - \bar{x}\bar{x}'$$

lo que se puede ver fácilmente ya que

$$S = X_c^{'} D_p X_c = (X - 1_p \bar{x}')' D_p (X - 1_p \bar{x}') =$$
$$= X' D_p X - \left(X' D_p 1_p\right)\bar{x}' - \bar{x}\left(1_p^{'} D_p X\right) + \bar{x}\left(1_p^{'} D_p 1_p\right)\bar{x}' =$$
$$= X' D_p X - \bar{x}\bar{x}' - \bar{x}\bar{x}' + \bar{x}\bar{x}' = X' D_p X - \bar{x}\bar{x}'$$

Propiedades del Espacio Factorial de las Filas:

Para obtener a partir del espacio de las filas un espacio más simple, de menor dimensión, que permita visualizar las relaciones de proximidad entre las modalidades representadas en la nube de puntos, aplicaremos el Análisis de Componentes Principales.

Para ello sabemos que las direcciones características de las componentes principales son las marcadas por los autovectores de la matriz de varianzas y covarianzas de los datos, y que la dispersión explicada por cada componente es equivalente a su autovalor asociado.

a) Los autovalores de la matriz de varianzas y covarianzas S, cumplen que:

- $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_{q-1} \geq 0$,con autovectores asociados $v_1, v_2, \ldots\ldots, v_{q-1}$
- $\lambda_q = 0$ asociado al autovector "centroide" $\bar{x}$

  lo que será cierto, si y solo sí

  $$S\bar{x} = 0\bar{x} = 0_q \Leftrightarrow (X' D_p X - \bar{x}\bar{x}')\bar{x} = 0_q \Leftrightarrow X' D_p X\bar{x} = \bar{x}\bar{x}'\bar{x} = \bar{x}$$

  y, efectivamente,

$$X' D_p X\bar{x} = \left(D_p^{-1} F D_q^{-1/2}\right)^{'} D_p \left(D_p^{-1} F D_q^{-1/2}\right)\left(D_q^{1/2} 1_q\right) =$$
$$= D_q^{-1/2} F^{'} D_p^{-1} D_p D_p^{-1} F D_q^{-1/2} D_q^{1/2} 1_q = D_q^{-1/2} F^{'} D_p^{-1} (F 1_q) = D_q^{-1/2} F^{'} D_p^{-1} D_p 1_p =$$
$$= D_q^{-1/2} (F^{'} 1_p) = D_q^{-1/2} D_q 1_q = D_q^{1/2} 1_q = \bar{x}$$

b) Si se define la matriz V=((Vlk)), siendo

$$V_{lk} = \sum_{i=1}^{p} \frac{f_{il} f_{ik}}{f_{i\cdot} \sqrt{f_{\cdot l}} \sqrt{f_{\cdot k}}}$$, o matricialmente $$V = X^{'} D_p X$$

entonces:

- el mayor autovalor de V es 1 y está asociado al autovector "centroide" $\bar{x}$

  lo que será cierto, si y solo si $V\bar{x} = 1\bar{x} \Leftrightarrow X' D_p X\bar{x} = \bar{x}$; lo cual ya hemos demostrado en a).

- los demás autovalores y autovectores de V coinciden con los correspondientes a los $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_{q-1}$ de S.

Y efectivamente, si *u* es otro autovector (distinto del centroide $\bar{x}$) de *S* asociado a un autovalor λ, es decir, que cumple que $Su=\lambda u$, entonces

$$Su = \lambda u \Leftrightarrow (X' D_p X - \bar{x}\bar{x}')u = \lambda u \Leftrightarrow X' D_p Xu - \bar{x}(\bar{x}'u) = \lambda u$$

y como $\bar{x}$ y u son autovectores (ortogonales), su producto escalar es nulo y, por tanto,

$$Su = \lambda u \Leftrightarrow X' D_p Xu = \lambda u \Leftrightarrow Vu = \lambda u$$

Análisis Factorial de Correspondencias Simple: Espacio de las Columnas.

En el espacio de dimensión *p* (Espacio de las columnas o modalidades de B) se procede de manera análoga.

