Análisis Multivariado I
Análisis de Correspondencias Simples

Mathias Bourel

DMMC - Facultad de Ciencias Económicas y Administración, Universidad de la República, Uruguay
IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

29 de junio de 2016

Introducción

El Análisis de Correspondencias sigue un procedimiento análogo a la técnica de componentes principales, esta vez a través del estudio de tablas de contingencias. En las mismas se consideran las frecuencias de aparición de variables *cualitativas* en un conjunto de elementos. La idea consiste en estudiar qué modalidades de una variable están asociadas con qué modalidades de la otra variable. También vamos a querer ver qué categorías de una misma variable son parecidas entre sí.

Por ejemplo podemos querer conocer la opinión de consumidores de tv cable en un barrio de Montevideo en función del cable que tienen, I posibilidades, y su conformidad en cuanto a la programación, J posibles opiniones, obteniendose de esta manera una matriz $X \in \mathcal{M}_{I \times J}$.

Cable	Poco Conforme	Conforme	Muy conforme	Total
Nuevo Siglo	3	47	178	228
Montecable	24	56	20	100
TCC	12	8	23	43
DirectTV	2	14	88	104
Total	41	125	309	475

$$X = \begin{pmatrix} 3 & 47 & 178 \\ 24 & 56 & 20 \\ 12 & 8 & 23 \\ 2 & 14 & 88 \end{pmatrix} \in \mathcal{M}_{4 \times 3}$$

Cada persona aparece en una sola casilla de la tabla.

El objetivo del Análisis de Correspondencias Simples es de estudiar las relaciones entre las modalidades de dos variables cualitativas. Para eso se reduce la dimensión, como en componentes

principales, descomponiendo las nubes de puntos filas y la ube de puntos columnas de la tabla de contingencia asociada a las modalidades de ambas variables.

$X \mid Y$	y_1	$\dots$	y_j	$\dots$	y_J	
x_1			$\vdots$			
			$\vdots$			
x_i	$\dots$	$\dots$	n_{ij}	$\dots$	$\dots$	$n_{i\cdot}$
			$\vdots$			
x_I			$\vdots$			
			$n_{\cdot j}$			

Cuadro: Tabla de contingencias

- Suponemos que $I > J$.
- $x_1, x_2, \dots, x_I$ representan las modalidades de la variables X.
- $y_1, y_2, \dots, y_J$ representan las modalidades de la variables Y.
- n es la cantidad de individuos
- n_{ij} es la cantidad de individuos que cumplen la modalidad i de la variable X y la modalidad j de la variable Y.
- $n_{i\cdot}$ es la cantidad de individuos que cumplen la modalidad i de la variable X
- $n_{\cdot j}$ es la cantidad de individuos que cumplen la modalidad j de la variable Y.

	Hotel	Locación	Res.Secund	Padres	Amigos	Camping	Grupo Viaje	Otros	Total
Prod. Rurales	195	62	1	499	44	141	49	65	1056
Jefes	700	354	229	959	185	292	119	140	2978
Ejecutivo sup	961	471	633	1580	305	360	162	148	4620
Ejecutivo prom	572	537	279	1689	206	748	155	112	4298
Empleado	441	404	166	1079	178	434	178	92	2972
Obrero	783	1114	387	4052	497	1464	525	387	9209
Otras prof.	142	103	210	1133	132	181	46	59	2006
Inactivos	741	332	327	1789	311	236	102	102	3940
Total	4535	3377	2232	12780	1858	3856	1336	1105	31079

Matriz de frecuencias relativas

Podemos trabajar con la matriz F de frecuencias relativas, pensada como matriz de probabilidades, que se obtiene de la matriz anterior dividiendo cada casilla de la tabla de contingencia por n el total de valores observados

$X\|Y$	y_1	$\dots$	y_j	$\dots$	y_J	
x_1			$\vdots$			
			$\vdots$			
x_i	$\dots$	$\dots$	$f_{ij} = \frac{n_{ij}}{n}$	$\dots$	$\dots$	$f_{i\cdot} = \frac{n_{i\cdot}}{n}$
			$\vdots$			
x_I			$\vdots$			
			$f_{\cdot j} = \frac{n_{\cdot j}}{n}$			

- A $f_{i\cdot}$ se le llama *frecuencia marginal de la modalidad* i.
- A $f_{\cdot j}$ se le llama *frecuencia marginal de la modalidad* j.
- A $f_i^j = \frac{n_{ij}}{n_{i\cdot}} = \frac{f_{ij}}{f_{i\cdot}}$ se le llama *frecuencia condicional* j sabiendo i.
- A $f_j^i = \frac{n_{ij}}{n_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}$ se le llama *frecuencia condicional* i sabiendo j.

Cada fila puede considerarse como un punto en el espacio euclideo $\mathbb{R}^J$. La idea consistirá en buscar un subespacio de $\mathbb{R}^J$ de dimensión menor donde podamos ver la distancia entre estos puntos, de manera que las filas que tienen estructura parecidas estén cercas y las que tienen estructuras muy distintas alejadas.
La distancia euclidea no es una buena medida de proximidad de los puntos en la matriz de frecuencia F:

A	0.03	0.06	0.15	0.06	0.3
B	0.07	0.14	0.35	0.14	0.7
T	0.1	0.2	0.5	0.2	1

Si hacemos la distancia euclidea entre las filas obtenemos un valor alto. Sin embargo las dos filas tienen exactamente la misma estructura relativa pues si dividimos cada casillero por la frecuencia relativa de la fila $f_{i\cdot}$ obtenemos:

A	0.1	0.2	0.5	0.2	1
B	0.1	0.2	0.5	0.2	1

La operación matricial par pasar de una tabla a otra es

$$R = D_f^{-1} F$$

donde D_f es una matriz diagonal $I \times I$ que tiene en su diagonal principal al vector

$$\mathbf{f} = (f_{1\cdot}, f_{2\cdot}, \ldots, f_{I\cdot})$$

Esto motiva el poder considerar las tablas de los perfiles fila y columna.

Tabla de perfiles por filas

Podemos trabajar con la matriz de perfiles filas que se obtiene de la tabla de contingencia dividiendo cada casilla de la fila i por la suma $n_{i\cdot}$ del total de valores observados para ella.

