Índice general

Tema 1

Análisis de Correspondencias Simple

El Análisis de Correspondencias es una técnica de análisis de datos desde un punto de vista gráfico. Esta técnica se centra en variables cualitativas, nominales u ordinales. Se trata de representar gráficamente tablas de datos, de manera que el analista pueda visualizar las relaciones de dependencia e independencia entre dichas de variables.

El Análisis de Correspondencias simple se utiliza principalmente para representar gráficamente datos que vienen dados en tablas de contingencia, es decir, tablas de doble entrada dónde tenemos dos variables cualitativas, nominales u ordinales. En las tablas de contingencia las categorías de una variable vendrán dadas en las filas y las de la otra vendrán dadas en las columnas. En el centro de la tabla aparecen las frecuencias, k_{ij}, es decir, el número de individuos que presentan la modalidad i de la variable I, y la modalidad j de la variable J.

Ejemplo: *La siguiente tabla muestra los datos recogidos de una muestra de 100 personas que se han clasificado según si padecen cáncer de pulmón o no (variable I, filas), y si son fumadores o no (variable J, columnas)*

	Si	**No**
Si	32	18
No	12	38

En este ejemplo hay 18 personas de la muestra que padecen cáncer de pulmón y no fuman. Es decir k_{22} =18.

En la mayoría de aplicaciones del Análisis de Correspondencias, las tablas de contingencia que estamos tratando tienen multitud de filas o columnas, por lo que es necesario reducir la dimensión para poder visualizar e interpretar mejor la información contenida en los datos. La esencia de este tipo de análisis es la identificación de subespacios de pocas dimensiones dónde se puedan proyectar simultáneamente los puntos filas y los puntos columna, de manera que sea más fácil extraer conclusiones sobre las relaciones entre las variables. En definitiva, se trata de resumir una gran cantidad de datos en un número reducido de dimensiones, en este sentido, su objetivo es parecido al de los métodos factoriales pero en el caso de variables categóricas u ordinales.

Por último, en el Análisis de Correspondencias el estudio de filas y columnas es similar, el Análisis de Correspondencias las trata de forma simétrica.

1.1. Nubes de puntos $\mathbb{R}^p$ y en $\mathbb{R}^n$: Perfiles fila y perfiles columna

Partimos de una tabla de contingencia que clasifica a un grupo K de individuos según dos variables: La variable I con n modalidades y la variable J con p modalidades.

		Modalidad J					
		1	$\cdots$	j	$\cdots$	p	Total
Modalidad I	1			$\vdots$			
	$\vdots$			$\vdots$			
	i	$\cdots$	$\cdots$	k_{ij}	$\cdots$	$\cdots$	$k_{i\cdot}$
	$\vdots$			$\vdots$			
	n			$\vdots$			
	Total			$k_{\cdot j}$			K

Consideramos ahora la tabla dónde hemos dividido cada una de las frecuencias k_{ij} por el total de individuos de la muestra K, es decir, es la tabla que contiene las frecuencias relativas. A esta tabla se la conoce también como tabla (matriz) de correspondencias.

		Modalidad J					
		1	$\cdots$	j	$\cdots$	p	Total
Modalidad I	1			$\vdots$			
	$\vdots$			$\vdots$			
	i	$\cdots$	$\cdots$	f_{ij}	$\cdots$	$\cdots$	$f_{i\cdot}$
	$\vdots$			$\vdots$			
	n			$\vdots$			
	Total			$f_{\cdot j}$			1

siendo,

$$K = \sum_{ij} k_{ij} \tag{1.1}$$

$$f_{ij} = \frac{k_{ij}}{K} \tag{1.2}$$

$$f_{i\cdot} = \sum_{j=1}^{p} f_{ij} = \sum_{j=1}^{p} \frac{k_{ij}}{K} \tag{1.3}$$

$$f_{\cdot j} = \sum_{i=1}^{n} f_{ij} = \sum_{i=1}^{n} \frac{k_{ij}}{K} \tag{1.4}$$

El Análisis de Correspondencias trabaja con las frecuencias relativas, no con las absolutas. Las frecuencias relativas nos permiten comparar mejor las diferencias existentes entre las categorías de

las variables. Por ejemplo, supongamos que tenemos los siguientes puntos correspondientes a dos filas (dos categorías de la variable I):

$$(783; 1114; 387; 4052; 497; 1464; 525; 387)$$

$$(65; 43; 21; 294; 79; 57; 18; 6)$$

El primer punto constituye el reparto de 9209 individuos en 8 categorías de la variable J. El segundo punto muestra como se reparten 583 individuos en esas 8 categorías. Si para entender mejor las diferencias de comportamiento entre las dos categorías de la variable I, medimos la distancia entre estos dos puntos usando la distancia euclídea, el valor de dicha distancia confirmaría la diferencia en la cantidad de individuos que hay en ambas categorías pero no permitiría ver la diferencia entre los tipos de comportamiento que reflejan estas filas, es decir, las diferencias en función de cómo se reparten esos individuos en las distintas categorías de la variable J. Esto es debido a que las distancias entre filas (o columnas) están condicionadas por sus masas (cantidad de individuos que hay en cada fila o columna). Para hacer las distintas categorías comparables surgen los conceptos de **perfiles-fila** y **perfiles-columna**, dónde se consideran las frecuencias relativas condicionadas de filas y columnas, respectivamente.

En una tabla de contingencia con n filas y p columnas, los **perfiles-fila** y **perfiles-columna** se definen así:

Perfiles-fila $\left\{\left(\frac{f_{ij}}{f_{i\cdot}}\right); j = 1, \cdots, p\right\}_i$, para $i = 1, \cdots, n$. En este caso de lo que se trata es de dividir las frecuencias de cada fila por el total de la fila, $f_{i\cdot}$

Perfiles-columna $\left\{\left(\frac{f_{ij}}{f_{\cdot j}}\right); i = 1, \cdots, n\right\}_j$, para $j = 1, \cdots, p$. Aquí dividimos las frecuencias de las columnas por el total de cada columna, $f_{\cdot j}$

Como podemos ver cada fila o columna de la tabla está afectada por un peso proporcional a su importancia en el conjunto de datos, $f_{i\cdot}$ y $f_{\cdot j}$ respectivamente. Este peso es conocido como *masa*. De esta manera, podemos expresar las tablas de contingencia en términos de los perfiles fila o de los perfiles columna como sigue:

1. Tabla de Perfiles-Fila

	1	$\cdots$	j	$\cdots$	p	Masas
1	$\frac{f_{11}}{f_{1\cdot}}$	$\cdots$	$\frac{f_{1j}}{f_{1\cdot}}$	$\cdots$	$\frac{f_{1p}}{f_{1\cdot}}$	$f_{1\cdot}$
$\vdots$			$\vdots$			$\vdots$
i	$\frac{f_{i1}}{f_{i\cdot}}$	$\cdots$	$\frac{f_{ij}}{f_{i\cdot}}$	$\cdots$	$\frac{f_{ip}}{f_{i\cdot}}$	$f_{i\cdot}$
$\vdots$			$\vdots$			$\vdots$
n	$\frac{f_{n1}}{f_{n\cdot}}$	$\cdots$	$\frac{f_{nj}}{f_{n\cdot}}$	$\cdots$	$\frac{f_{np}}{f_{n\cdot}}$	$f_{n\cdot}$

2. Tabla de Perfiles-Columna

	1	$\cdots$	j	$\cdots$	p
1	$\frac{f_{11}}{f_{\cdot 1}}$	$\cdots$	$\frac{f_{1j}}{f_{\cdot j}}$	$\cdots$	$\frac{f_{1p}}{f_{\cdot p}}$
$\vdots$			$\vdots$		
i	$\frac{f_{i1}}{f_{\cdot 1}}$	$\cdots$	$\frac{f_{ij}}{f_{\cdot j}}$	$\cdots$	$\frac{f_{ip}}{f_{\cdot p}}$
$\vdots$			$\vdots$		
n	$\frac{f_{n1}}{f_{\cdot 1}}$	$\cdots$	$\frac{f_{nj}}{f_{\cdot j}}$	$\cdots$	$\frac{f_{np}}{f_{\cdot p}}$
Masas	$f_{\cdot 1}$	$\cdots$	$f_{\cdot j}$	$\cdots$	$f_{\cdot p}$

Los perfiles se pueden considerar como coordenadas en un espacio multidimensional. Así pues tenemos dos nubes de puntos: una constituida por n puntos en $\mathbb{R}^p$ de coordenadas:

$$\left(\frac{f_{i1}}{f_{i\cdot}};\frac{f_{i2}}{f_{i\cdot}};\cdots,\frac{f_{ij}}{f_{i\cdot}};\cdots;\frac{f_{ip}}{f_{i\cdot}}\right);\ i=1,\cdots,n \tag{1.5}$$

Y otra constituida por p puntos en $\mathbb{R}^n$ de coordenadas:

$$\left(\frac{f_{1j}}{f_{\cdot j}};\frac{f_{2j}}{f_{\cdot j}};\cdots,\frac{f_{ij}}{f_{\cdot j}};\cdots;\frac{f_{nj}}{f_{\cdot j}}\right);\ j=1,\cdots,p \tag{1.6}$$

Notas:

1. La masa también afecta al perfil medio de filas y columnas, es decir, afecta al **centro de gravedad** o **centroide** desplazándolo, de manera que no se sitúa en el centro greográfico de la nube de puntos, si no que tiende a situarse cerca de los puntos de mayor masa.

2. El conjunto de masas $f_{i\cdot}$, $i=1,\cdots,n$, corresponde con la columna promedio, o lo que es lo mismo el centro de gravedad de las columnas.

3. El conjunto de masas $f_{\cdot j}$, $j=1,\cdots,p$, corresponde con la fila promedio, o lo que es lo mismo el centro de gravedad de las filas.

Por otra lado, los n puntos-fila anteriores (o perfiles-fila) están situados en realidad en un subespacio $p-1$ dimensional de $\mathbb{R}^p$, ya que existe entre las coordenadas de cualquiera de ellos la relación baricéntrica:

$$\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}}\right)=1\ \ ;\ \ i=1,\cdots,n$$

Y lo mismo cabe decir de los p puntos-columna (perfiles-columna)

$$\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{\cdot j}}\right)=1\ \ ;\ \ j=1,\cdots,p$$

de modo que estos p puntos en $\mathbb{R}^n$, lo están en un subespacio $(n-1)$ dimensional de $\mathbb{R}^n$.

1.2. Distancias en $\mathbb{R}^p$ y $\mathbb{R}^n$: Distancia chi-cuadrado

En Análisis Factorial se utiliza la distancia euclídea para medir la distancia entre los puntos-fila y los puntos-columna de $\mathbb{R}^p$ y de $\mathbb{R}^n$ respectivamente.

En el caso del Análisis de Correspondencias, consideramos los perfiles-fila y perfiles-columna, que siguen siendo puntos de $\mathbb{R}^p$ y $\mathbb{R}^n$ respectivamente, pero que están afectados por sus masas. Por este motivo, la forma de medir distancias entre los puntos no puede ser la distancia euclídea. La distancia con la que se trabaja en el análisis de correspondencias clásico es una distancia euclídea ponderada, conocida como chi-cuadrado (χ^2). Estas distancias son las que el análisis de correspondencias trata de reproducir en sus representaciones gráficas.

Las distancias χ^2 de los perfiles fila y columna vienen dadas por las siguientes expresiones:

1. Distancia entre perfiles fila de los individuos i e i'

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 \frac{1}{f_{\cdot j}} \tag{1.7}$$

2. Distancia entre perfiles columna de los individuos j e j'

$$d^2(j,j') = \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2 \frac{1}{f_{i\cdot}} \tag{1.8}$$

Como podemos ver cada sumando está ponderado por el inverso de la masa de la columna o de la fila, respectivamente, para no favorecer a aquellas columnas o filas que tienen más masa.

1.2.1. Principio de Equivalencia Distribucional

La distancia χ^2, tiene ciertas propiedades, entre ellas una fundamental para el análisis de correspondencias, lo que se conoce como **Principio de Equivalencia Distribucional**. Dicho principio nos dice que si dos puntos-fila en $\mathbb{R}^p$ se agrupan en uno solo con masa la suma de las masas de ambos, entonces son invariantes las distancias entre los demás puntos en $\mathbb{R}^p$ y en $\mathbb{R}^n$. (Y lo mismo si se agrupan puntos-columna en $\mathbb{R}^n$).