Así, para medir el grado de proximidad entre las modalidades de B en el espacio p-dimensional de las columnas se utiliza la siguiente distancia de Benzecri:

$$d^2(j,j') = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2$$

Obsérvese que si proyectamos cada modalidad B_j del atributo B en un punto del espacio euclideo *p*-dimensional de coordenadas:

$$\left(\frac{f_{1j}}{f_{\cdot j}\sqrt{f_{1\cdot}}}, ..., \frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}}, ..., \frac{f_{pj}}{f_{\cdot j}\sqrt{f_{p\cdot}}} \right) \quad j=1,2,...,q$$

el conjunto de modalidades de B quedaría representado en dicho espacio euclídeo por la nube de los *q* puntos proyectados de forma análoga, siendo la distancia euclidea entre cada dos de esos puntos justamente la distancia de Benzecri calculada entre las modalidades correspondientes:

$$d^2(j,j') = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2 = \sum_{i=1}^{p} \left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} - \frac{f_{ij'}}{f_{\cdot j'}\sqrt{f_{i\cdot}}} \right)^2$$

La matriz de coordenadas de las q modalidades columnas (de B) en el espacio p-dimensional proyectado será:

$$X^*_{q\cdot p} = (X_1^*, X_2^*, ..., X_p^*) = \left(\left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} \right) \right)_{\substack{j=1,2,...,q \\ i=1,2,...,p}} = D_q^{-1} F' D_p^{-1/2}$$

Propiedades del Espacio de las Columnas:

a) La nube de puntos está contenida en el hiperplano de dimensión p-1:

$$\sum_{i=1}^{p}\sqrt{f_{i\cdot}}\cdot X_i^* = 1 \text{, o matricialmente } X^{*}D_p^{1/2}1_p = 1_q$$

b) El centroide de la nube de puntos es:

$$\bar{x}^* = (\sqrt{f_{1\cdot}},...,\sqrt{f_{i\cdot}},...,\sqrt{f_{p\cdot}})' \Rightarrow \bar{x}^{*\prime}\bar{x}^* = 1 \text{, o matricialmente } \quad \bar{x}^* = D_p^{1/2}1_p$$

c) La matriz de datos centradas será entonces:

$$X_c^* = \left(\left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} - \sqrt{f_{i\cdot}}\right)\right)_{\substack{j=1,2,...,q \\ i=1,2,...,p}} = X^* - 1_q\bar{x}^{*\prime} = X^* - 1_q 1_p' D_p^{1/2}$$

d) La matriz de Varianzas y Covarianzas de la nube de puntos, $S^*=((S^*_{lk}))$, tiene por elementos:

$$S_{lk}^* = \sum_{j=1}^{q}\frac{f_{lj}f_{kj}}{f_{\cdot j}\sqrt{f_{l\cdot}}\sqrt{f_{k\cdot}}} - \sqrt{f_{l\cdot}}\cdot\sqrt{f_{k\cdot}} \quad \text{o matricialmente} \quad S^* = X_c^{*\prime}D_q X_c^* = X^{*\prime}D_q X^* - \bar{x}^*\bar{x}^{*\prime}$$

Propiedades del Espacio Factorial de las Columnas:

Con relación a la aplicación del Análisis de Componentes Principales, se verifican las siguientes propiedades:

a) Los autovalores de la matriz de varianzas y covarianzas S^* son:

- $\mu_1 \geq \mu_2 \geq ..... \geq \mu_{p-1} \geq 0$,con autovectores asociados $u_1, u_2, ..., u_{p-1}$, y
- $\mu_p = 0$ asociado al autovector "centroide" $\bar{x}^*$

b) Si definimos la matriz $V^* =((V^*_{lk}))$, siendo

$$V_{lk}^* = \sum_{j=1}^{q}\frac{f_{lj}f_{kj}}{f_{\cdot j}\sqrt{f_{l\cdot}}\sqrt{f_{k\cdot}}} \text{, o matricialmente} \qquad V^* = X^{*\prime}D_q X^*$$

entonces:

- el mayor autovalor de V^* es 1, y está asociado al autovector centroide $\bar{x}^*$, y

- los demás autovalores y autovectores de V^* coinciden con los correspondientes a los $\mu_1 \geq \mu_2 \geq \ldots\ldots \geq \mu_{p-1}$ de S^*.

Análisis Factorial de Correspondencias Simple: Espacio Común.

Hasta ahora hemos representado las modalidades de los dos atributos enfrentados en la Tabla de Contingencia en dos espacios métricos, diferentes incluso en su dimensión, derivados de utilizar las filas o las columnas de aquélla.

Ahora veremos que, después de aplicar sendos Análisis de Componentes Principales sobre las nubes de puntos de las modalidades filas y columnas proyectadas en ambos espacios, en el de las filas de dimensión *q* y en el de las columnas de dimensión *p*, las respectivas soluciones se encuentran relacionadas de forma que existe un relación lineal que permite pasar de una a otra fácilmente y, en consecuencia, permite representar todas las modalidades, tanto filas como columnas, en un único espacio que llamaremos *espacio común*.