$X \mid Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_J	
x_1	$\frac{f_{11}}{f_{1\cdot}}$	$\frac{f_{12}}{f_{1\cdot}}$	$\dots$	$\frac{f_{1j}}{f_{1\cdot}}$	$\dots$	$\frac{f_{1J}}{f_{1\cdot}}$	1
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\vdots$
i	$\frac{f_{i1}}{f_{i\cdot}}$	$\frac{f_{i2}}{f_{i\cdot}}$	$\dots$	$\frac{f_{ij}}{f_{i\cdot}}$	$\dots$	$\frac{f_{iJ}}{f_{i\cdot}}$	1
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\vdots$
x_I	$\frac{f_{I1}}{f_{I\cdot}}$	$\frac{f_{I2}}{f_{I\cdot}}$	$\dots$	$\frac{f_{Ij}}{f_{I\cdot}}$	$\dots$	$\frac{f_{IJ}}{f_{I\cdot}}$	1

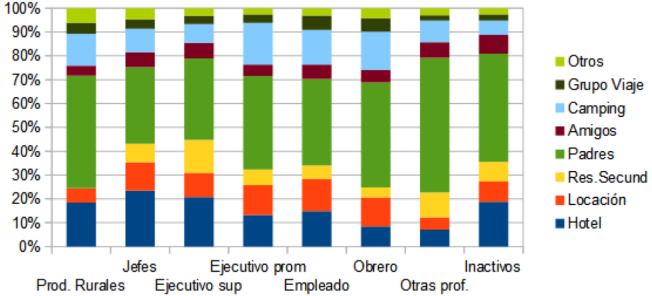
Cuadro: Tabla de perfiles filas; cada fila suma 1

La misma permite comparar la repartición de los valores de Y en las distintas modalidades de X.

Como todas las filas suman 1, todos los puntos i están sobre un hiperplano de dimensión $J-1$ de $\mathbb{R}^J$.

Ejemplo

	Hotel	Locación	Res.Secund	Padres	Amigos	Camping	Grupo Viaje	Otros	Total
Prod. Rurale	0,184659091	0,058712121	0,00094697	0,472537879	0,041666667	0,133522727	0,046401515	0,06155303	1
Jefes	0,235057085	0,118871726	0,076897246	0,322028207	0,06212223	0,098052384	0,039959704	0,047011417	1
Ejecutivo su	0,208008658	0,101948052	0,137012987	0,341991342	0,066017316	0,077922078	0,035064935	0,032034632	1
Ejecutivo pro	0,133085156	0,124941833	0,064913913	0,392973476	0,047929269	0,174034435	0,036063285	0,026058632	1
Empleado	0,148384926	0,135935397	0,055854643	0,363055182	0,059892328	0,14602961	0,059892328	0,030955585	1
Obrero	0,085025519	0,120968618	0,042024107	0,440004344	0,053968943	0,158974916	0,057009447	0,042024107	1
Otras prof.	0,070787637	0,051345962	0,104685942	0,564805583	0,065802592	0,090229312	0,022931206	0,029411765	1
Inactivos	0,188071066	0,084263959	0,082994924	0,454060914	0,07893401	0,059898477	0,025888325	0,025888325	1
Total	1,253079138	0,796987669	0,565330733	3,351456926	0,476333356	0,938663939	0,323210747	0,294937493	8



El 23 % de los jefes van al hotel y en las otras profesiones el 56 % van a la casa de los padres.

Podemos trabajar con la matriz de perfiles columnas que se obtiene de la tabla de contingencia dividiendo cada casilla de la columna j por la suma $n_{.j}$ del total de valores observados para ella.

$X\|Y$	y_1	$\dots$	y_j	$\dots$	y_J
1	$\frac{f_{11}}{f_{.1}}$	$\vdots$	$\frac{f_{1j}}{f_{.j}}$	$\vdots$	$\frac{f_{1J}}{f_{.J}}$
2	$\frac{f_{21}}{f_{.2}}$	$\vdots$	$\frac{f_{2j}}{f_{.j}}$	$\vdots$	$\frac{f_{2J}}{f_{.J}}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$\frac{f_{i1}}{f_{.1}}$	$\vdots$	$\frac{f_{ij}}{f_{.j}}$	$\vdots$	$\frac{f_{iJ}}{f_{.J}}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	$\frac{f_{I1}}{f_{.1}}$	$\vdots$	$\frac{f_{Ij}}{f_{.j}}$	$\vdots$	$\frac{f_{IJ}}{f_{.J}}$
	1	$\dots$	1	$\dots$	1

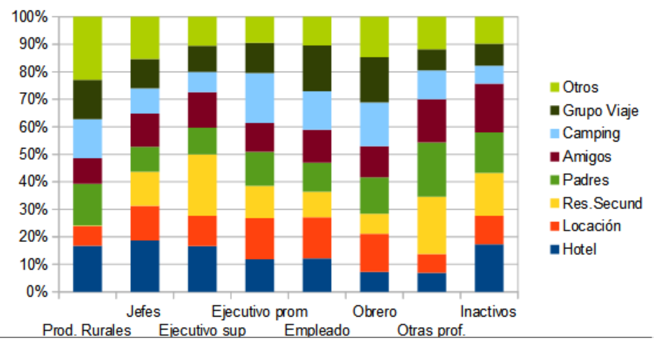
Cuadro: Tabla de perfiles columnas; cada columna suma 1.

La misma permite comparar la repartición de los valores de X en las distintas modalidades de Y.

Idem acá: todas las columnas suman 1, todas los puntos j están sobre un hiperplano de dimensión $I-1$ de $\mathbb{R}^I$.

Ejemplo

	Hotel	Locación	Res.Secund	Padres	Amigos	Camping	Grupo Viaje	Otros	Total
Prod. Rurale	0,042998897	0,018359491	0,000448029	0,039045383	0,023681378	0,03656639	0,036676647	0,058823529	0,256599744
Jefes	0,154355017	0,104826769	0,102598566	0,075039124	0,099569429	0,075726141	0,089071856	0,126696833	0,827883735
Ejecutivo su	0,211907387	0,139472905	0,283602151	0,123630673	0,164155005	0,093360996	0,121257485	0,133936652	1,271323253
Ejecutivo pro	0,126130099	0,159016879	0,125	0,132159624	0,110871905	0,193983402	0,116017964	0,101357466	1,06453734
Empleado	0,09724366	0,11963281	0,07437276	0,084428795	0,095801938	0,112551867	0,133233533	0,083257919	0,800523282
Obrero	0,172657111	0,32987859	0,173387097	0,317057903	0,267491927	0,37966805	0,392964072	0,350226244	2,383330994
Otras prof.	0,031312018	0,030500444	0,094086022	0,088654147	0,071044133	0,046939834	0,034431138	0,053393665	0,450361401
Inactivos	0,16339581	0,098312111	0,146505376	0,139984351	0,167384284	0,06120332	0,076347305	0,092307692	0,94544025
Total	1	1	1	1	1	1	1	1	8