Este principio tiene bastante trascendencia en el análisis de correspondencia de una tabla de contingencia. Cuando se establecen las clases de las características I y J que definen la Tabla, hay un grado de arbitrariedad más o menos grande, de modo que podrían en un principio definirse clases en I o/y en J muy próximas entre sí. Por tanto, algunas de las clases que se definieron en un principio podrían agruparse en una sola por diversos motivos. En estos casos, deberíamos usar una distancia lo menos sensible ante esos agrupamientos de clases, en el sentido de que no se alterasen las distancias entre puntos ya calculadas y que son entre puntos a los que no afecta en principio esas agrupaciones. La distancia chi-cuadrado dada por [1.7] y [1.8] cumple estas cualidades.

Nota: *La distancia χ^2 es la natural considerando el espacio de puntos de los perfiles fila (o el de las columnas) como un espacio euclídeo ponderado con métrica definida por la matriz $\mathbb{D}_c^{-1} = \operatorname{diag}\left(\frac{1}{f_{\cdot j}}\right)$, estando los puntos del correspondiente espacio afectados por las masas (ponderaciones) asociadas a cada punto, y antes indicadas.*

1.2.2. Inercia de la nube de puntos

La inercia de la nube de puntos-fila (perfiles-fila) respecto a su centro de gravedad, G_I, es una medida de dispersión de la nube de puntos. Se calcula como la suma ponderada de las distancias entre los punto-fila (perfiles-fila) y su centro de gravedad, usando como ponderación la masa de cada perfil-fila y como métrica la distancia χ^2:

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

De igual forma se obtiene la inercia de la nube de puntos-columna respecto a su centro de gravedad, G_J:

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

Como puede observarse, las inercias de las nubes de los puntos-fila y los puntos-columna son iguales. Además estas inercias coinciden con el estadístico χ^2 de Pearson dividido por el total de individuos K. Recordemos que el estadístico χ^2 de Pearson nos sirve para contrastar si existe independencia entre dos variables categóricas, por tanto, la inercia (o varianza/heterogeneidad de los datos) será mayor cuanto más se alejen los puntos de su centro de gravedad o visto de otro modo, cuanto más se alejen las frecuencias de la condición de independencia.

1.3. Ajustes a la nube de perfiles en $\mathbb{R}^p$ y en $\mathbb{R}^n$

Como mencionamos al inicio del tema, en la mayoría de aplicaciones del Análisis de Correspondencias, las tablas de contingencia que estamos tratando tienen muchas filas o columnas. Se hace necesario reducir la dimensionalidad del problema. La esencia del Análisis de Correspondencias es la identificación de espacios de pocas dimensiones dónde poder proyectar esos perfiles-fila y perfiles-columna con la menor pérdida de información. Se conocen como subespacios óptimos. En esta sección tratamos el ajuste de las nubes de puntos, perfiles-fila y perfiles-columna, a los correspondientes subespacios óptimos.

En principio podríamos pensar en aplicar directamente la metodología que se utiliza en el Análisis de Componentes Principales, pero la distancia entre perfiles-fila (o perfiles-columna) no es una suma de cuadrados simplemente, sino que esa suma está ponderada. En el Análisis de Correspondencias simple los espacios $\mathbb{R}^p$ y $\mathbb{R}^n$ son espacios euclídeos ponderados. Por tanto, para poder aplicar lo que se ha visto en el Análisis de Componentes Principales debemos hacer una tranformación sobre las nubes de puntos, como vemos a continuación:

1.3.1. Ajuste, en $\mathbb{R}^p$, de la nube de perfiles-fila

Sabemos que los n puntos, perfiles-fila, en $\mathbb{R}^p$ que constituyen la nube de puntos respectiva, son aquellos cuyas coordenadas son:

$$\left\{\frac{f_{ij}}{f_{i\cdot}}\ ;\ j=1,\cdots,p\right\}\ ;\ i=1,\cdots,n$$

Estos puntos se pueden manejar a través de las siguientes coordenadas transformadas:

$$\left\{\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}\ ;\ j=1,\cdots,p\right\}\ ;\ i=1,\cdots,n$$

Dichas coordenadas son tales que la distancia euclídea al cuadrado entre dos de esos puntos coincide con la distancia χ^2 entre ellos, como podemos ver a continuación.

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}\cdot\sqrt{f_{\cdot j}}} - \frac{f_{i'j}}{f_{i'\cdot}\cdot\sqrt{f_{\cdot j}}} \right)^2 = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 \frac{1}{f_{\cdot j}} \tag{1.9}$$

Recordemos que la distancia χ^2 es la distancia según la métrica de la que anteriormente hemos dotado al espacio de los perfiles fila $\mathbb{R}^p$.

Pero ¿qué significa manejar los perfiles-fila en $\mathbb{R}^p$, por las coordenadas $\left\{ \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \; ; \mathrm{J} = 1, \cdots, p \right\}$?. Lo que se hace al pasar de las coordenadas $\frac{f_{ij}}{f_{i\cdot}}$ a las $\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}$, no es otra cosa que un cambio de escala en los ejes de $\mathbb{R}^p$ $\left(f_{ij} \longrightarrow \frac{f_{ij}}{\sqrt{f_{\cdot j}}} \right)$.

Por tanto el problema que se planteaba para poder aplicar el Análisis de Componentes Principales a la nube de perfiles-fila en $\mathbb{R}^p$, queda resuelto si manejamos dichos perfiles a través de las coordenadas transformadas, ya que la distancia euclídea entre estos puntos transformados equivale a la distancia χ^2 entre perfiles-fila.

A continuación, aplicamos el Análisis de Componentes Principales a la nube de puntos-fila dados por las coordenadas:

$$\left\{ \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \; ; \; j = 1, \cdots, p \right\} \; ; \; i = 1, \ldots, n \tag{1.10}$$

en $\mathbb{R}^p$, dotado de la distancia euclídea habitual.

Para aplicar el Análisis de Componentes Principales, como previamente hemos dicho, necesitamos trasladar el origen del sistema de referencia al centro de gravedad de la respectiva nube de puntos en $\mathbb{R}^p$ (recordemos que, a su vez, el Análisis de Componentes Principales se resolvía aplicando el Análisis Factorial General, por lo tanto, ajustando subespacios vectoriales, lo que requiere obviamente el centramiento en media).

¿Cuál es el centro de gravedad de la nube de puntos [1.10]?. Como los puntos están afectados de una masa $f_{i\cdot}$,

$$\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \longrightarrow \frac{1}{f_{i\cdot}} \frac{f_{ij}}{\sqrt{f_{\cdot j}}}$$

dicho centro de gravedad viene dado por:

$$\left\{ \sum_{i=1}^{n} f_{i\cdot} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \right) = \sqrt{f_{\cdot j}} \;\; ; \;\; j = 1, \cdots, p \right\} \tag{1.11}$$

Por tanto, al trasladar el origen al centro de gravedad, las coordenadas de los puntos (perfiles-fila) en $\mathbb{R}^p$, pasan a ser

$$\left\{ \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \;\; ; \;\; j = 1, \cdots, p \right\} \;\; ; \;\; i = 1, \cdots, n \tag{1.12}$$

Siguiendo la metodología del Análisis Factorial General al proyectar esta nube de puntos transformados dada por [1.12] sobre el subespacio vectorial definido por el vector unitario u, $u'u = 1$ y $\sum_{j=1}^{p} u_j^2 = 1$, uno de los puntos de dicha nube, por ejemplo el i-ésimo, con $i = 1, \cdots, n$, proporcionaría una proyección $\widehat{\Psi}_i$ dada por:

$$\widehat{\Psi}_i = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) u_j \tag{1.13}$$

en donde u_j es la j-ésima componente del vector unitario u en $\mathbb{R}^p$.

Recordemos que el punto i-ésimo de la nube está dotado de una masa $f_{i\cdot}$. Por tanto, la **inercia** de la nube [1.12], es decir, la suma ponderada por $f_{i\cdot}$ de todas las proyecciones al cuadrado, valdrá:

$$\sum_{i=1}^{n} f_{i\cdot} \widehat{\Psi}_i^2 \tag{1.14}$$

Por tanto, el subespacio definido por el vector u tal que se verifica

$$\operatorname*{Max}_{u} \sum_{i=1}^{n} f_{i\cdot} \widehat{\Psi}_i^2$$

define la primera Componente Principal según la metodología del Análisis Factorial General.

Continuando dicha metodología, se llega a la conclusión de que la matriz de covarianzas a diagonalizar para obtener todas las componentes principales es la de término general

$$t_{jj'} = \sum_{i=1}^{n} f_{i\cdot} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) \left(\frac{f_{ij'}}{f_{i\cdot}\sqrt{f_{\cdot j'}}} - \sqrt{f_{\cdot j'}} \right) \tag{1.15}$$

que define la matriz $p \times p$, que notamos por T.

Notas:

1. Puede comprobarse que esta matriz T puede ponerse como $T = X^{'}X$, donde X es una matriz $n \times p$, con término general x_{ij} dado por

$$x_{ij} = \frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \tag{1.16}$$

Finalmente, una vez calculados los autovalores λ_α de la matriz simétrica T dada por [1.15], la coordenadas-proyección de un punto i sobre el eje factorial α-ésimo u_α vendrá dada por:

$$\widehat{\Psi}_{\alpha i} = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) u_{\alpha j} \tag{1.17}$$

en donde u_α es el autovector de T asociado a λ_α, con $\alpha = 1, \cdots, p$

1.3.2. Simplificación en los cálculos de las proyecciones $\widehat{\Psi}_{\alpha i}$

Tenemos, por tanto, que calcular los autovalores de la matriz $T \equiv X'_{p\times n} X_{n\times p}$. Puede comprobarse que:

1. El vector $u_p = \left(\sqrt{f_{\cdot 1}}, \cdots, \sqrt{f_{\cdot j}}, \cdots, \sqrt{f_{\cdot p}}\right)'$ es un vector propio de T, con respecto al autovalor 0.

   Esto quiere decir que $(T - \lambda I)u = 0$ es cierto para

   $$T_{p\times p} u_{p(p\times 1)} = 0_{p\times 1}$$

   Es decir,

   $$T = \left(\sqrt{f_{\cdot 1}}, \cdots, \sqrt{f_{\cdot j}}, \cdots, \sqrt{f_{\cdot p}}\right)' = 0_{p\times 1}$$

2. El hecho de que $u_p = \left(\sqrt{f_{\cdot j}}\ ;\ j = 1, \cdot, p\right)$ sea autovector de T trae consecuencias interesantes que permiten simplificar los cálculos para obtener las proyecciones de los puntos sobre los ejes factoriales. En primer lugar, podemos ver que dado cualquier otro de los autovectores u_α, y denotando por $u_{\alpha j}$ a su j-ésima componente, se verifica:

   $$\sum_{j=1}^{p} u_{\alpha j}\sqrt{f_{\cdot j}} = 0 \tag{1.18}$$

   en virtud de la ortogonalidad entre autovectores.

   En segundo lugar, si consideramos la proyección del punto i-ésimo de la nube sobre el eje α, $\widehat{\Psi}_{\alpha i}$, dada por [1.17], esta expresión se puede simplificar. En efecto

   $$\widehat{\Psi}_{\alpha i} = \sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}}\right)u_{\alpha j} = \sum_{j=1}^{p}\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}u_{\alpha j} - \sum_{j=1}^{p}\sqrt{f_{\cdot j}}u_{\alpha j}$$

   de donde

   $$\widehat{\Psi}_{\alpha i} = \sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}\right)u_{\alpha j} \tag{1.19}$$

3. Por otra parte puede comprobarse lo siguiente: si u_α es autovector de T, distinto del antes denotado u_p, lo es también de $T^* = X^{*'}X^*$, donde

   $$X^* = (x^*_{ij}) \qquad x^*_{ij} = \frac{f_{ij}}{\sqrt{f_{i\cdot}f_{\cdot j}}} \tag{1.20}$$

   y respecto del mismo autovalor.

   En este caso, X^* no es centrada en contraposición de la X que lo era. Pero es más fácil realizar el análisis de correspondencias sobre $T^* = X^{*'}X^*$ que sobre $T = X'X$.

   Nota: Debemos de tener encuenta, no obstante, que esto es cierto para todos los autovectores excepto para el autovector respecto del autovalor 0 de T, u_p. Este autovector u_p lo es también de la nueva matriz, T^*, pero lo es respecto del autovalor 1.