Teorema: Si notamos por Y la matriz

$$Y = \left(\left(\frac{f_{ij}}{\sqrt{f_{i\cdot}}\sqrt{f_{\cdot j}}} \right) \right)_{p \times q} = D_p^{-1/2} F D_q^{-1/2}$$

entonces, las matrices V y V^*, a partir de las cuales se podían obtener las componentes principales de los espacios de filas y de columnas respectivamente, cumplen que:

a) $V=Y'Y$ y $V^*=YY'$

ya que

$$\begin{aligned} Y'Y &= \left(D_p^{-1/2} F D_q^{-1/2}\right)' \left(D_p^{-1/2} F D_q^{-1/2}\right) = D_q^{-1/2} F' D_p^{-1/2} D_p^{-1/2} F D_q^{-1/2} = \\ &= D_q^{-1/2} F' D_p^{-1} F D_q^{-1/2} = \left(D_q^{-1/2} F' D_p^{-1}\right) D_p \left(D_p^{-1} F D_q^{-1/2}\right) = \\ &= \left(D_p^{-1} F D_q^{-1/2}\right)' D_p \left(D_p^{-1} F D_q^{-1/2}\right) = X' D_p X = V \end{aligned}$$

y

$$\begin{aligned} YY' &= \left(D_p^{-1/2} F D_q^{-1/2}\right) \left(D_p^{-1/2} F D_q^{-1/2}\right)' = D_p^{-1/2} F D_q^{-1/2} D_q^{-1/2} F' D_p^{-1/2} = \\ &= D_p^{-1/2} F D_q^{-1} F' D_p^{-1/2} = \left(D_p^{-1/2} F D_q^{-1}\right) D_q \left(D_q^{-1} F' D_p^{-1/2}\right) = \\ &= \left(D_q^{-1} F' D_p^{-1/2}\right)' D_q \left(D_q^{-1} F' D_p^{-1/2}\right) = X^{*'} D_q X^* = V^* \end{aligned}$$

b) Los autovalores no nulos de Y'Y (V) y de YY' (V^*) son iguales:

$$1=\lambda_1=\mu_1 \geq \lambda_2=\mu_2 \geq \ldots\ldots \geq \lambda_k=\mu_k ,$$

siendo $k \leq \min(p,q)$ y el resto de autovalores nulos.

ya que si λ es un autovalor no nulo de V=Y'Y asociado a un autovector *u*, entonces:

$$Vu = \lambda u \Leftrightarrow Y'Yu = \lambda u \Leftrightarrow YY'Yu = Y\lambda u \Leftrightarrow V^*Yu = \lambda Yu$$

de donde se deduce que es también un autovalor de V^* asociado al autovector *Yu*.

Y, recíprocamente, si μ es un autovalor no nulo de V^*=YY' asociado a un autovector *v*, entonces:

$$V^*v = \mu v \Leftrightarrow YY'v = \mu v \Leftrightarrow Y'YY'v = Y'\mu v \Leftrightarrow VY'v = \mu Y'v$$

de donde se deduce que es también un autovalor de V asociado al autovector *Y'v*.

Y si todos los autovalores no nulos lo son de ambas matrices V y V^*, si los autovectores de estas matrices son los mismos que los de las matrices S y S* con excepción del autovalor 1 asociado a sus respectivos centroides, y si sabemos que éstas matrices S y S* a lo sumo tienen respectivamente p-1 y q-1 autovectores no nulos, entonces concluimos que

$$k \leq \min\{1+(p-1),1+(q-1)\}=\min(p,q).$$

c) En consecuencia, las modalidades de ambos atributos A y B pueden ser representadas en un mismo espacio común, ya que los autovectores característicos de las componentes principales en los espacios filas y columnas (autovectores) se relacionan así se la siguiente forma:

si (λ,v) son autovalor y autovector asociados de V, entonces (λ,Yv) lo son de V^*, y si (μ,u) son autovalor y autovector asociados de V^*, entonces $(\mu,Y'u)$ lo son de V.

Es decir, todos los autovalores de V y V^* coinciden: coinciden en que el autovalor máximo que siempre vale uno ($\lambda_1=\mu_1=1$); el segundo autovalor en el espacio fila coincide con el segundo autovalor en el espacio columna ($\lambda_2=\mu_2$); el tercero coincide con el tercero ($\lambda_3=\mu_3$) ,...., y así sucesivamente hasta uno de ellos, el *k*-ésimo del espacio fila que coincide también con el *k*-ésimo del espacio columna ($\lambda_k=\mu_k$); y a partir de aquí, todos demás autovalores serán nulos.