El 15,4 % de las personas que van al hotel son jefes. Dentro de las personas que van al hotel hay una mayoría de ejecutivos superiores, pero estos ultimos prefieren ir en lo de los padres (ver perfil

Prueba de independencia entre las variables

Antes de comenzar el estudio debemos ver si las dos variables son independientes. Recordamos que dos variables X e Y son independientes si para todo par (i, j) se tiene que

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i)\mathbb{P}(Y = y_j) \ (*)$$

Esto equivale, cuando se puede hablar de probabilidad condicional, a que

- para todo par (i, j), $\mathbb{P}(X = x_i|Y = y_j) = \mathbb{P}(X = x_i)$.
- para todo par (i, j), $\mathbb{P}(Y = y_j|X = x_i) = \mathbb{P}(Y = y_j)$.

La expresión $(*)$ que se traduce en la tabla de frecuencias relativas por

$$f_{ij} = f_{i.}f_{.j} \ \ \forall\, i = 1, \ldots, I,\ j = 1, \ldots, J.$$

Si no se cumple esta igualdad para todo i y para todo j hay algún grado de asociación entre ambas variables.

Observar que

$$\frac{f_{ij}}{f_{i.}} = \frac{n_{ij}}{n_{i.}}$$

por lo que la propiedad de independencia se traduce como

$$\frac{n_{ij}}{n} = \frac{n_{i.}}{n}\frac{n_{.j}}{n} \ \forall\, i,j \Rightarrow \underbrace{n_{ij}}_{\text{valor observado}} = \underbrace{\frac{n_{i.}n_{.j}}{n}}_{\text{valor teórico}} \quad \forall\, i,j$$

Por último:

- Si $f_{ij} > f_{i.}f_{.j}$ decimos que las modalidades i y j se atraen.
- Si $f_{ij} < f_{i.}f_{.j}$ decimos que las modalidades i y j se repelen.

Prueba de independencia entre las variables

Se hace necesario definir un test estadístico global que mida de alguna manera la distancia entre lo observado y lo que uno espera, dado que se cumple la hipótesis nula de independencia entre las variables.

(H0): X e Y son independientes
(H1): X e Y no son independientes

El estadístico es:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - \frac{n_{i}.n_{.j}}{n}\right)^2}{\frac{n_{i}.n_{.j}}{n}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(nf_{ij} - nf_{i}.f_{.j}\right)^2}{nf_{i}.f_{.j}} = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(\text{fr. observadas - fr esperadas})^2}{\text{fr. esperadas}}$$

El valor n_{ij} es el valor observado en la celda i/j y el cociente $\frac{n_{i}.n_{.j}}{n}$ es el valor esperado de la celda ij bajo (H0). El estadístico χ^2 mide el desvío entre lo que se observa y lo esperado en caso de independencia y sigue asintoticamente una distribución χ^2 con $(I-1)\times(J-1)$ grados de libertad.
Si el valor de χ^2 es grande entonces tenemos una dependencia grande entre las variables.
Sin embargo al ser un indicador global, es insuficiente para medir las asociaciones entre las modalidades.

Prueba de independencia entre las variables

Veamos en el ejemplo siguiente por qué se divide por el valor teórico.

$$\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Jugamos a cara o cruz 10 veces. Se gana 1 vez y la diferencia es 4. Jugamos a cara o cruz 100 veces. Se gana 46 veces y la diferencia es 4.

Teórico	Observado	Diferencia	Diferencia^2	$\frac{\text{Diferencia}^2}{\text{Teórico}}$
5	1	4	16	3.2
50	46	4	16	0.32

Claramente no es lo mismo ganar 1 vez en 10 que ganar 46 veces en 100.

```
> base=read.table("vacaciones.csv",sep=",",header=TRUE,row.names=1)
> k=chisq.test(base)
> k$observed
                Hotel Locación Res.Secund Padres Amigos Camping Grupo.Viaje Otros
Prod. Rurales     195       62          1    499     44     141          49    65
Jefes             700      354        229    959    185     292         119   140
Ejecutivo sup     961      471        633   1580    305     360         162   148
Ejecutivo prom    572      537        279   1689    206     748         155   112
Empleado          441      404        166   1079    178     434         178    92
Obrero            783     1114        387   4052    497    1464         525   387
Otras prof.       142      103        210   1133    132     181          46    59
Inactivos         741      332        327   1789    311     236         102   102
> k$expected
                    Hotel  Locación Res.Secund    Padres    Amigos   Camping Grupo.Viaje     Otros
Prod. Rurales    154.0899  114.7435   75.83873  434.2379  63.13099  131.0189    45.39451  37.54561
Jefes            434.5452  323.5853  213.87097 1224.5838 178.03417  369.4832   128.01596 105.88146
Ejecutivo sup    674.1433  502.0026  331.79446 1899.7909 276.19808  573.2076   198.60098 164.26204
Ejecutivo prom   627.1576  467.0146  308.66939 1767.3812 256.94791  533.2568   184.75910 152.81348
Empleado         433.6697  322.9333  213.44007 1222.1165 177.67547  368.7388   127.75804 105.66814
Obrero          1343.7632 1000.6369  661.36259 3786.8342 550.54287 1142.5691   395.86937 327.42189
Otras prof.      292.7124  217.9691  144.06487  824.8875 119.92497  248.8863    86.23238  71.32244
Inactivos        574.9188  428.1148  282.95891 1620.1680 235.54555  488.8394   169.36967 140.08494
> k

        Pearson's Chi-squared test

data:  base
X-squared = 2292.148, df = 49, p-value < 2.2e-16
```

En la primera tabla vemos los n_{ij} y en la segunda tabla los productos $n_{.j}n_{.j}/n$.
Conclusión: Hay una dependencia significativa entre las variables.