1.3.3. Ajuste en $\mathbb{R}^n$, de la nube de perfiles-columna

Siguiendo el método usado para la nube de perfiles-fila en $\mathbb{R}^p$, basta realizar una permutación de los índices i, j a todo lo que se ha visto en el ajuste en $\mathbb{R}^p$. Esto es así en el Análisis de Correspondencias Simple pues estamos aplicando técnicas factoriales (el Análisis Factorial General o el Análisis de Componentes Principales) a una matriz inicial de datos que constituye, estadísticamente hablando, una tabla de contingencia, en la que el papel de filas y columnas es intercambiable, al contrario que ocurre en otras tablas, como por ejemplo en las tablas de medidas (variables-observaciones) a las que les hemos aplicado el Análisis de Componentes Principales anteriormente.

Nota: Decir que los índices son intercambiables, no quiere decir que al hacer el ajuste en $\mathbb{R}^n$ a los perfiles-columna, obtengamos el mismo ajuste. Lo que queremos decir es que, técnicamente hablando, el ajuste en $\mathbb{R}^n$ se puede obtener intercambiando i, j simplemente.

Ahora consideramos los puntos j de la nube en $\mathbb{R}^n$, que tienen por coordenadas:

$$\left\{ \frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} \; i = 1 \cdots, n \; ; \; f_{\cdot j} \right\} \quad j = 1, \cdots, p$$

en donde $f_{\cdot j}$ es la masa de cada punto.

El centro de gravedad de los p puntos anteriores es el punto de coordenadas $\left(\sqrt{f_{i\cdot}} \; i = 1, \cdots, n\right)$. Este es el centro de gravedad de la nube en $\mathbb{R}^n$. A partir de esta nube y centro de gravedad todo el desarrollo es paralelo al realizado en el caso de la nube en $\mathbb{R}^p$.

En concreto el Análisis Factorial General (o el Análisis de Componentes Principales si se prefiere) se aplicará en este caso a una matriz S (que antes habíamos denotado por T), cuyos autovalores y autovectores resuelven el problema. Esta matriz S tendrá como elemento genérico

$$s_{ii'} = \sum_{j=1}^{p} f_{\cdot j} \left[\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} - \sqrt{f_{i\cdot}} \right] \left[\frac{f_{i'j}}{f_{\cdot j}\sqrt{f_{i'\cdot}}} - \sqrt{f_{i'\cdot}} \right] \quad i, i^{'} = 1, \cdots, n$$

S es, evidentemente, de dimensión $n \times n$, mientras que T lo era $p \times p$.

Análogamente al caso de los perfiles fila, podemos escribir S como producto de la forma $X^{'}X$, que ahora vendrá dada en términos de una matriz análoga a la que aparece en [1.16] pero permutando i por j. Si llamamos a esta matriz análoga X^*, y por otro lado, considerando $XX^{'}$ en lugar de $X^{'}X$ (según el Análisis Factorial General y/o el Análisis de Componentes Principales), tenemos:

$$S_{n\times n} = X^*_{n\times p} X^{*'}_{p\times n}$$

Igual que en el caso de la nube en $\mathbb{R}^p$, aquí puede simplificarse el cálculo de autovalores y autovectores, porque, como es fácil comprobar, podemos calcularlo sobre una matriz no centrada, S^* más fácil de manejar, que tiene los mismos autovalores y autovectores que la matriz S :

$$S^* = X^{**} X^{**'}$$

en donde $X^{**} = (x^{**}_{ij})$, siendo

$$x^{**}_{ij} = \frac{f_{ij}}{\sqrt{f_{i\cdot} f_{\cdot j}}}$$

En función de estas consideraciones, se define la Conclusión Final sobre el ajuste en $\mathbb{R}^n$ siguiente:

Sean $\widehat{\Phi}_{\alpha j}$ las coordenadas-proyección sobre el α-ésimo eje v_α en $\mathbb{R}^n$, del punto-perfil-columna j-ésimo. Entonces:

$$\widehat{\Phi}_{\alpha j} = \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} \right) v_{\alpha i} \tag{1.21}$$

en donde $v_{\alpha i}$ es la i-ésima componente del autovector v_α de la matriz S, siendo v_α el α-ésimo vector unitario factorial en $\mathbb{R}^n$ cuyo soporte es el α-ésimo eje factorial en $\mathbb{R}^n$.

1.4. Relaciones entre las nubes ajustadas en $\mathbb{R}^p$ y en $\mathbb{R}^n$

En la sección anterior hemos ajustado a las nubes de puntos-fila y punto-columna, en $\mathbb{R}^p$ y en $\mathbb{R}^n$ respectivamente, los respectivos subespacios óptimos, usando la teoría general del Análisis Factorial General. Ahora, siguiendo la misma metodología, vamos a analizar cómo se formulan en el Análisis de Correspondencias, las relaciones entre los subespacios ajustados y las subsiguientes relaciones entre las coordenadas de los puntos-fila y puntos-columna cuando éstos se refieren al sistema de referencia dado por los ejes factoriales en ambos subespacios ajustados.

1.4.1. Relaciones generales entre los dos espacios ajustados en $\mathbb{R}^p$ y en $\mathbb{R}^n$

Cuando se analizó este tipo de relación en el Análisis Factorial General, se obtuvo el siguiente resultado general:

1. Las dos nubes ajustadas (perfiles-fila en $\mathbb{R}^p$; perfiles-columna en $\mathbb{R}^n$), se definen a partir, respectivamente, de las matrices $(X^{'}X)_{p\times p}$ y $(XX^{'})_{n\times n}$
2. Sean $(\lambda_\alpha, u_\alpha)$ y (μ_α, v_α) las parejas autovalores-autovectores de $X^{'}X$ y $XX^{'}$ respectivamente. Recordemos del Análisis Factorial General que $\lambda_\alpha = \mu_\alpha$, para todos aquellos autovalores que no son nulos.
3. Los autovectores u_α y v_α, que constituyen, respectivamente, vectores unitarios cuyos soportes son los ejes factoriales en $\mathbb{R}^p$ y en $\mathbb{R}^n$, están relacionados entre sí, mediante las relaciones

$$\begin{aligned} u_\alpha &= \frac{1}{\sqrt{\lambda_\alpha}} X^{'} v_\alpha \\ v_\alpha &= \frac{1}{\sqrt{\lambda_\alpha}} X u_\alpha \end{aligned} \tag{1.22}$$

Ahora se trata de ver cómo las expresiones [1.22] se particularizan al Análisis de Correspondencias Simple. Recórdemos que en la sección anterior, hemos concluido cuales son las matrices $X^{'}X$ y $XX^{'}$ sobre las que se obtienen los ejes factoriales en $\mathbb{R}^p$ y en $\mathbb{R}^n$, en el Análisis de Correspondencias Simple. No obstante, recordemos que para simplificar los cálculos utilizamos las matrices $X^{*'}X^{*}$ y $X^{*}X^{*'}$, en donde:

$$X^{*} = \left(x^{*}_{ij}\right) \quad x^{*}_{ij} = \frac{f_{ij}}{\sqrt{f_{\cdot j} f_{i\cdot}}}$$

En consecuencia, las relaciones [1.22] se transcriben así

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X^* u_\alpha$$
$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X^{*'} v_\alpha \tag{1.23}$$

1.4.2. Relaciones entre las coordenadas de los puntos sobre los ejes factoriales en ambos espacios

Como es sabido del Análisis Factorial General, las coordenadas de los puntos-fila (puntos-columna) en el sistema de ejes factoriales u_α (v_α), establecido en $\mathbb{R}^p$ ($\mathbb{R}^n$) vienen dadas por las relaciones Xu_α ($X^{'}v_\alpha$).

En el caso del Análisis de Correspondencias Simple, y utilizando la formulación simplificada, estas coordenadas son, respectivamente X^*u_α ($X^{*'}v_\alpha$)

Nota: Obsérvese que X^*u_α es un vector $n \times 1$ (X^* es $n \times p$ y u_α, vector unitario en $\mathbb{R}^p$ es $(p \times 1)$); sus componentes nos dan las coordenadas de los n puntos-fila respecto del u_α que se vaya considerando. Si denotamos por $\widehat{\Psi}_{\alpha i}$ a la i-ésima componente de ese vector n-dimensional X^*u_α, $\widehat{\Psi}_{\alpha i}$ indica la coordenada del i-ésimo punto-fila, respecto del eje factorial α en $\mathbb{R}^p$.

Análogamente, $X^{*'}v_\alpha$ es un vector $p \times 1$, resultado de multiplicar la matriz $X^{*'}$, $p \times n$, por v_α, autovector unitario según el eje factorial α-ésimo en $\mathbb{R}^n$, es decir un vector $n \times 1$. Las p componentes, denotadas $\widehat{\Phi}_{\alpha j}$, de ese vector $X^{*'}v_\alpha$, son las coordenadas del j-ésimo punto-columna respecto del α-ésimo eje factorial ajustado en $\mathbb{R}^n$.

Por tanto, las coordenadas citadas son:

$$\widehat{\Psi}_{\alpha i} = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \right) u_{\alpha j} \quad i = 1, \cdots, n \tag{1.24}$$

$$\widehat{\Phi}_{\alpha j} = \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} \right) v_{\alpha i} \quad j = 1, \cdots, p \tag{1.25}$$

¿Qué relación existirá entre las $\widehat{\Psi}_{\alpha i}$ y $\widehat{\Phi}_{\alpha j}$?

Para encontrar esta relación consideremos que las expresiones anteriores [1.24] y [1.25] dan dichas coordenadas en términos de u_α y v_α respectivamente. Y por otro lado las relaciones entre u_α y v_α están dadas por [1.23].

Combinando ambas relaciones se concluye lo siguiente: La componente i-ésima de v_α, a partir de [1.23] es

$$v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \left(\frac{f_{ij}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \right) u_{\alpha j}$$

de donde

$$v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sqrt{f_{i\cdot}}\, \widehat{\Psi}_{\alpha i} \tag{1.26}$$

De igual forma se deduce que

$$u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sqrt{f_{\cdot j}} \widehat{\Phi}_{\alpha j} \tag{1.27}$$

Por tanto:

$$\widehat{\Psi}_{\alpha i} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \frac{1}{\sqrt{\lambda_\alpha}} \sqrt{f_{\cdot j}} \widehat{\Phi}_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}} \widehat{\Phi}_{\alpha j} \tag{1.28}$$

Y análogamente

$$\widehat{\Phi}_{\alpha j} = \sum_{i=1}^{n} \frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} \frac{1}{\sqrt{\lambda_\alpha}} \sqrt{f_{i\cdot}} \widehat{\Psi}_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{\cdot j}} \widehat{\Psi}_{\alpha i} \tag{1.29}$$

La conclusión es evidente: las coordenadas, por ejemplo, de los puntos-fila (las $\widehat{\Psi}_{\alpha i}$) se obtienen, con la homotecia $\frac{1}{\sqrt{\lambda_\alpha}}$, mediante una combinación baricéntrica de coeficientes $\frac{f_{ij}}{f_{i\cdot}}$, de todas las coordenadas, respecto del eje α-ésimo, de los puntos-columna. Esto equivale a manejar la matriz de componentes $\frac{f_{ij}}{f_{i\cdot}}$, que en la práctica suele construirse en la forma $\frac{f_{ij}}{f_{i\cdot}} \times 100$.

Por otra parte, la combinación baricéntrica $\sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}} \widehat{\Phi}_{\alpha j}$, define el baricentro de las coordenadas $\widehat{\Phi}_{\alpha j}$ de todos los puntos-columna ($j = 1, \cdots, p$) respecto del eje α, con la ponderación $f_{i\cdot}$ del punto-fila i-ésimo del que se trate (Principio Baricéntrico del cálculo de las Coordenadas en Análisis de Correspondencias Simple).

De igual forma se puede razonar sobre las $\widehat{\Phi}_{\alpha j}$.

Nota: De las relaciones [1.23] obtenemos la siguiente conclusión: Todos los autovalores λ_α son iguales o menores que la unidad. Recordemos que el autovalor unidad, $\lambda = 1$, es autovalor de $T^* = X^{*'}X^*$ y su autovector asociado es $u_p = (\sqrt{f_{\cdot j}}\ ;\ j = 1, \cdots, p)$ (según vimos en la sección 1,3,2.). Por tanto, probar lo que aquí proponemos equivale a probar que el mayor autovalor de T^* es precisamente $\lambda_1 = 1$.