Y aunque los autovalores sean iguales, los autovectores no tienen por qué serlos, ya que los espacios son distintos; de hecho, tienen dimensiones distintas. Sin embargo, resulta que los autovectores del espacio de las columnas se obtienen como el producto de la matriz Y por los autovectores del espacio de las filas; y los del espacio de las filas se obtienen como el producto de la transpuesta de la matriz Y por los autovectores del espacio de las columnas. Es decir, ya que mediante esta forma podemos obtener directamente unos a partir de los otros, podremos pasar de un espacio a otro a través de esta matriz de transformaciones Y; lo que nos permitirá, visualizar los correspondientes mapas que el Análisis de

Componenentes Principales proporciona para los espacios de los factores, en un mismo espacio común.

De esta forma:

- modalidades de un mismo atributo que estén cercanas indicarán comportamientos similares (producen efectos similares sobre el otro atributo)
- modalidades procedentes de atributos distintos que estén cercanas indicarán atracción entre ellas mientras que modalidades procedentes de variables distintas que se presentes muy separadas indicarán repulsión entre ellas.

Las Inercias en el Análisis Factorial de Correspondencias

Consideremos una nube de n puntos $x_i=(x_{i1},x_{i2},...,x_{ip})$ $i=1,2,...,n$, sobre los que actúan sendos pesos w_i, en un cierto espacio de dimensión p.

Se define la *inercia de un punto x_i, sobre el que actúa un peso w_i, con respecto de otro punto O*, a la cantidad:

$$I(x_i;O)=\sum_{j=1}^{p} w_i(x_{ij}-O_j)^2$$

y se calcula la *inercia de la nube de puntos, con respecto de otro punto O*, a la suma de las respectivas inercias de cada punto que la compone:

$$I(O)=\sum_{i=1}^{n} I(x_i;O)=\sum_{i=1}^{n}\sum_{j=1}^{p} w_i(x_{ij}-O_j)^2=\sum_{j=1}^{p}\sum_{i=1}^{n} w_i(x_{ij}-O_j)^2=\sum_{j=1}^{p} I_j(O)$$

recibiendo el nombre de *inercia de la dimensión j-ésima (a lo largo de la dimensión j-ésima)*, a la cantidad:

$$I_j(O)=\sum_{i=1}^{n} w_i(x_{ij}-O_j)^2$$

y conociéndose, finalmente, por *contribución absoluta del punto x_i a la inercia de la dimensión j-ésima*, así por *contribución absoluta de la dimensión j-ésima a la inercia del punto x_i* a la cantidad $w_i(x_{ij}-O_j)^2$.

Si comparamos estos conceptos con la expresión de la suma de las varianzas en cada dimensión *j-ésima* de la nube de puntos de las modalidades filas (ponderadas por sus frecuencias marginales) en el espacio q-dimensional de las filas, observamos que puede ser identificada con la inercia de la nube de puntos con respecto de su centroide, considerando como peso de cada modalidad fila su respectiva frecuencia marginal relativa $f_{i\cdot}$

$$\sum_{j=1}^{q}\lambda_j=\sum_{j=1}^{q}\sum_{i=1}^{p}f_{i\cdot}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)^2=\sum_{i=1}^{p}f_{i\cdot}\sum_{j=1}^{q}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)^2=\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(f_{ij}-f_{i\cdot}f_{\cdot j}\right)^2}{f_{i\cdot}f_{\cdot j}}=\frac{\chi^2}{n}$$

lo que puede identificarse con:

$$\sum_{j=1}^{q}\lambda_j=\frac{\chi^2}{n}=I(\bar{x})=\sum_{j=1}^{q}I_j(\bar{x})=\sum_{i=1}^{p}I(x_i;\bar{x})$$

Y recíprocamente, si comparamos aquellos conceptos con la expresión de la suma de las varianzas en cada dimensión *i-ésima* de la nube de puntos de las modalidades columnas (ponderadas por sus frecuencias marginales) en el espacio p-dimensional de las columnas, observamos que puede ser identificada con la inercias de la nube de puntos con respecto de su centroide, considerando como pesos de cada modalidad columna su respectiva frecuencia marginal relativa $f_{\cdot j}$.