Tenemos entonces dos nubes de perfiles:

- Una nube N^I de I puntos filas en $\mathbb{R}^J$.
  Cada punto de N^I está afectado por un peso que representa la importancia de la modalidad i de X: $f_{i.} = \frac{n_{i.}}{n}$
- Una nube N^J de J puntos columnas en $\mathbb{R}^I$
  Cada punto de N^J está afectado por un peso que representa la importancia de la modalidad j de Y: $f_{.j} = \frac{n_{.j}}{n}$

Entonces:

1. El centro de gravedad de la nube de filas N^I es $G_I = (g_{x_1}, g_{x_2}, \cdots, g_{x_J})$ donde
$$g_{x_j} = \sum_{i=1}^{I} \frac{n_{i.}}{n} \frac{n_{ij}}{n_{i.}} = \sum_{i=1}^{I} f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j}$$
   Entonces si G_I es el baricentro de las filas se tiene que
$$G_I = (f_{.1}, f_{.2}, \ldots, f_{.J}) \in \mathbb{R}^J$$
2. Análogamente si G_J es el baricentro de las columnas se tiene que
$$G_J = (f_{1.}, f_{2.}, \ldots, f_{I.}) \in \mathbb{R}^I$$

Distancia chi-cuadrado

Para obtener comparaciones razonables debemos tener en cuenta la frecuencia relativa de aparición del atributo estudiado y por lo tanto el peso de cada fila y de cada columna.

En lugar de mirar la diferencia entre las filas r_a y r_b con $\sum_{j=1}^{J}\left(\frac{f_{aj}}{f_{a\cdot}}-\frac{f_{bj}}{f_{b\cdot}}\right)^2$ miraremos

$$D^2(r_a, r_b) = \sum_{j=1}^{J} \frac{1}{f_{\cdot j}} \left(\frac{f_{aj}}{f_{a\cdot}}-\frac{f_{bj}}{f_{b\cdot}}\right)^2 = \sum_{j=1}^{J} \frac{(r_a - r_b)^2}{f_{\cdot j}} = (r_a - r_b)' D_c^{-1}(r_a - r_b)$$

donde D_c es una matriz diagonal $J \times J$ con los términos $f_{\cdot 1}, f_{\cdot 2}, \ldots, f_{\cdot J}$

De esta manera ponderamos con la frecuencia relativa que tiene cada columna j, equilibrando la influencia de las columnas sobre la distancia entre las filas, aumentando los términos, en principio menores, de modalidades raras. Este procedimiento es análogo al de dividir por el desvío estandar en las variables continuas.

La distancia se hace más grande si a y b están asignadas de manera distintas según las modalidades de Y. Por ejemplo la distancia entre dos categorías socio profesionales es más importante si están en distintos lugares para vacacionar.

Distancia chi-cuadrado

De la misma manera, entre las columnas:

$$D^2(c_a, c_b) = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ia}}{f_{.a}} - \frac{f_{ib}}{f_{.b}} \right)^2 = (c_a - c_b) D_f^{-1} (c_a - c_b)$$

Distancia chi-cuadrado

La distancia χ^2 así definida tiene la propiedad de *equivalencia distribucional*, esto es que si se cumple que dos filas (columnas) son proporcionales entonces la distancia entre dos columnas (filas) cualesquiera no se modifica agrupando las dos filas en una sola con peso igual a la suma de los pesos.
En el ejemplo siguiente, las dos primeras filas son proporcionales y entonces la distancia entre las dos primeras columnas es

	X	Y	Z	T	Total
A	4	3	2	1	10
B	12	9	6	3	30
C	2	4	8	6	20
Total	18	16	16	10	60

$$\frac{60}{10}\left(\frac{4}{18}-\frac{3}{16}\right)^2+\frac{60}{30}\left(\frac{12}{18}-\frac{9}{16}\right)^2+\frac{60}{20}\left(\frac{2}{18}-\frac{4}{16}\right)^2=\frac{25}{188}$$

y si agrupamos las dos primeras filas:

	X	Y	Z	T	Total
A'	16	12	8	4	40
C	2	4	8	6	20
Total	18	16	16	10	60

$$\frac{60}{40}\left(\frac{16}{18}-\frac{12}{16}\right)^2+\frac{60}{20}\left(\frac{2}{18}-\frac{4}{16}\right)^2=\frac{25}{188}$$

Nube de perfiles

1. Las I filas forman una nube de puntos N^I (nube de filas) en un subespacio de dimensión $J-1$ de $\mathbb{R}^J$. Los ejes son las columnas y G_I al baricentro de la nube N^I. La inercia de la nube de filas es:

$$\text{Inercia}\,(N_I) = \sum_{i=1}^{I} \underbrace{d^2_{\chi^2}(\text{fila } i,\ G_I) \times \text{ peso fila } i}_{\text{inercia}(i)} = \sum_{i=1}^{I}\left[\sum_{j=1}^{J}\left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2 \frac{1}{f_{.j}}\right] f_{i.}$$

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

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

   Entonces si χ^2 es grande, no hay posiblemente independencia y la inercia es grande.

2. Análogamente para las J columnas: forman una nube de puntos N^J (nube de columnas) en un subespacio de dimensión $I-1$ de $\mathbb{R}^I$. Los ejes son las filas y si G_J al baricentro de la nube N^J, la inercia es

$$\text{Inercia}\,(N_J) = \frac{\chi^2}{n}$$

3. La inercia es nula cuando todos los perfiles filas (resp. columnas) son iguales al centro de gravedad sii todas las distribuciones de Y sabiendo que X es i (resp. X sabiendo que Y es j) son iguales e iguales a la distribución marginal de Y (resp. de X) sii X e Y son independientes.

Descomposición factorial

Las filas y las columnas de la tabla de contingencias son dos particiones de los mismos individuos. Juegan entonces papeles simétricos. El Análisis de Correspondencias consiste en

- comparar la distribución de Y entre las distintas modalidades de X (estudiar los perfiles fila).
- comparar la distribución de X entre las distintas modalidades de Y (estudiar los perfiles columna).
- Identificar aquellas casillas de la tabla de contingencia en las cuales difieren de los efectivos teóricos bajo la hipótesis de independencia

El Análisis de Correspondencias es una Análisis de Componentes Principales sobre la tabla de contingencias usando la distancia χ^2.