A continuación resumimos esqueméticamente lo analizado la sección 1.4.2

$\mathbb{R}^p$	$\mathbb{R}^n$	Observaciones
$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X' v_\alpha$	$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X u_\alpha$	Relaciones generales en el AFG
$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X^{*'} v_\alpha$	$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X^* u_\alpha$	Particularización al Análisis de Correspondencias Simple con X^*
$X^* u_\alpha$	$X^{*'} v_\alpha$	Coordenadas de los puntos-fila y columna en el Análisis de Correspondencias Simple
$\widehat{\Psi}_{\alpha i} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} u_{\alpha j}$	$\widehat{\Phi}_{\alpha j} = \sum_{i=1}^{n} \frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} u_{\alpha i}$	$\widehat{\Psi}_{\alpha i}$ = coordenada del i-ésimo punto-fila en el eje α-ésimo en $\mathbb{R}^p$
$i = 1, \cdots, n$	$j = 1, \cdots, p$	$\widehat{\Phi}_{\alpha j}$ = coordenada del j-ésimo punto-columna en el α-ésimo eje en $\mathbb{R}^n$
$\widehat{\Psi}_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}} \widehat{\Phi}_{\alpha j}$	$\widehat{\Phi}_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{\cdot j}} \widehat{\Psi}_{\alpha i}$	Relaciones analáticas entre las coordenadas

Nota: En Análisis de Correspondencias Simple se tiene

$$X^* = (x^*_{ij}) \quad \text{con} \quad x^*_{ij} = \frac{f_{ij}}{\sqrt{f_{\cdot j} f_{i\cdot}}}$$

1.5. Ayudas a la interpretación en el Análisis de Correspondencias Simple

Desde los inicios del Análisis de Datos por Bencekri y colaboradores, al tratar cualquiera de sus técnicas, se incluye como etapa final, una vez terminado su desarrollo, una serie de cuestiones que se engloban bajo el título genérico de Ayudas a la Interpretación. Estas ayudas no son más que una serie de consecuencias derivadas del propio desarrollo de la técnica correspondiente, que sirven en la práctica de ayuda para la interpretación de los resultados obtenidos.

En concreto, a nivel de las técnicas factoriales asociadas al Análisis Factorial General del Análisis de Datos, las ayudas expresan interpretaciones, normalmente interpretables también gráficamente, asociadas a ciertas propiedades de los factores deducidos en cada caso, de modo que podamos medir e interpretar la importancia de los factores y lo que ellos significan.

En el Análisis de Componentes Principales, como caso especial del Análisis Factorial General, ya vimos una medida asociada a los factores obtenidos, que mide la importancia de ellos en la estructura

factorial construida. Recordemos la tasa de inercia general del Análisis Factorial General y su versión en el Análisis de Componentes Principales, por ejemplo normalizado, en donde

$$\sum_{j=1}^{p} \lambda_\alpha = \text{tr}\left(X^{'}X\right) = \text{tr}\left(XX^{'}\right) = \text{tr(matriz de correlaciones)} = \sum_i \sum_j x_{ij}^2 = p$$

De tal modo que se define la tasa de inercia asociada a los primeros q factores como:

$$\tau_q = \frac{\sum_{i=1}^{q} \lambda_\alpha}{\sum_{i=1}^{p} \lambda_\alpha}$$

A continuación vamos a desarrollar esta cuestión en el Análisis de Correspondencias Simple.

1.5.1. Medidas básicas para la interpretación en el Análisis de Correspondencias Simple

Se manejan en la práctica del Análisis de Correspondencias Simple dos medidas dirigidas a facilitar la interpretación de la estructura factorial que el Análisis de Correspondencias Simple asocia al espacio de puntos-fila y al de puntos-columna. Estas dos medidas son, en cierto sentido, duales en la interrelación punto-factor. Veámoslas a continuación.

1. Dado un eje factorial cualquiera, digamos el α-ésimo, los n puntos-fila, proyectados sobre dicho eje, definen sobre él un conjunto de n valores proyectados, los $\widehat{\Psi}_{\alpha i}$ obviamente, cuya dispersión vamos a medir mediante su varianza. Esta varianza, si el origen está en el centro de gravedad, vendrá dada por la expresión reducida:

$$\sum_{i=1}^{n} f_{i\cdot} \widehat{\Psi}_{\alpha i}^2 \tag{1.30}$$

   En efecto, sabemos que los puntos-fila son manejados por los perfiles correspondientes, de masa $f_{i\cdot}$ y si el origen está, en efecto, en el centro de gravedad, entonces la media (ponderada) será:

$$\overline{\widehat{\Psi}}_{\alpha i} = \sum_{i=1}^{n} f_{i\cdot} \widehat{\Psi}_{\alpha i} = 0$$

   En efecto:

$$\sum_{i=1}^{n} f_{i\cdot} \widehat{\Psi}_{\alpha i} = \sum_{i=1}^{n} f_{i\cdot} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} u_{\alpha j} = \sum_{j=1}^{p} \frac{u_{\alpha j}}{\sqrt{f_{\cdot j}}} \sum_{i=1}^{n} f_{ij} = \sum_{j=1}^{p} \frac{u_{\alpha j}}{\sqrt{f_{\cdot j}}} f_{\cdot j} = \sum_{j=1}^{p} u_{\alpha j} \sqrt{f_{\cdot j}} = 0$$

   La última igualdad se da en virtud de lo visto en el párrafo **??** anterior.

   Por tanto, la varianza considerada es la dada por [1.30]. Si esta expresión se calcula explícitamente, se tendrá:

$$\sum_{i=1}^{n} f_{i\cdot}\widehat{\Psi}_{\alpha i}^2 = \sum_{i=1}^{n} f_{i\cdot}\left(\frac{\sqrt{\lambda_\alpha}v_{\alpha i}}{\sqrt{f_{i\cdot}}}\right)^2 = \sum_{i=1}^{n}\lambda_\alpha v_{\alpha i}^2 = \lambda_\alpha\sum_{i=1}^{n} v_{\alpha i}^2 = \lambda_\alpha$$

sin más que tener en cuenta la ecuación [1.26] del párrafo 1.4.2 y el hecho de ser unitario el vector v_α, en la dirección del eje factorial α-ésimo.

En definitiva, en Análisis de Correspondencias Simple, la inercia (varianza) de los puntos-fila proyectados, en el eje α-ésimo, vale el respectivo autovalor λ_α asociado al autovector u_α que define a dicho eje. Es decir:

$$\sum_{i=1}^{n} f_{i\cdot}\widehat{\Psi}_{\alpha i}^2 = \lambda_\alpha$$

Nota: Compárese este resultado con el del Análisis Factorial General, en donde la importancia de un eje viene medida en términos de la varianza (inercia) asociada a dicho eje que, por otra parte, es igual al autovalor correspondiente. En función de esto definíamos la tasa de inercia explicada por cada eje, o por el conjunto de los q primeros ejes factoriales τ_q.

Lo anterior da pie a introducir una medida de la contribución de un punto-fila cualquiera, el i-ésimo, a la inercia asociada al eje factorial α-ésimo, definida así:

$$Ca_\alpha(i) = \left(\frac{f_{i\cdot}\widehat{\Psi}_{\alpha i}^2}{\lambda_\alpha}\right) \tag{1.31}$$

sumando debido al punto i-ésimo en [1.30] que suele llamarse contribución absoluta del elemento i-ésimo, a la inercia (varianza) explicada por el eje α-ésimo.

Nota: Otros autores llaman a $Ca_\alpha(i)$ la contribución relativa del punto i al momento α-ésimo de inercia (o, si se quiere, a la varianza explicada por el eje α-ésimo). Se denota $CTR_\alpha(i)$, según estos autores.

Una repetición de los cálculos anteriores, para el j-ésimo punto-columna, de masa $f_{\cdot j}$, conduce a definir la Contribución de dicho punto-columna a la inercia asociada al α-ésimo eje factorial en $\mathbb{R}^n$. En efecto, sin más que tener en cuenta ahora la expresión [1.27] del párrafo 1.4.2, se llega a

$$Ca_\alpha(j) = \left(\frac{f_{\cdot j}\widehat{\Phi}_{\alpha j}^2}{\lambda_\alpha}\right) \tag{1.32}$$

Nota: Tanto [1.31] como [1.32] se suelen expresar en la práctica en porcentajes, multiplicándolos por 100. A veces se multiplican por 1000.

2. Con las contribuciones absolutas medimos la importancia de una fila o de una columna, a la hora de explicar la inercia a su vez explicada por un determinado factor (por decirlo de otra manera, en las expresiones [1.31] y [1.32] de las contribuciones absolutas, se fija un α y se analiza variando i ó j, la contribución de ellos al α). Cabe plantear la situación al revés, es decir, tratar de medir el efecto (contribución) de los factores en la posición de un punto-fila o de un punto-columna determinados. Estas medidas se realizan mediante unos coeficientes denominados contribuciones relativas del factor α-ésimo en el punto-fila i-ésima (o punto-columna j-ésimo).

Veamos cómo se construyen estas contribuciones relativas.

1. Consideremos en primer lugar la nube de puntos-fila y sea el i-ésimo punto. Se sabe que el centro de gravedad de esta nube es el punto de coordenadas $\left(\sqrt{f_{\cdot j}}\ ;\ j=1,\cdots,p\right)$ en $\mathbb{R}^p$. Entonces la distancia al cuadrado de dicho punto i al centro de gravedad viene dada por

$$d^2(i;cg)=\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)^2=\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}}-f_{\cdot j}\right)^2\frac{1}{f_{\cdot j}} \tag{1.33}$$

(recuérdese, párrafo 1.3.1 ,expresión [1.9], la distancia entre dos puntos-fila en $\mathbb{R}^p$ y la expresión [1.11] que da el centro de gravedad de la nube de puntos-fila).
Definimos entonces

$$Cr_\alpha(i)=\frac{\widehat{\Psi}^2_{\alpha i}}{d^2(i;cg)}\quad ;\quad \mathbb{R}^p \tag{1.34}$$

que es llamada contribución relativa del eje factorial α-ésimo al punto fila i. Se suele denotar también $COR_\alpha(i)$.
De modo que $Cr_\alpha(i)$ puede interpretarse como el coseno al cuadrado del punto i con el eje α.

Por otra parte es fácil probar que la suma de la contribución relativa de todos los ejes factoriales al punto i vale la unidad. En efecto:

$$\sum_{\alpha=1}^{p}Cr_\alpha(i)=\sum_{\alpha=1}^{p}\frac{\widehat{\Psi}^2_{\alpha i}}{d^2(i;cg)}=\frac{1}{d^2(i;cg)}\sum_{\alpha=1}^{p}\widehat{\Psi}^2_{\alpha i}$$

Calculemos $\displaystyle\sum_{\alpha=1}^{p}\widehat{\Psi}^2_{\alpha i}$. Según la expresión [1.17] de 1.3.1 se tiene

$$\sum_{\alpha=1}^{p}\widehat{\Psi}^2_{\alpha i}=\sum_{\alpha=1}^{p}\left(\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)u_{\alpha j}\right)^2=$$

$$=\sum_{\alpha=1}^{p}\left[\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)u_{\alpha j}\right]\left[\sum_{j'=1}^{p}\left(\frac{f_{ij'}}{f_{i\cdot}\sqrt{f_{\cdot j'}}}-\sqrt{f_{\cdot j'}}\right)u_{\alpha j'}\right]=$$

$$\sum_{\alpha=1}^{p}\left\{\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)^2u^2_{\alpha j}\right\}+\sum_{\alpha=1}^{p}\left\{\sum_{\substack{j,j'=1\\ j\neq j'}}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)\left(\frac{f_{ij'}}{f_{i\cdot}\sqrt{f_{\cdot j'}}}-\sqrt{f_{\cdot j'}}\right)u_{\alpha j}u_{\alpha j'}\right\}=$$

$$=\left\{\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)^2\right\}\sum_{\alpha=1}^{p}u^2_{\alpha j}+\left\{\sum_{\substack{j,j'=1\\ j\neq j'}}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{\cdot j}}\right)\left(\frac{f_{ij'}}{f_{i\cdot}\sqrt{f_{\cdot j'}}}-\sqrt{f_{\cdot j'}}\right)\right\}\sum_{\substack{\alpha=1\\ j\neq j'}}^{p}u_{\alpha j}u_{\alpha j'}$$

Pero teniendo en cuenta que, dados los autovectores u_α (unitarios en la dirección del eje factorial respectivo F_α), estos son ortogonales (son correspondientes a raíces características distintas), entonces la matriz que los contiene como columnas es ortogonal. Es decir

$$(u_1|u_2|,\cdots,|u_\alpha|,\cdots,|u_p)\,(u_1|u_2|,\cdots,|u_\alpha|,\cdots,|u_p)^{'} = I_{p\times p}$$