$$\sum_{h=1}^{k}\mu_h=\sum_{i=1}^{p}\sum_{j=1}^{q}f_{\cdot j}\left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}}-\sqrt{f_{i\cdot}}\right)^2=\sum_{i=1}^{p}f_{\cdot j}\sum_{j=1}^{q}\left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}}-\sqrt{f_{i\cdot}}\right)^2=\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(f_{ij}-f_{\cdot j}f_{i\cdot}\right)^2}{f_{\cdot j}f_{i\cdot}}=\frac{\chi^2}{n}$$

lo que puede identificarse con:

$$\sum_{i=1}^{p}\mu_j=\frac{\chi^2}{n}=I(\bar{x}^*)=\sum_{i=1}^{p}I_i(\bar{x}^*)=\sum_{j=1}^{q}I(x_j^*;\bar{x}^*)$$

Así pues, las sumas de varianzas en las dimensiones de los espacios filas y columnas pueden interpretarse como las inercias totales de las nubes de puntos en ambos espacios filas y columnas con respecto a sus centroides, y coinciden con el coeficiente de contingencia cuadrático medio de Pearson; pudiéndose, a partir de las expresiones anteriores, identificar claramente las distintas contribuciones absolutas de los puntos a las inercias de las nubes, de los puntos a las inercias de las dimensiones y de las dimensiones a la inercia de la nube; para, a partir de ellas, obtener por simples cocientes las contribuciones relativas de cada punto o dimensión a dichas inercias.

Análisis Factorial de Correspondencias Múltiple

Hasta ahora hemos hablado del Análisis Factorial de Correspondencias Simple que se aplica cuando tratamos de analizar comportamientos de las modalidades de dos variables. Si tratamos de analizar simultáneamente el comportamiento de más de dos variables, deberemos emplear en Análisis Factorial de Correspondencias Múltiple.

El procedimiento más inmediato para aplicar el Análisis Factorial de Correspondencias Múltiple utilizando lo ya estudiado para el Análisis Factorial de Correspondencias Simple consiste justamente en elaborar la llamada *Tabla de Burr,* que recoge ordenadamente todas

las correspondientes tablas de contingencia simples formadas con cada posible combinación de 2 variables de entre todas las que intervienen en el análisis, tal y como aparece en el siguiente diagrama:

		Var1			Var2				Var3		
		m_1	...	m_k	n_1	...	n_p	...	R_1	...	r_q
Var1	m_1		0	0	T12			...	T13		
	...	0		0				...			
	m_k	0	0					...			
Var2	n_1	T21				0	0	...	T23		
	...				0		0	...			
	n_p				0	0		...			
	...	...	...	...	...	...	...	...	...	...	...
Var3	r_1	T31			T32			...		0	0
								...	0		0
	r_q							...	0	0	

Esta Tabla de Burr es una tabla también de dimensión dos en la que se funden ordenadamente todas las tablas simples en las que se enfrentan cada una de las variables con otra variable. Todas las tablas que no están en la diagonal principal son tablas de contingencia que obteníamos al cruzar una variable con otra. Por ejemplo, si enfrentamos la variable 1 con la variable 2, tendremos la tabla de contingencia simple T12; si enfrentamos la variable 1 con la variable 3, tendremos la tabla de contingencia T13; y así sucesivamente. En la diagonal de la Tabla de Burr, aparecen unas tablas de contingencia especiales que se obtienen al enfrentar una variable consigo misma. En su diagonal aparecerían las frecuencias de aparición de cada una de sus modalidades y el resto de sus posiciones serían ceros.

Así, la Tabla de Burr sería una especie de tabla de contingencia general que permitiría analizar la proximidad y asociación entre todas esas modalidades de todas las variables analizadas, tal y como se ha hecho en el Análisis de Correspondencias Simple. Por la propia estructura de la Tabla de Burr que contiene tanto en sus filas como en sus columnas todas las modalidades de todas las variables, cualquiera de los tres espacios estudiados en la aplicación del Análisis Factorial de Correspondencias Simple (espacio de las filas, espacio de las columnas y espacio común) reflejarán todas las relaciones de proximidad y asociación existentes entre aquéllas.

Como consecuencia de que todas las modalidades de todas las variables se encuentran tanto en las filas como en las columnas, si enfocamos el análisis de dos variables cualitativas a través de una tabla de Burr, los espacios filas y columnas que esta tabla genera "duplica" la dimensión de los espacios filas y columnas del enfoque del análisis factorial de correspondencias simple; por lo que puede verse que los autovalores deducidos por este enfoque serán las raíces cuadradas de los autovalores obtenidos mediante el enfoque simple.