Procedemos en hacer las transformaciones siguientes:

$$x_{ij} = \frac{f_i^j}{\sqrt{f_{.j}}} \qquad y_{ij} = \frac{f_j^i}{\sqrt{f_{i.}}}$$

y las correspondientes nubes de puntos N^I y N^J.
Entonces con este cambio queda:

1. El centro de gravedad $G_X = (g_{x_1}, \ldots, g_{x_J})$ de N^I tiene como coordenadas

$$g_{x_j} = \sum_{i=1}^{I} f_{i.} x_{ij} = \sum_{i=1}^{I} f_{i.} \frac{f_i^j}{\sqrt{f_{.j}}} = \sqrt{f_{.j}}$$

2. El centro de gravedad $G_Y = (g_{y_1}, \ldots, g_{y_I})$ de N^J tiene como coordenadas

$$g_{y_i} = \sum_{j=1}^{J} f_{.j} y_{ij} = \sum_{j=1}^{J} f_{.j} \frac{f_j^i}{\sqrt{f_{i.}}} = \sqrt{f_{i.}}$$

3. La inercia para las dos nubes de puntos es $I = \frac{\chi^2}{n}$ (práctico)

Proyección factorial de las nubes

Lso ejes pasan por G_X; consideramos la tabla de perfiles transformados y centrados

$$\tilde{X} = \begin{pmatrix} x_{11} - g_{x_1} & \dots & x_{1j} - g_{x_j} & \dots & x_{1J} - g_{x_J} \\ x_{21} - g_{x_1} & \dots & x_{2j} - g_{x_j} & \dots & x_{2J} - g_{x_J} \\ \vdots & & \vdots & & \vdots \\ x_{i1} - g_{x_1} & \dots & x_{ij} - g_{x_j} = \frac{f_i^j - f_{.j}}{\sqrt{f_{.j}}} & \dots & x_{iJ} - g_{x_J} \\ \vdots & & \vdots & & \vdots \\ x_{I1} - g_{x_1} & \dots & x_{Ij} - g_{x_j} & \dots & x_{IJ} - g_{x_J} \end{pmatrix}$$

El i-esimo perfil de $\tilde{X}$ (la fila i) tiene peso $f_{i.}$.

Sea $S = \tilde{X}'P\tilde{X} \in \mathcal{M}_{J\times J}$ la matriz de varianzas y covarianzas de $\tilde{X}$ donde

$$P^{1/2} = diag(\sqrt{f_{1.}}, \dots, \sqrt{f_{I.}}) = D_f^{1/2}$$

Observar que:

$$I = \sum_{i=1}^{I} f_{i.} d^2(\tilde{X}^{(i)}, G_X) = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{i.}(x_{ij} - g_{x_j})^2 = tr(S) \quad \text{(ya lo vimos con ACP)}$$

Si definimos $C = P^{1/2}\tilde{X} \in \mathcal{M}_{I\times J}$ entonces $S = C'C$ y $c_{ij} = (x_{ij} - g_{x_j})\sqrt{f_{i.}}$

Recordar que en ACP, $S = \tilde{X}'P\tilde{X} = C'C$ con $c_{ij} = \frac{x_{ij} - g_{x_j}}{\sqrt{n}}$

Proyección factorial de las nubes

Buscamos las direcciones donde nos alejamos más de la independencia y vamos a querer maximizar la inercia de los puntos proyectados

De vuelta, si $u_1 = (u_{11}, \dots, u_{1J})$ es la dirección buscada, de norma 1, proyecto cada perfil fila $\tilde{X}^i$ sobre u_1:

$$c_1(i) = \langle \tilde{X}^i, u_1 \rangle = \sum_{j=1}^{J} \frac{f_i^j - f_{.j}}{\sqrt{f_{.j}}} u_{1j}$$

$$c_1 = \tilde{X} u_1 = P^{1/2} C u_1$$

y la inercia sobre el primer eje es:

$$I_1 = \sum_{i=1}^{I} f_{i.} c_1(i)^2 = ||P^{1/2} C||^2 = c_1^{'} P c_1 = u_1^{'} \tilde{X}' P \tilde{X} u_1 = u_1^{'} S u_1$$

Entonces buscamos el primer eje u_1 de manera que

$$\begin{cases} u_1^{'} S u_1 \text{ sea máximo} \\ ||u_1|| = 1 \end{cases}$$

y los demás ejes se buscan como en ACP: se diagonaliza S, los ejes factoriales pasan por G_X y sus direcciones son los vectores propios asociados a valores propios de S y son ortogonales dos a dos.

La inercia sobre el primer eje es entonces

$$I_1 = \sum_{i=1}^{I} f_{i.} c_1(i)^2 = u_1^{'} S u_1 = \lambda_1$$

En el práctico se probará que si S la matriz de varianzas-covarianzas de $\tilde{X}$, es decir $S = \tilde{X}'P\tilde{X}$ donde $P = D_f$, entonces:

1. los valores propios de S son todos menores o iguales a 1.
2. S tiene como mayor valor propio a 1 y como vector propio asociado $D_c^{1/2}$.

Esto último nos indica que esta solución es trivial y no da información sobre la estructura de las filas. Tomaremos entonces como mayor valor propio menor que 1 y su vector propio asociado u_1. Entonces proyectamos $\tilde{X}$ en la dirección de u_1 obteniendo

$$c_1 = \tilde{X}u_1 = D_f^{-1}FD_c^{-1/2}u_1$$

y el vector c_1 es la mejor representación de las filas de la tabla de contingencia en una dimensión. De la misma manera, si seguimos con el vector propio asociado al segundo valor propio más grande, entonces podemos representar las filas en un espacio de dimensión 2.

$$C_f = \tilde{X}[u_1, u_2]$$

donde $[u_1, u_2]$ es una matriz $I \times 2$ cuyas columnas son los vectores propios de S asociado a los dos mayores valores propios. Las dos coordenadas de cada fila nos da la mejor representación de las filas de F en un espacio de dimensión dos.

Proyección de la nube de filas

En resumen, para buscar una buena representación de las filas de la tabla de contingencia:

1. Consideramos las filas por sus frecuencias relativas condicionadas y las consideramos como puntos en el espacio.
2. La distancia es la distancia χ^2
3. Proyectamos sobre las direcciones de máxima variabilidad, teniendo en cuenta el peso relativo de cada fila

Para ello:

1. Calculamos la matriz S de varianzas-covarianzas de $\tilde{X}$
2. Consideramos los $J-1$ vectores propios $u_1, \ldots, u_{J-1}$ asociados a los valores propios menores que 1.
3. Calculamos las proyecciones $\tilde{X}u_j$ para todo $j = 1, \ldots, J-1$.