De aquí, mediante fáciles cálculos, se comprueba que:

$$\sum_{\alpha=1}^{p} u_{\alpha j}^2 = 1 \quad ; \quad \sum_{\substack{\alpha=1\\ j\neq j'}}^{p} u_{\alpha j} u_{\alpha j'} = 0$$

En resumen

$$\sum_{\alpha=1}^{p} \widehat{\Psi}_{\alpha i}^2 = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right)^2 = d^2(i;cg)$$

Por tanto, en efecto:

$$\sum_{\alpha=1}^{p} Cr_\alpha(i) = 1$$

Nota: Otra interpretación de la $Cr_\alpha(i)$ es como coeficiente de correlación al cuadrado entre el eje α y el punto i-ésimo fijado

2. Análogamente, cabe actuar de manera paralela, en $\mathbb{R}^n$, con la nube de puntos-columna. En definitiva, se establecen la contribuciones relativas de los ejes factoriales α en $\mathbb{R}^n$ a cada punto-columna ($j = 1,\cdots,p$):

$$Cr_\alpha(j) = \frac{\widehat{\Phi}_{\alpha j}^2}{d^2(j;cg)} \quad ; \quad \mathbb{R}^n \tag{1.35}$$

con análogas interpretaciones que las $Cr_\alpha(i)$. Aquí el centro de gravedad, cg, es el da la nube de puntos-columna y obviamente nos situamos en $\mathbb{R}^n$, donde los puntos-columna son los puntos considerados

1.6. Análisis de Correspondencias Múltiple

El uso del Análisis de Correspondencias Simple se puede generalizar a más de dos variables. Una de las generalizaciones más utilizadas es el Análisis de Correspondencias Múltiple, que consiste en aplicar el Análisis de Correspondencias Simple a una tabla de contingencia dónde las filas representan a los individuos y las columnas corresponden a las distintas modalidades de variables categóricas que estamos estudiando sobre los individuos. Las modalidades de cada variable se codifican con unos y ceros dependiendo de si el individuo presenta o no la modalidad en cuestión. Es decir, supongamos que tenemos s variables categóricas. Sea la variable Z una de estas variables que presenta m modalidades distintas, cada una de las modalidades viene dada en una columna de la tabla de contingencia. Entonces, para el individuo $i-ésimo$ en la modalidad q de la variable Z tenemos:

$$Z_{iq} = \begin{cases} 1 & si\ el\ individuo\ i-ésimo\ presenta\ la\ modalidad\ q \\ 0 & si\ el\ individuo\ i-ésimo\ no\ presenta\ la\ modalidad\ q \end{cases}$$

Ejemplo: *Supongamos que estamos estudiando una serie de variables categóricas sobre una muestra de 5 individuos. Entre ellas tenemos las variables Y="Nivel de estudios", que presenta las modalidades: primarios, secudarios, superiores y otros, y la variable Z="Ingresos mensuales", que presenta las modalidades: Bajos, Medios y Altos. Entonces, estas variables se codifican en la tabla como sigue:*

	Pr.	**S.**	**Su.**	**O.**	**Baj.**	**Med.**	**Alt.**
1	1	0	0	0	1	0	0
2	0	1	0	0	0	1	0
3	0	1	0	0	1	0	0
4	0	0	0	1	0	0	1
5	0	0	1	0	0	1	0

En este ejemplo el primer individuo tiene estudios primarios e ingresos mensuales bajos. El segundo tiene estudios secundarios e ingresos medios, etc.

Supongamos ahora que tenemos una muestra de n individuos y un total s variables categóricas, cada una con m_s modalidades. Supongamos que $p = \sum_{i=1}^{s} m_i$, es decir, en total tenemos p categorías distintas. Reenumerando las categorías de las variables de 1 a p nuestra tabla de contingencia quedará como sigue:

		Categorías					
		1	2	j	$\cdots$	p	Total
Individuos	1	z_{11}	$\cdots$	$\cdots$	$\cdots$	z_{1p}	$z_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	i	z_{i1}	$\cdots$	z_{ij}	$\cdots$	z_{ip}	$z_{i\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	n	z_{n1}	$\cdots$	$\cdots$	$\cdots$	z_{np}	$z_{n\cdot}$
	Total	$z_{\cdot 1}$	$\cdots$	$z_{\cdot j}$	$\cdots$	$z_{\cdot p}$	K

A esta tabla de contingencia se la conoce como matriz disyuntiva.

En esta tabla los $z_{i\cdot}$ son todos iguales, e iguales al número de variables cualitativas que tenemos, es decir, a s. Por otro lado, los $z_{\cdot,j}$ representan el número de individuos que presentan cada uno de las categorías de las variables. El valor K por tanto, no es más que $n * s$, es decir, el número de individuos de nuestra muestra por el total de variables categóricas.

El Análisis de Correspondencias Múltiples consiste en aplicar lo que hemos visto para el Análisis de Correspondencias Simples a esta matriz disyuntiva.

Nota: Alternativamente, si multiplicamos la traspuesta de la matriz disyuntiva por la matriz disyuntiva, obtenemos lo que se conoce como matriz de Burt. Dicha matriz relaciona las distintas categorías de cada una de las variables.

Veamos qué quiere decir con el ejemplo anterior. Sea D la matriz que representa la tabla de datos del ejemplo anterior.

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Ahora multiplicamos $D^{'}$ por D y nos queda:

$$D^{'} * D = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

El resultado obtenido proporciona una tabla dónde tanto las filas como las columnas representan categorías de las variables bajo estudio. Es decir, enfrentamos las categorías de las variables unas con otras y vemos cuantos individuos hay en la combinación de dos categorías cualquiera.

En forma de tabla, lo que acabamos de obterner se ve de la siguiente forma:

	Pr.	**S.**	**Su.**	**O.**	**Baj.**	**Med.**	**Alt.**
Pr.	1	0	0	0	1	0	0
S.	0	2	0	0	1	1	0
Su.	0	0	1	0	0	1	0
O.	0	0	0	1	0	0	1
Baj.	1	1	0	0	2	0	0
Med.	0	1	1	0	0	2	0
Sup.	0	0	0	1	0	0	1

En esta tabla podemos observar que en la diagonal principal se encuentran las categorías de las variables enfrentadas a si mismas. Podemos observar en dicha diagonal que hay 1 individuo con estudios Primarios, 2 con estudios Secundarios, 1 con estudios Superiores y otro individuo en la categoría Otros. Además, se puede ver que 2 de esos individuos tienen ingresos mensuales Bajos, 2 tienen ingresos Medios y 1 ingresos Altos. Fuera de la diagonal tenemos las categorías de la variable Y="Nivel de estudios", enfrentadas a las de la variable Z="Ingresos mensuales". En este caso podemos ver, por ejemplo, que hay 1 individuo con estudios Primarios e ingresos Bajos, otro con estudios Secundarios e ingresos Bajos, otro con estudios Secundarios e ingresos Medios, etc.

Aplicar el Análisis de Correspondencias Simples a la matriz de Burt es equivalente a aplicarlo a la matriz disyuntiva.

1.6.1. Ayudas a la interpretación de los resultados

En esta sección se resumen algunas ideas que pueden facilitar la interpretación de los resultados del análisis de correspondencias múltiple.

- Las contribuciones absolutas ayudan a dar una interpretación a los ejes principales. Con ellas medidos la importancia que tiene cada variable en la construcción de cada eje. Una variable aporta suficiente inercia al eje α si su contribución absoluta es mayor que 1 entre el número total de variables s $(1/s)$.

- Las contribuciones relativas nos dicen lo bien representadas que están las categorías por cada eje.

  Se considera que una categoría está bien representada en el espacio de los ejes principales si la suma de sus contribuciones relativas con los ejes principales es de al menos un 60 %.

  Se considera que una categoría está bien representada por un eje concreto si su contribución relativa a ese eje es de al menos un 30 %.

- Otras

  - Las categorías que aparecen más cerca del centro de gravedad, es decir, el origen, son las categorías mas comunes para los individuos que se están estudiando. Por tanto, aquellas modalidades que hayan sido escogidas por pocos individuos se encontrarán alejadas del centro de gravedad.
  - Dos individuos estarán cercanos si coinciden en la mayoría de las mismas modalidades. Si tienen las mismas modalidades entonces estarán los dos en el mismo punto.
  - Dos modalidades escogidas por los mismos individuos valen lo mismo.

Tema 2

Aplicación en R

2.1. Análisis de Correspondencias Simple

Vamos a aplicar Análisis de Correspondencias Simple a una tabla de contingencia que contiene datos de un estudio que se ha realizado sobre la compra de 4 marcas de móviles (variable I) y el tipo de cliente (variable J). Se han investigado 4 marcas de móviles y se han identificado 3 tipos de cliente según ingresos familiares.

Nota: En R tenemos varios paquetes disponibles para aplicar el Análisis de Correspondencias, por ejemplo, los ***ade4*** y ***FactoMineR*** realizan dicho análisis. Instalamos y cargamos en R los paquetes ***ade4*** y ***factoextra*** para realizar este ejercicio. Este último paquete nos sirve para representar los resultados gráficamente.

En primer lugar abrimos el fichero **correspondencias.txt** que se encuentra en Prado y contiene la tabla de contingencia de este estudio:

```
> datos<-read.table("correspondencias.txt", header=TRUE)
> datos
      Tipo1 Tipo2 Tipo3
Marca1  20   20   115
Marca2  20   90   20
Marca3  60   20   20
Marca4  60   20   2
```

Lo primero que debemos hacer es investigar si las variables son independientes, si lo son, no tiene sentido continuar con el análisis. Realizamos un contraste Chi-cuadrado de independencia:

```
> chisq.test(datos)
Pearson's Chi-squared test

data:  datos
X-squared = 287.38, df = 6, p-value < 2.2e-16
```

Claramente rechazamos la independencia de las variables, por lo que podemos estudiar cómo se relacionan estas variables mediante el Análisis de Correspondencias Simple.

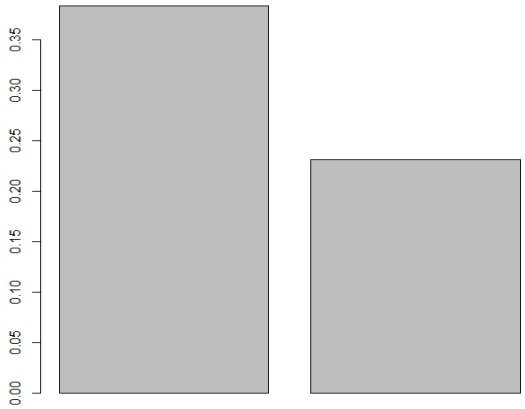
Aplicamos el procedimiento *dudi.coa* para obtener los resultados del Análisis de Correspondencias Simple. Al aplicar este procedimiento R nos pregunta con cuantas dimensiones queremos trabajar, para lo cual proporciona el gráfico de los autovalores que podemos ver debajo. Los autovalores y el porcentaje de varianza explicada por cada eje nos sirven para determinar cual es la dimensión del espacio en el que vamos a representar nuestros datos de manera que perdamos la menor cantidad de información.

En este ejemplo tan sencillo realmente con 2 dimensiones estamos representando el total de la varianza contenida en los datos, en realidad, no estamos haciendo ninguna reducción de la dimensión para representar los datos, es por ello que el gráfico de los autovalores muestra 2 autovalores. El espacio en el que vamos a trabajar y realizar las representaciones tiene obviamente 2 dimensiones.