Proyección nube columnas

Como en ACP vamos a querer proyectar también la nube de puntos columnas N^J. En vez de considerar $\tilde{X} = D_f^{-1} F D_c^{-1/2}$ y los valores y vectores propios de

$$S = \tilde{X}' P \tilde{X} = D_c^{-1/2} F' D_f^{-1} F D_c^{1/2} \quad \text{para } N^I$$

intercambiamos los roles de i y de j y trabajamos con las matrices $D_c^{-1} F' D_f^{-1/2}$ y

$$T = D_f^{-1/2} F D_c^{-1} F' D_f^{-1/2}$$

Observar que si escribimos $Z = D_f^{-1/2} F D_c^{-1/2}$ entonces

$$S = Z'Z \quad \text{y} \quad T = ZZ'$$

Como ya vimos $S \in \mathcal{M}_{I\times I}$ y $T \in \mathcal{M}_{J\times J}$ tienen los mismos valores propios no nulos.

Entonces la mejor representación de las columnas en un espacio de dimensión 1 será:

$$D_c^{-1} F' D_f^{1/2} v_1$$

siendo v_1 el vector propio asociado al mayor valor propio de ZZ', y la mejor representación de las columnas en dimensión 2 será

$$C_c = D_c^{-1} F' D_f^{1/2} [v_1, v_2]$$

Representación conjunta

Las matrices ZZ' y $Z'Z$ tienen los mismos valores propios no nulos y sus vectores propios la siguiente relación:
si u_i es vector propio de $Z'Z$ asociado a λ_i entonces

$$Z'Zu_i = \lambda_i u_i$$

y si multiplicamos por Z entonces

$$Z(Z'Zu_i) = \lambda_i Zu_i$$

y obtenemos que Zu_i es vector propio de ZZ' asociado a λ_i. Por lo tanto

$$v_i = Zu_i \quad \forall\, i$$

Como nos interesan los vectores propios de ZZ' y de $Z'Z$ es por eso que una descomposición como la SVD puede ser interesante de utilizar.

Resumen

El análisis de correspondencias de una tabla de contingencia $I \times J$ se hace de la siguiente manera:

1. Calculamos la tabla de frecuencias relativas F.
2. Calculamos la tabla estandarizada $Z = D_f^{-1/2} F D_c^{-1/2}$.
3. Calculamos los h vectores propios ligados a los valores propios mayores (distintos de 1). Si obtenemos los vectores propios u_i de $Z'Z$, los v_i de ZZ' se obtienen por $v_i = Zu_i$. Analogamente si se obtienen los v_i de ZZ', $u_i = Z'v_i$.

   Las I filas de la matriz se representan como I puntos en $\mathbb{R}^h$ y las coordenadas de cada fila vienen dadas por

   $$C_f = D_f^{-1/2} Z A_2$$

   donde A_2 tiene en columnas los vectores propios de $Z'Z$.

   Las J filas de la matriz se representan como J puntos en $\mathbb{R}^h$ y las coordendas de cada fila vienen dadas por

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

   donde B_2 tiene en columnas los vectores propios de ZZ'.

Relación entre las dos formulaciones

Las coordenadas de las nubes de puntos proyectadas sobre los ejes factoriales son:

- El factor fila α es el vector de coordenadas de las proyecciones de los perfiles files $X^{(i)}$ sobre el eje de rango α:

$$c_\alpha(i) = \tilde{X}^{(i)} u_\alpha = \sum_{j=1}^{J} \frac{f_i^j - f_{.j}}{\sqrt{f_{.j}}} u_\alpha(j) = \sum_{j=1}^{J} \frac{f_i^j}{\sqrt{f_{.j}}} u_\alpha(j)$$

  La última igualdad se prueba en un ejercicio de práctico (¡y en realidad centrar o no centrar la matriz da lo mismo!).

- El factor columna α es el vector de coordenadas de las proyecciones de los perfiles columnas $Y^{(j)}$ sobre el eje de rango α:

$$d_\alpha(j) = \tilde{Y}^{(j)} v_\alpha = \sum_{i=1}^{I} \frac{f_j^i - f_{i.}}{\sqrt{f_{i.}}} v_\alpha(i) = \sum_{i=1}^{I} \frac{f_j^i}{\sqrt{f_{i.}}} v_\alpha(i)$$

Relación entre las dos formulaciones

Fórmulas de transición:

$c_\alpha(i) = \sum_{j=1}^{J} \frac{f_i^j}{\sqrt{f_{.j}}} u_\alpha(j) = \sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} u_\alpha(j)$ y por otro lado,como vimos en Análisis Factorial,

$$v_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{J} \frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}} u_\alpha(j)$$

entonces

$$c_\alpha(i) = \sqrt{\lambda_\alpha}\frac{v_\alpha(i)}{\sqrt{f_{i.}}}$$

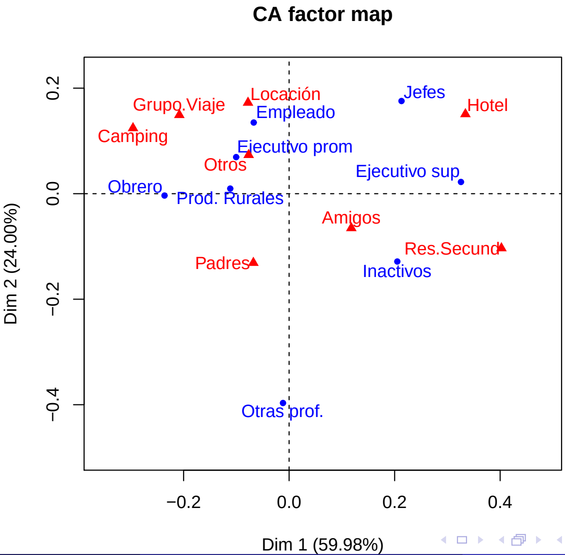
También en Análisis Factorial vimos que $u_\alpha(i) = \frac{\sqrt{f_{.j}}}{\sqrt{\lambda_\alpha}} d_\alpha(j)$, entonces

$$c_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}} d_\alpha(j)$$

Análogamente

$$d_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} c_\alpha(i)$$

```
> AFC=CA(base,ncp=5,graph=TRUE)
```



Ejemplo de ACS usando R

```
> AFC
**Results of the Correspondence Analysis (CA)**
The row variable has  8  categories; the column variable has 8 categories
The chi square of independence between the two variables is equal to 2292.148
(p-value =  0 ).
*The results are available in the following objects:

   name               description
1  "$eig"             "eigenvalues"  (valores propios de S)
2  "$col"             "results for the columns"
3  "$col$coord"       "coord. for the columns" (coord. de los puntos columnas)
4  "$col$cos2"        "cos2 for the columns"   (calidad repr. puntos columnas)
5  "$col$contrib"     "contributions of the columns" (contribucion puntos columnas)
6  "$row"             "results for the rows"
7  "$row$coord"       "coord. for the rows"
8  "$row$cos2"        "cos2 for the rows"
9  "$row$contrib"     "contributions of the rows"
10 "$call"            "summary called parameters"
11 "$call$marge.col" "weights of the columns" (pesos columnas)
12 "$call$marge.row" "weights of the rows"  (pesos filas)
```

Ejemplo de ACS usando R

```
> AFC$row$coord
                      Dim 1        Dim 2        Dim 3       Dim 4        Dim 5
Prod. Rurales   -0.11160583  0.009686625 -0.331079734 -0.05028123  0.108913985
Jefes            0.21302067  0.175665571 -0.083575888  0.01167763  0.019443713
Ejecutivo sup    0.32571537  0.022229111  0.092811557  0.02470341  0.037327118
Ejecutivo prom  -0.10038234  0.069364473  0.071450764 -0.10559460 -0.002748292
Empleado        -0.06710022  0.134872398  0.020813580  0.02593565 -0.049499681
Obrero          -0.23618313 -0.003534578  0.007116966  0.03767886  0.002723447
Otras prof.     -0.01164813 -0.396747383  0.048110957 -0.01057656  0.040091875
Inactivos        0.20505507 -0.128579628 -0.091696513 -0.01137359 -0.074098260
```

$c_\alpha = (c_\alpha(i))_{1 \leq i \leq I-1}$

```
>AFC$col$coord
                  Dim 1       Dim 2        Dim 3        Dim 4        Dim 5
Hotel        0.33415248  0.15081675 -0.099232613 -0.033916543  0.001276274
Locacion    -0.07791859  0.17251300  0.077248428  0.022111480 -0.055525029
Res.Secund   0.40241445 -0.10332602  0.233256062  0.016537122  0.048576831
Padres      -0.06774438 -0.13102386 -0.032448013 -0.008068563 -0.004162512
Amigos       0.11789513 -0.06519633 -0.011860467  0.047072701 -0.078723277
Camping     -0.29589905  0.12427663  0.066085060 -0.065485958  0.029642374
Grupo.Viaje -0.20809792  0.14919893  0.002080871  0.147259167  0.003222590
Otros       -0.07666056  0.07345940 -0.137879867  0.102860597  0.139506799
```

$d_\alpha = (d_\alpha(j))_{1 \leq j \leq J-1}$

Ejemplo de ACS usando R

```
> AFC$eig
        eigenvalue percentage of variance cumulative percentage of variance
dim 1 4.423429e-02            59.97683631                          59.97684
dim 2 1.769798e-02            23.99651566                          83.97335
dim 3 7.652040e-03            10.37532452                          94.34868
dim 4 2.240295e-03             3.03759299                          97.38627
dim 5 1.683756e-03             2.28298833                          99.66926
dim 6 2.430850e-04             0.32959658                          99.99885
dim 7 8.449133e-07             0.00114561                         100.00000
```

$(\lambda_\alpha(j))_{1\leq\alpha\leq I-1}$. Estos valores propios son todos menores que 1.

Los 3 primeros ejes captan el 94 % de la inercia total.

contribución filas Para cada eje retenido y cada nube de puntos, miramos los puntos de N^I y los de N^J que contribuyen más a la formación del eje α. Para los filas, son los puntos cuya contribución es mayor que la media $\frac{1}{I}$. La contribución se mide por

$$ctr_\alpha(i) = \frac{f_{i.}c_\alpha^2(i)}{\sum_{i=1}^{I} f_{i.}c_\alpha^2(i)} = \frac{f_{i.}c_\alpha^2(i)}{\lambda_\alpha}$$

```
> AFC$row$contrib
                    Dim 1       Dim 2      Dim 3      Dim 4       Dim 5
Prod. Rurales   0.95677882  0.01801433 48.6726341  3.8344548 23.93783931
Jefes           9.82974105 16.70733721  8.7466638  0.5832593  2.15147752
Ejecutivo sup  35.65269429  0.41504523 16.7340797  4.0493338 12.30111902
Ejecutivo prom  3.15032176  3.75967087  9.2264771 68.8299839  0.06203631
Empleado        0.97335351  9.82888382  0.5413749  2.8712499 13.91577299
Obrero         37.36664509  0.02091687  0.1961363 18.7774144  0.13052815
Otras prof.     0.01979782 57.40745717  1.9524263  0.3222906  6.16165515
Inactivos      12.05066765 11.84267451 13.9302079  0.7320133 41.33957155
```

$ctr_\alpha(i) \geq 1/8 = 12{,}5\,\%$. El primer eje opone los ejecutivos superiores a los obreros en cuanto al perfil de sus lugares de vacaciones.

contribución columnas

La contribución se mide por

$$ctr_\alpha(j) = \frac{f_{.j} d_\alpha^2(j)}{\sum_{j=1}^{J} f_{.j} d_\alpha^2(j)} = \frac{f_{.j} d_\alpha^2(j)}{\lambda_\alpha}$$

Para las columnas retenemos los puntos cuya contribución es mayor que la media $\frac{1}{J}$.

```
> AFC$col$contrib
                 Dim 1     Dim 2       Dim 3      Dim 4       Dim 5
Hotel       36.8332935 18.753645 18.777677965  7.4925265  0.01411624
Locacion     1.4913757 18.271910  8.473563758  2.3713439 19.89584445
Res.Secund  26.2914900  4.332343 51.064114529  0.8766816 10.06482646
Padres       4.2662989 39.887876  5.657998150  1.1949537  0.42315126
Amigos       1.8785006  1.435824  0.109901862  5.9130533 22.00420966
Camping     24.5582852 10.827425  7.081077367 23.7498699  6.47465617
Grupo.Viaje  4.2083884  5.406885  0.002432498 41.6101211  0.02651369
Otros        0.4723676  1.084092  8.833233871 16.7914500 41.09668206
```

$ctr_\alpha(j) \geq 1/8 = 12{,}5\,\%$. El primer eje opone las residencias secundarias o los hoteles a los campings.