Veamos el gráfico de los autovalores.

```
>acs<-dudi.coa(datos)
```

Figura 2.1: Gráfico de los autovalores



Ahora le pedimos que nos muestre los objetos del análisis *acs*

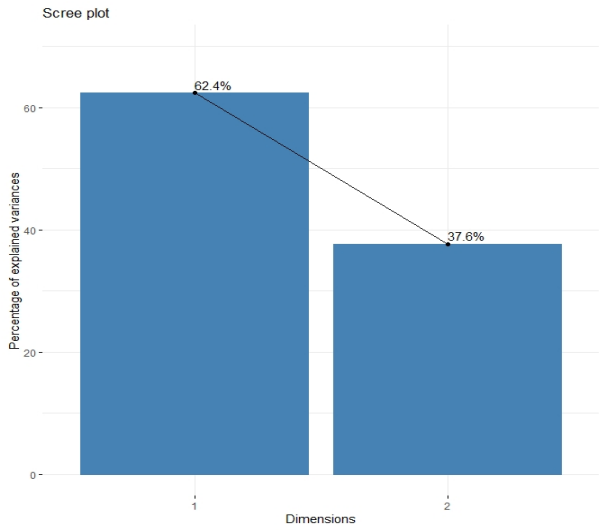
```
> acs
Duality diagramm
class: coa dudi
$call: dudi.coa(df = datos)

$nf: 2 axis-components saved
$rank: 2
eigen values: 0.3837 0.2316
  vector length mode    content
1 $cw    3      numeric column weights
2 $lw    4      numeric row weights
3 $eig   2      numeric eigen values

  data.frame nrow ncol content
1 $tab       4    3    modified array
2 $li        4    2    row coordinates
3 $l1        4    2    row normed scores
4 $co        3    2    column coordinates
5 $c1        3    2    column normed scores
other elements: N
```

En esta salida podemos observar los valores de la inercia, 0.3837 y 0.2316. El primer eje tiene más poder clasificatorio que el segundo.

Mediante el siguiente gráfico podemos ver el porcentaje de la varianza de los datos explicada por cada eje:

```
> fviz_screeplot(acs, addlabels=TRUE,ylim=c(0,70))
```

Figura 2.2: Gráfico del porcentaje de varianza explicada



Podemos observar que entre los dos ejes se explica el total de la varianza. Como se ha mencionado anteriormente, en este ejemplo tenemos pocas modalidades de cada variable, y en realidad no estamos haciendo una proyección sobre un subespacio de menor dimensión.

A continuación obtenemos los pesos de las filas en primer lugar y de las columnas en segundo lugar:

```
> acs$lw
   Marca1    Marca2    Marca3    Marca4
0.3319058 0.2783726 0.2141328 0.1755889

> acs$cw
    Tipo1     Tipo2     Tipo3
0.3426124 0.3211991 0.3361884
```

Dentro de las filas, la modalidad más influyente es la Marca1, y dentro de las columnas las influencias son similares, siendo la más influyente la modalidad Tipo1.

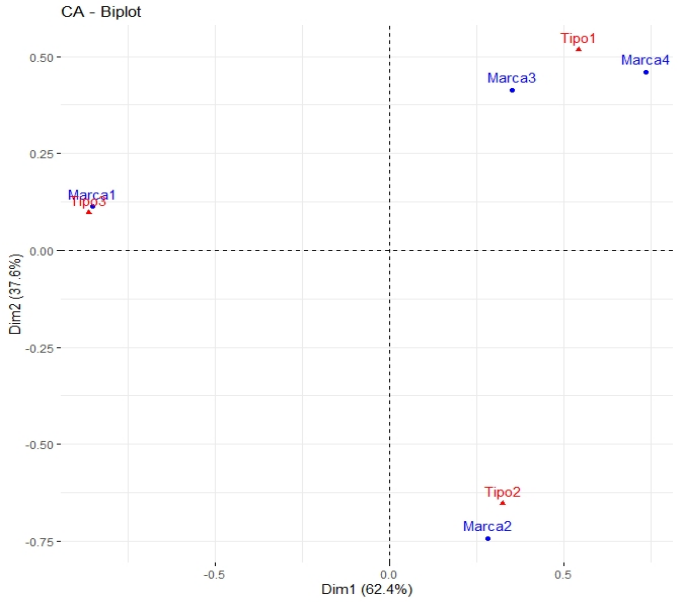
Las coordenadas por filas y columnas sobre los ejes obtenidos por el ACS son las siguientes:

```
> acs$lw
   Marca1    Marca2    Marca3    Marca4
0.3319058 0.2783726 0.2141328 0.1755889
> acs$cw
    Tipo1     Tipo2     Tipo3
0.3426124 0.3211991 0.3361884
> acs$li
            Axis1      Axis2
Marca1 -0.8514957  0.1129835
Marca2  0.2824273 -0.7429064
Marca3  0.3511712  0.4134526
Marca4  0.7335265  0.4600017
> acs$l1
              RS1        RS2
Marca1 -1.3745684  0.2347550
Marca2  0.4559220 -1.5435974
Marca3  0.5668952  0.8590642
Marca4  1.1841308  0.9557831
> acs$co
           Comp1      Comp2
Tipo1  0.5418040  0.5169624
Tipo2  0.3237476 -0.6528781
Tipo3 -0.8614698  0.0969282
> acs$c1
             CS1        CS2
Tipo1  0.8746334  1.0741352
Tipo2  0.5226253 -1.3565383
Tipo3 -1.3906697  0.2013956
```

Realizamos ahora una serie de gráficos para determinar cómo se relacionan las variables.
El primer gráfico es un gráfico simétrico y muestra un patrón global dentro de los datos. En este gráfico las filas y las columnas se representan en el mismo espacio utilizando las coordenadas principales. En este caso, solo se puede interpretar realmente la distancia entre puntos de fila o la distancia entre puntos de columna.

```
> fviz_ca_biplot(acs)
```

Figura 2.3: Gráfico conjunto de filas y columnas



En el gráfico anterior las filas están representadas por puntos azules y las columnas por triángulos rojos. La distancia entre cualquier punto de fila o columna da una medida de su similitud (o disimilitud). Lo que podemos decir mirando este gráfico es que los clientes de Tipo1 tienden a comprar las Marcas 3 y 4 de móviles. Los clientes de Tipo 2 tienden a comprar la Marca 2. Finalmente los clientes de Tipo 3 tieden a comprar la Marca 1.

Como hemos mencionado anteriormente, en este tipo de gráfico las distancias entre los puntos fila y columna no se puede interpretar, para interpretar esa distancia, los perfiles de columna deben presentarse en el espacio de fila o viceversa. Este tipo de mapa se llama biplot asimétrico y lo vemos a continuación.

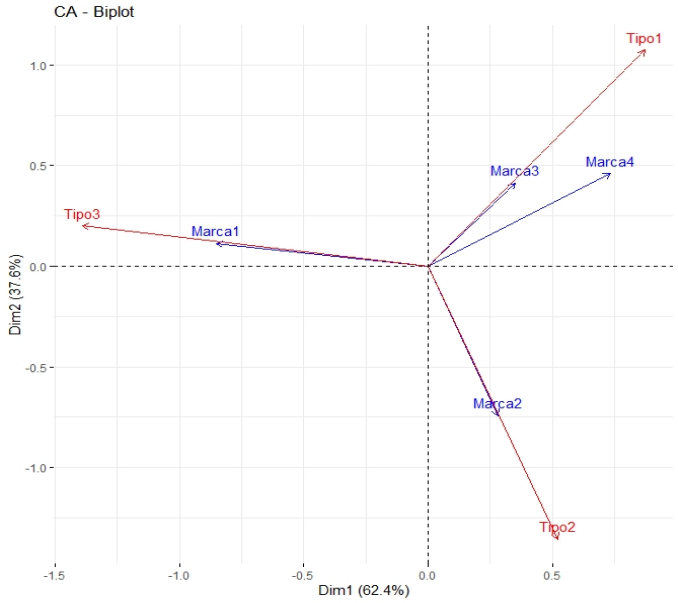
Para hacer un gráfico asimétrico, los puntos de las filas (o columnas) se trazan a partir de las coordenadas estándar (S) y los perfiles de las columnas (o las filas) se trazan a partir de las coordenadas principales (P) .

El segundo gráfico representa los perfiles columna en el espacio de las filas.

```
>fviz_ca_biplot(acs, map ="rowprincipal", arrow = c(TRUE, TRUE))
```

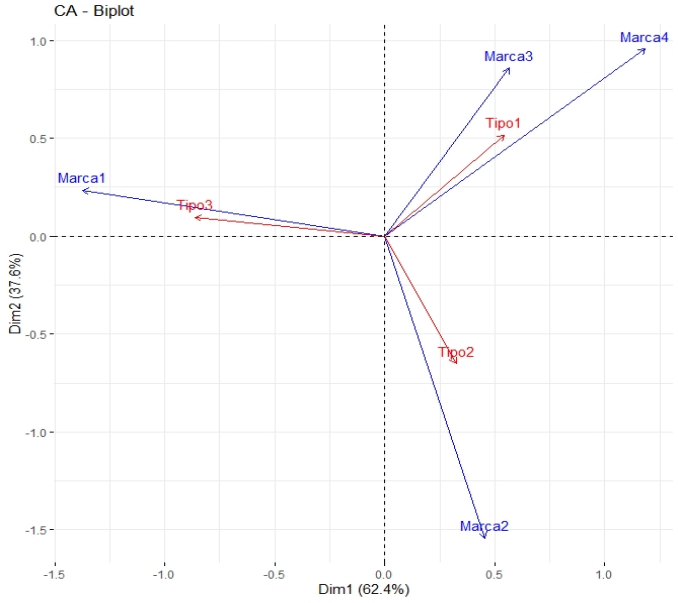
En este gráfico, si el ángulo entre dos flechas es agudo, entonces existe una fuerte asociación entre la fila y la columna correspondientes. Además para interpretar la distancia entre filas y una columna, se debe proyectar perpendicularmente puntos de fila en la flecha de la columna.

Figura 2.4: Gráfico conjunto de filas y columnas. Columnas en el espacio de las filas




```
>fviz_ca_biplot(acs, map ="colprincipal", arrow = c(TRUE, TRUE))
```

Figura 2.5: Gráfico conjunto de filas y columnas. Filas en el espacio de las columnas



Este último gráfico tiene una interpretación similar al anterior.

2.2. Análisis de Correspondencias Múltiple

Ahora vamos a analizar una base de datos que contiene varias variables categóricas. En concreto, contiene 8 variables categóricas. Esta base de datos la podemos encontrar en R. Es la base de datos **DogsBreeds**.

Nota: Para realizar esta práctica es necesario intarlar y cargar en R los siguientes paquetes: *ade4, factoextra, FactoMineR.*

La base de datos DogBreeds viene dada como una tabla de contingencia con 27 filas y 7 columnas. Las filas representan los tipos de razas de perro que se han investigados, las columnas se refieren a 7 caracteísticas que describen cada una de las razas:

- Size (tamaño): Small, Medium, Large
- Weight (peso): Light, Medium, Heavy
- Speed (velocidad): Low, Medium, High
- Intelligence (inteligencia): Low, Medium, High
- Afecttivity (afectividad): Low, High
- Aggressiviness (agresividad): Low, High
- Function (función): Company, Hunter, Utility

Nosotros vamos a considerar sólo las variables relacionadas con el carácter del perro, es decir, las seis primeras variables.

Cargamos el paquete de datos

```
> data("DogBreeds")
> DogBreeds
     SIZE WEIG SPEE INTE AFFE AGGR FUNC
bass  sma  lig  low  low  low  hig  hun
beau  lar  med  hig  med  hig  hig  uti
boxe  med  med  med  med  hig  hig  com
buld  sma  lig  low  med  hig  low  com
bulm  lar  hea  low  hig  low  hig  uti
cani  sma  lig  med  hig  hig  low  com
chih  sma  lig  low  low  hig  low  com
cock  med  lig  low  med  hig  hig  com
coll  lar  med  hig  med  hig  low  com
dalm  med  med  med  med  hig  low  com
dobe  lar  med  hig  hig  low  hig  uti
dogo  lar  hea  hig  low  low  hig  uti
foxh  lar  med  hig  low  low  hig  hun
foxt  sma  lig  med  med  hig  hig  com
galg  lar  med  hig  low  low  low  hun
gasc  lar  med  med  low  low  hig  hun
labr  med  med  med  med  hig  low  hun
masa  lar  med  hig  hig  hig  hig  uti
mast  lar  hea  low  low  low  hig  uti
peki  sma  lig  low  low  hig  low  com
podb  med  med  med  hig  hig  low  hun
podf  lar  med  med  med  low  low  hun
poin  lar  med  hig  hig  low  low  hun
sett  lar  med  hig  med  low  low  hun
```

```
stbe  lar  hea  low  med  low  hig  uti
teck  sma  lig  low  med  hig  low  com
tern  lar  hea  low  med  low  low  uti
```

Creamos un nuevo fichero de datos que contiene información de las 6 primeras variables. Mostramos un resumen de estas variables.

```
>Datosrazas<-DogBreeds[,-7]
>summary(Datosrazas)
  SIZE     WEIG     SPEE     INTE     AFFE     AGGR
 lar:15   hea: 5   hig: 9   hig: 6   hig:14   hig:13
 med: 5   lig: 8   low:10   low: 8   low:13   low:14
 sma: 7   med:14   med: 8   med:13
```

Ahora procedemos con el análisis de correspondencias múltiple. Para el análisis de correspondencias múltiples el número total de dimensiones que se generan es igual al número total de categorías (p=16) menos el número de variables cualitativas (s=6) consideradas en el análisis. Por tanto, en este caso tenemos p-s=16-6=10 dimensiones.