Conclusión: El eje 1 opone los obreros, que frecuentemente van a los campings, a los ejecutivos superiores que van a residencias secundarias u hoteles.

1. Como en ACP, si un punto está muy alejado del resto de la nube y determina por sí solo la dirección del primer eje, se puede hacer de vuelta el análisis y tratarlo como punto suplementario (será representado pero no contribuirá a la formación de los ejes).
2. Si los puntos i y j contribuyen a la inercia de un solo eje, entonces:
   - Si $c_\alpha(i)d_\alpha(j) > 0$ la celda ij está más "cargada" que en la tabla teórica ($f_{ij} > p_{ij}$)).
   - Si $c_\alpha(i)d_\alpha(j) < 0$ la celda ij está más "cargada" que en la tabla teórica ($f_{ij} < p_{ij}$)).

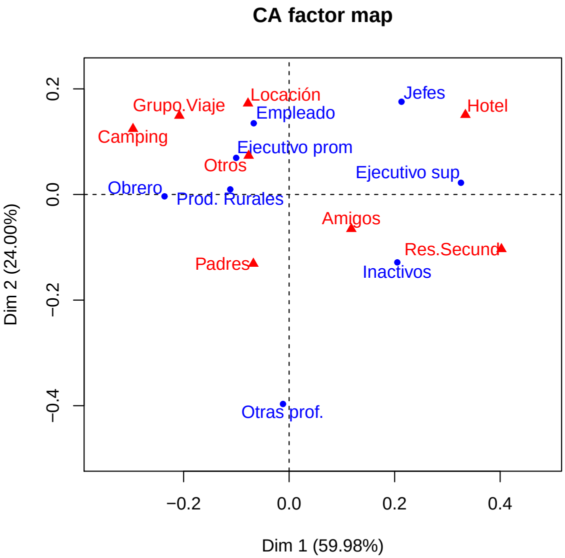
   Por ejemplo otras.prof*padres está más cargada que bajo la hipotesis de independencia, estas dos modalidades se atraen
3. Para interpretar bien los puntos, deben estar bien proyectados, y el indicador es el $\cos^2$.

Interpretación 2do eje

```
> AFC$row$contrib
                     Dim 1       Dim 2      Dim 3      Dim 4       Dim 5
Prod. Rurales   0.95677882  0.01801433 48.6726341  3.8344548 23.93783931
Jefes           9.82974105 16.70733721  8.7466638  0.5832593  2.15147752
Ejecutivo sup  35.65269429  0.41504523 16.7340797  4.0493338 12.30111902
Ejecutivo prom  3.15032176  3.75967087  9.2264771 68.8299839  0.06203631
Empleado        0.97335351  9.82888382  0.5413749  2.8712499 13.91577299
Obrero         37.36664509  0.02091687  0.1961363 18.7774144  0.13052815
Otras prof.     0.01979782 57.40745717  1.9524263  0.3222906  6.16165515
Inactivos      12.05066765 11.84267451 13.9302079  0.7320133 41.33957155
```

```
> AFC$col$contrib
                 Dim 1     Dim 2        Dim 3      Dim 4       Dim 5
Hotel        36.8332935 18.753645 18.777677965  7.4925265  0.01411624
Locacion      1.4913757 18.271910  8.473563758  2.3713439 19.89584445
Res.Secund   26.2914900  4.332343 51.064114529  0.8766816 10.06482646
Padres        4.2662989 39.887876  5.657998150  1.1949537  0.42315126
Amigos        1.8785006  1.435824  0.109901862  5.9130533 22.00420966
Camping      24.5582852 10.827425  7.081077367 23.7498699  6.47465617
Grupo.Viaje   4.2083884  5.406885  0.002432498 41.6101211  0.02651369
Otros         0.4723676  1.084092  8.833233871 16.7914500 41.09668206
```

El segundo eje es característico de los otras profesiones que van más a la casa de los padres que las otras categorías socioprofesionales.


CA factor map
Dim 1 (59.98%)
Dim 2 (24.00%)
−0.2
0.0
0.2
0.4
−0.4
Jefes
Hotel
Locación
Empleado
Grupo.Viaje
Camping
Ejecutivo prom
Otros
Ejecutivo sup
Obrero
Prod. Rurales
Amigos
Res.Secund
Inactivos
Padres
Otras prof.

Interpretación

```
> AFC$row$cos2
                      Dim 1        Dim 2        Dim 3        Dim 4        Dim 5
Prod. Rurales   0.0903790916 0.0006808308 0.7953520482 0.0183445078 0.0860719225
Jefes           0.5365926163 0.3649005228 0.0825967931 0.0016125393 0.0044705357
Ejecutivo sup   0.9044719383 0.0042127220 0.0734383161 0.0052027427 0.0118786613
Ejecutivo prom 0.3233486333 0.1543940559 0.1638212209 0.3577995679 0.0002423719
Empleado        0.1654966075 0.6686322656 0.0159233608 0.0247249553 0.0900629089
Obrero          0.9732215612 0.0002179661 0.0008836976 0.0247690548 0.0001294053
Otras prof.     0.0008396854 0.9741640461 0.0143249113 0.0006922976 0.0099475626
Inactivos       0.5790524131 0.2276780652 0.1157930516 0.0017814424 0.0756123744
```

```
> AFC$col$cos2
                 Dim 1      Dim 2        Dim 3       Dim 4        Dim 5
Hotel        0.76763693 0.15637439 6.769792e-02 0.007908418 1.119836e-05
Locacion     0.13234572 0.64874111 1.300790e-01 0.010657697 6.720558e-02
Res.Secund   0.70511343 0.04648697 2.369069e-01 0.001190780 1.027473e-02
Padres       0.20046843 0.74989464 4.599130e-02 0.002843753 7.568512e-04
Amigos       0.51888221 0.15868040 5.251467e-03 0.082720902 2.313574e-01
Camping      0.77781583 0.13720456 3.879683e-02 0.038096581 7.805767e-03
Grupo.Viaje 0.48249604 0.24802171 4.824459e-05 0.241614476 1.157093e-04
Otros        0.09456419 0.08683155 3.059035e-01 0.170247535 3.131652e-01
```

Referencias

- D. Peña, *Analisis de Datos Multivariantes*, Mac Graw Hill, 2002.
- A. I. Izenman, *Modern Multivariate Statistical Techniques*, Springer, 2008.
- Laurence Reboul, Transparences de cours *Analyse de données*, Université Aix-Marseille.