Nota: Para extraer conclusiones del ACM, con el fin de disminuir la complejidad del problema a analizar, se consideran solo los resultados correspondientes a unas pocas dimensiones, que serían aquellas que recojan un porcentaje adecuado de la inercia total del total. Normalmente tratamos de quedarnos con 2 o 3 dimensiones, es decir, tratamos de representar los puntos en $\mathbb{R}^2$ o $\mathbb{R}^3$. Una forma usual de proceder es analizar aquellas dimensiones cuya inercia sea mayor que la inercia media por dimensión (inercia total/número de dimensiones). La inercia total se obtiene haciendo uso de las modalidades y el número de variables categóricas, ya que se obtiene como el número de modalidades (p=16) entre el número de variables cualitativas (s=6) menos 1.

```
> p<-16
> s<-6
> Intotal<-(p/s)-1
> Intotal
[1] 1.666667
```

Por tanto, analizaremos aquellas dimensiones con una inercia mayor a $1{,}666667/6 = 0{,}2777778$.

A continuacíon realizamos el ańalisis de correspondencias múltiples y se muestra un resumen de dicho análisis. Por defecto se muestra la información referente a las 5 primeras dimensiones.

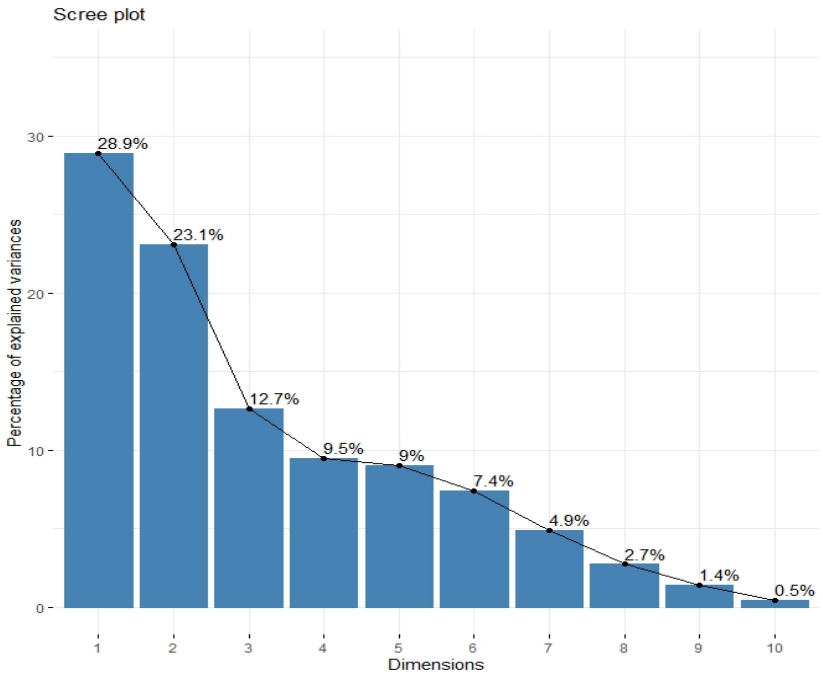
```
> acm<-dudi.acm(Datosrazas, nf=10,scannf = FALSE)
> summary(acm)
Class: acm dudi
Call: dudi.acm(df = Datosrazas, scannf = FALSE, nf = 10)

Total inertia: 1.667

Eigenvalues:
    Ax1     Ax2     Ax3     Ax4     Ax5
 0.4816  0.3847  0.2110  0.1576  0.1501

Projected inertia (%):
```

```
    Ax1     Ax2     Ax3     Ax4     Ax5
 28.896  23.084  12.657   9.453   9.008

Cumulative projected inertia (%):
    Ax1   Ax1:2   Ax1:3   Ax1:4   Ax1:5
  28.90   51.98   64.64   74.09   83.10

(Only 5 dimensions (out of 10) are shown)
```

Ahora vemos el gráfico de la inercia explicada por cada eje.

```
> fviz_screeplot(acm, addlabels = TRUE, ylim=c(0,35))
```

Figura 2.6: Gráfico de la inercia explicada por cada eje



Observando los autovalores y el gráfico parece que dos ejes serían suficientes para realizar el análisis. Los dos primeros ejes explican el 51.98 % de la inercia. El primer eje explica el 28,9 % de la inercia y el segundo el 23,1 %, después de eso hay un disminución considerable, el tercer eje explica el 2,7 % de la inercia de los datos.

A continuación obtenemos las coordenadas por filas y por columnas que nos proporciona el análisis. Se muestran sólo los seis primeros ejes:

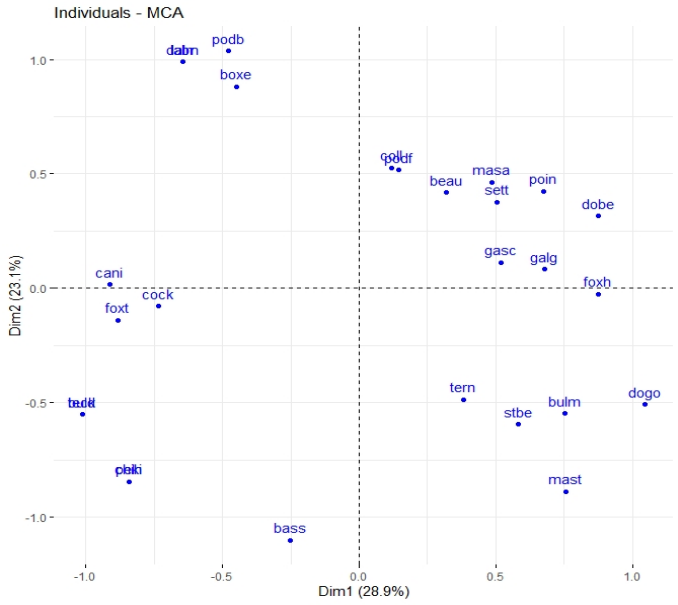
```
> acm$li
           Axis1       Axis2       Axis3       Axis4         Axis5        Axis6
bass -0.2541098 -1.10122699 -0.19070097  0.29263727 -0.5240085192 -0.039894681
beau  0.3172001  0.41770130 -0.10146771 -0.21143628 -0.1185095442  0.844917274
boxe -0.4473649  0.88177794  0.69201580  0.26000184 -0.4555898365  0.213745841
buld -1.0133522 -0.54987949 -0.16342320 -0.34991927  0.3307864847  0.201414177
bulm  0.7525745 -0.54691183  0.49757307  0.65515266  0.7219463555 -0.117925813
cani -0.9123015  0.01618767 -0.57656972  0.62813340  0.4340165309 -0.386066396
chih -0.8407994 -0.84385216 -0.46994714 -0.08634287 -0.1773464546 -0.197088251
cock -0.7332953 -0.07907317  0.66223042  0.18974319 -0.1046272364  0.626726024
coll  0.1173252  0.52610765 -0.33489373 -0.65775454  0.1921301958  0.374168009
dalm -0.6472398  0.99018429  0.45858978 -0.18631642 -0.1449500965 -0.257003424
dobe  0.8732102  0.31548110 -0.45231373  0.51008713  0.2398797276  0.254209144
dogo  1.0470168 -0.50695768  0.16503476  0.06288820 -0.3165215706  0.001701668
foxh  0.8765675 -0.02523985 -0.36217150 -0.01519800 -0.6626648082  0.132859151
foxt -0.8816221 -0.13896696  0.05352247  0.28559012 -0.2710348055  0.361835304
galg  0.6766927  0.08316651 -0.59559752 -0.46151625 -0.3520250682 -0.337890114
gasc  0.5173377  0.11340393  0.04402869  0.24097247 -0.8179229680 -0.418446461
labr -0.6472398  0.99018429  0.45858978 -0.18631642 -0.1449500965 -0.257003424
masa  0.4863955  0.46444958 -0.49813388  0.57742525  0.2759020523  0.567764838
mast  0.7559318 -0.88763278  0.58771530  0.12986753 -0.1805981803 -0.239275806
peki -0.8407994 -0.84385216 -0.46994714 -0.08634287 -0.1773464546 -0.197088251
podb -0.4780443  1.03693257  0.06192362  0.60254511  0.2494614999 -0.534155860
podf  0.1449101  0.51578295  0.11712661 -0.46892219  0.0008497112 -0.490693297
poin  0.6733354  0.42388745 -0.68573975  0.06376888  0.5505194676 -0.216540121
sett  0.5041399  0.37713917 -0.28907358 -0.72509266  0.1561078711  0.060612314
stbe  0.5833790 -0.59366011  0.89423924 -0.13370887  0.3275347590  0.159226622
teck -1.0133522 -0.54987949 -0.16342320 -0.34991927  0.3307864847  0.201414177
tern  0.3835042 -0.48525376  0.66081322 -0.58002713  0.6381744990 -0.311522643

> acm$co
             Comp1       Comp2       Comp3       Comp4       Comp5       Comp6
SIZE.lar  0.8366753  0.02057846 -0.05121744 -0.17022176  0.11266304  0.04996469
SIZE.med -0.8510880  1.23171972  1.01605178  0.34245635 -0.31004022 -0.11829709
SIZE.sma -1.1849557 -0.92389650 -0.61599962  0.12014924 -0.01996350 -0.02256927
WEIG.hea  1.0151341 -0.97390062  1.22159452  0.06760494  0.61451838 -0.28923175
WEIG.lig -1.1689180 -0.82434462 -0.35877044  0.16488382 -0.05122143  0.20335954
WEIG.med  0.3054053  0.81887572 -0.23127208 -0.11836395 -0.19020146 -0.01290840
SPEE.hig  0.8920999  0.37183247 -0.76308752 -0.23984823 -0.01008873  0.53218070
SPEE.low -0.3199406 -1.04490006  0.40172878 -0.08033130  0.30590834  0.02448794
SPEE.med -0.6036867  0.88781355  0.35631249  0.37024339 -0.37103561 -0.62931321
INTE.hig  0.3350656  0.45948302 -0.59992378  1.27524863  1.06319128 -0.20538874
INTE.low  0.3490450 -0.80855486 -0.35151126  0.02423769 -1.03505999 -0.46104962
INTE.med -0.3694426  0.28550314  0.49320252 -0.60349180  0.14625633  0.37851765
AFFE.hig -0.7754964  0.26693613 -0.06079688  0.07721587  0.04032182  0.31806714
AFFE.low  0.8351500 -0.28746968  0.06547357 -0.08315555 -0.04342350 -0.34253384
```

```
AGGR.hig  0.4315386 -0.20919553  0.33354829  0.55115649 -0.37446406  0.51425491
AGGR.low -0.4007145  0.19425299 -0.30972341 -0.51178817  0.34771663 -0.47752242
```

Ahora realizamos una serie de gráficos que nos ayudan a interpretar cómo se relacionan las variables. En primer lugar representamos los perfiles-fila, es decir, los puntos correspondientes a las razas de perro. Los representamos en los dos primeros ejes (gráfico 2.7). En segundo lugar representamos de forma conjunta las razas de perro y sus características (gráfico 2.8)

```
>fviz_mca_ind(acm)
>fviz_mca_biplot(acm)
```

Figura 2.7: Nube de puntos perfiles-fila

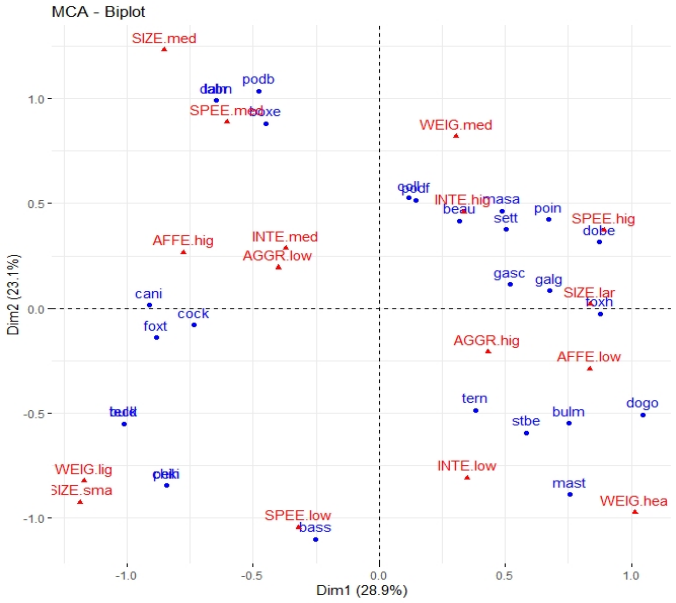


En el gráfico 2.7 podemos ver que existen razas cuyas caracteristicas son prácticamente iguales, los puntos están superpuestos. Esto ocurre por ejemplo con la razas chihuahua y pekines; las razas teckel y Bulldog; los dalmatas y labradores; etc. Además, hay otras razas que están bastante próximas en relación a su caracter, por ejemplo coll y podf.

En el gráfico conjunto (2.8) podemos observar, entre otras relaciones, las siguientes:

- Podemos ver que las modalidades de inteligencia alta, velocidad alta están cercanas a las razas Dalmata, Podb, Boxer y Labrador.
- Las razas de Doberman, Galgo y foxn son más veloces y más altos.
- Las modalidades tamaño pequeño, poco peso e inteligencia baja están relacionados con las razas Teckel, Chihuahua, Pekines y buld,
- Razas con poca afectividad y con peso alto son los dogos, los mastines y los bulm.
- Las razas Collie, Podf, Beauceron, Masa y Setter, tienden a tener pesos medianos, alto grado de inteligencia y tamaños altos.

Figura 2.8: Gráfico conjunto de razas de perro y características



Por último mostramos una serie de gráficos sobre la contribución de las distintas categorías a las tres primeras dimensiones:

El primer gráfico muestra la contribución de las categrías al plano de las dos primeras dimensiones (gráfico 2.9).

```
fviz_contrib(acm, choice="var", axes = c(1,2), fill="blue")
```

Los siguientes gráficos son una versión mejorada de las nubes de puntos de los categorías en subespacios de dos dimensiones dónde el color de las categorías depende de su contribución a los subespacios de dos dimensiones dónde se representan. El primero de estos gráficos es la representación de esa nube de puntos sobre los dos primeros ejes (gráfico 2.10), el segundo es sobre los ejes 1 y 3 (gráfico 2.11), y el último es sobre los ejes 2 y 3 (gráfico 2.12).

```
fviz_mca_var(acm, axes=c(1,2), choice="var.cat", repel=T, gradient.cols=c("yellow","orange","red
fviz_mca_var(acm, axes=c(1,3), choice="var.cat", repel=T, gradient.cols=c("yellow","orange","red
fviz_mca_var(acm, axes=c(2,3), choice="var.cat", repel=T, gradient.cols=c("yellow","orange","red
```

Observando los gráficos 2.9 y 2.10 podríamos decidir eliminar las categorías que no están bien representadas por los ejes principales, como ocurre en las dos primeras dimensiones con las categorías SPEED.high, FUNC.hun, AGGR.high, AGGR.low, INTE.med, INTE.high. Esto es recomendable cuando tenemos muchas categorías y se hace difícil interpretar los gráficos. Además, interpretar categorías que no están bien representadas por los ejes no tiene sentido.

Figura 2.9: Contribución de las categorías al subespacio 1-2

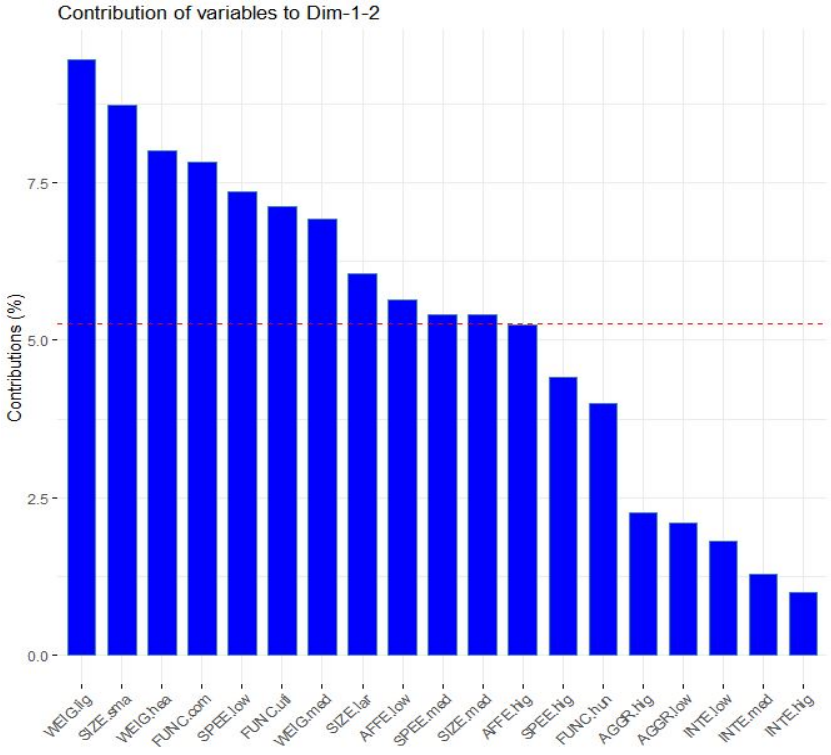


Figura 2.10: Nube proyectada de las categorías al subespacio 1-2

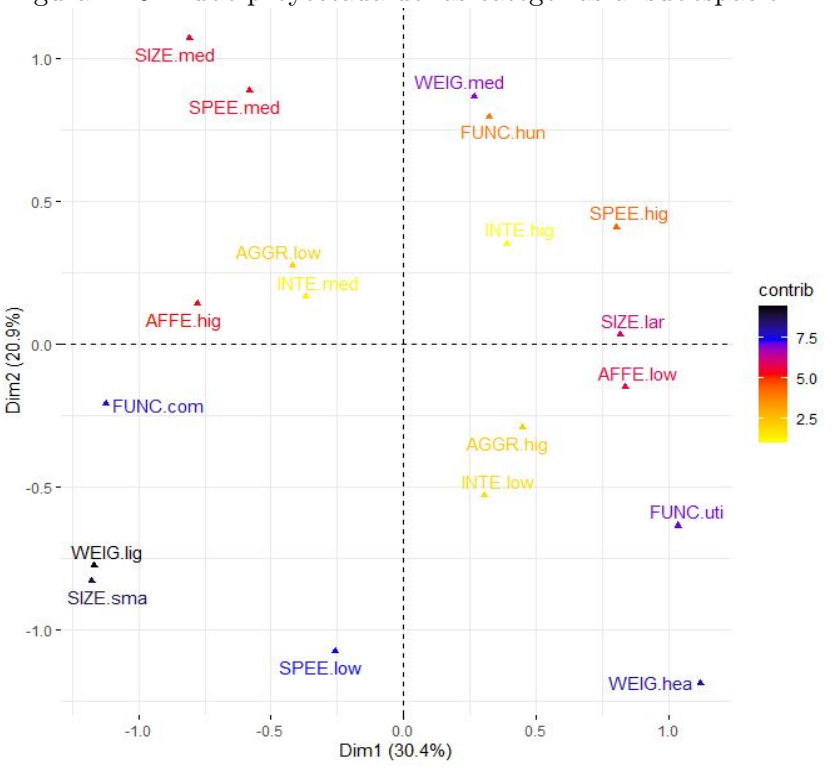


Figura 2.11: Nube proyectada de las categorías al subespacio 1-3

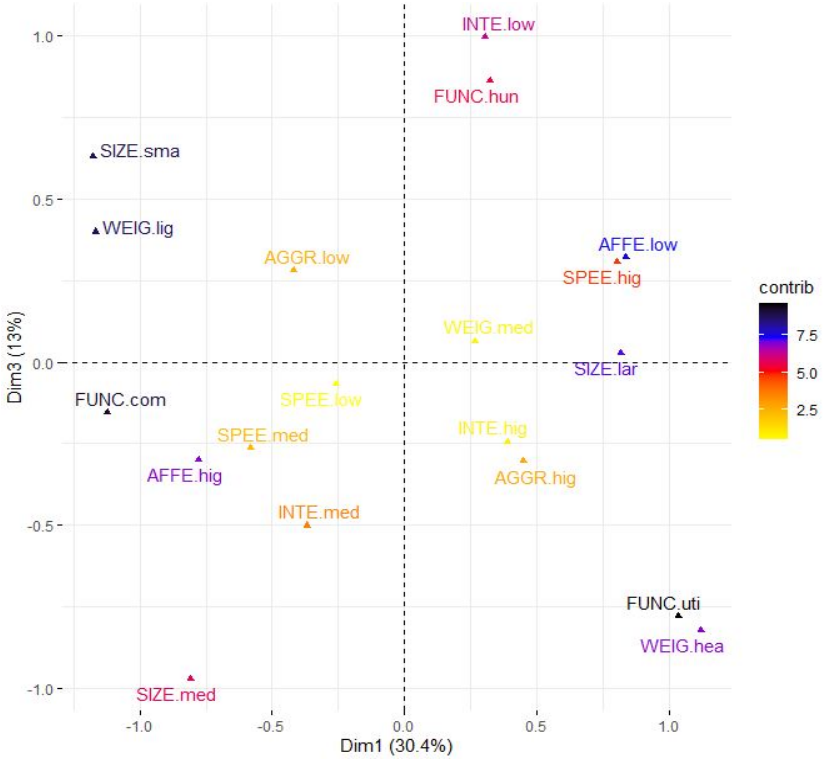
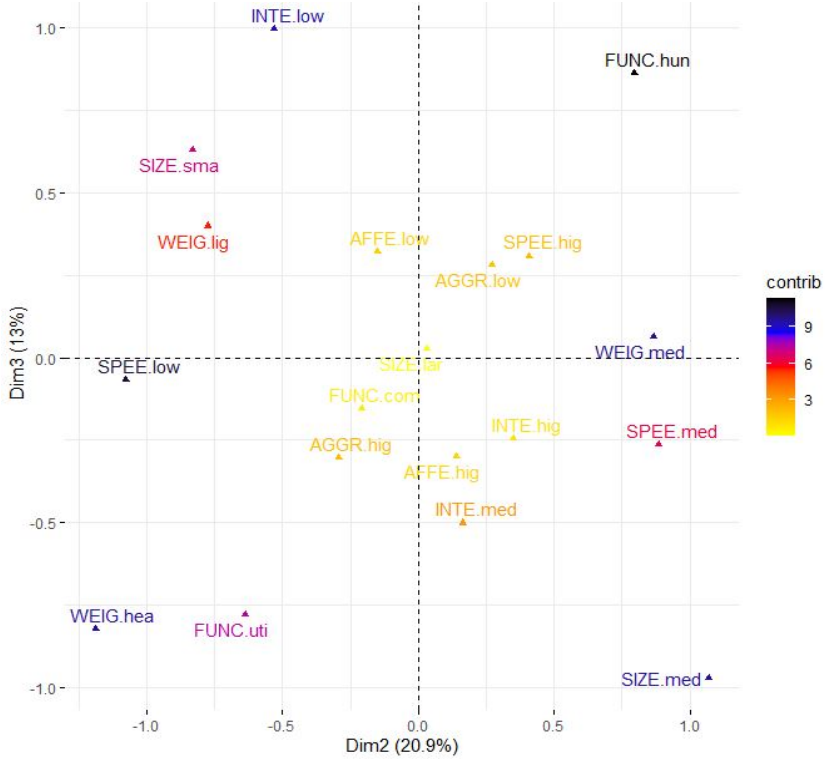


Figura 2.12: Nube proyectada de las categorías al subespacio 2-3



Como podemos ver en estos gráficos, las categorías que aparecen cerca del centro de gravedad (0,0), es decir, las que más se repiten, no están bien representadas por los ejes, por lo que sería conveniente eliminarlas.

En el gráfico 2.10, por ejemplo, podemos observar que las categorías SIZE.med, SPEED.med y WEIGH.med se encuentran próximas. Lo mismo ocurre con WEIG.LIG y SIZE.sma, así como AFFE.hig y FUNC.com, o WEIGH.med y FUNC.hun. Parece que estas categorías suelen aparecer juntas en las razas de perro estudiadas. Por otro lado, en ese mismo gráfico, podemos ver característas que se encuentran en lados opuestos de los ejes, es decir, que son de alguna manera contrarias, por ejemplo, SPEE.hig y WEIG.hea, FUNC.uti y WEIG.lig o SIZE.sma. WEIG.med y SPEE.low, AFFE.hig y SIZE.lar, etc.