Análisis de Correspondencias

Introducción

El *análisis de correspondencias* es una técnica multivariada *exploratoria* para analizar tablas de frecuencias multidimensionales, esto es, tablas de clasificación cruzada de dos o más variables categóricas. El desarrollo del método de correspondencias se centra, generalmente, en las tablas de dos dimensiones; no obstante, el análisis de tablas multidimensionales depende en mucho de las mismas ideas que se desarrollan para las tablas bidimensionales. Esta técnica es el equivalente de componentes principales para variables cualitativas.

Entonces, por su similitud con el análisis de componentes principales, podemos decir que el análisis de correspondencias es una técnica para desplegar de forma gráfica datos categóricos multivariados (generalmente bidimensionales), derivando coordenadas para representar las categorías de las variables que constituyen los *renglones y columnas de una tabla de contingencia*, para plasmar gráficamente la asociación entre estas variables. Entonces, el análisis de correspondencias **(AC)** es:

• Técnica exploratoria para analizar tablas multidimensionales de contingencia o clasificación cruzada, entre dos o más variables categóricas.

• El objetivo es desplegar en una gráfica las asociaciones entre las categorías de una tabla de contingencia. Asociaciones tanto entre renglones, columnas y renglones y columnas. Para descubrir qué categorías están asociadas.

• Es una técnica de reducción de dimensión. Idealmente esperaríamos representar estas asociaciones entre columnas y renglones, en gráficas en dos o tres dimensiones, siempre que con estas pocas dimensiones, se logre una buena representación de ellas.

Análisis de tablas bidimensionales

En este caso, la información está constituída por una matriz de dimensiones $\mathbf{I} \times \mathbf{J}$, que representa las frecuencias absolutas observadas de dos variables cualitativas en una muestra de n elementos. La primera variable representa los renglones de esta tabla, que toma $\mathbf{I}$ valores posibles distintos, y la segunda representa las columnas, y toma $\mathbf{J}$ valores posibles distintos.

Ejemplo. Esta tabla presenta la clasificación de $n = 5387$ escolares escoceses por el color de sus ojos, con cuatro categorías posibles: $\mathbf{I} = 4$, y el color de su cabello, con cinco categorías: $\mathbf{J} = 5$. Esta tabla tiene interés histórico ya que fue utilizada por Fisher en *1940* para ilustrar un método de análisis de tablas de contingencia que está muy relacionado con el que aquí presentamos.

	Color de cabello					
Color de ojos	Rubio	Pelirrojo	Castaño	Obscuro	Negro	Total
claros	688	116	584	188	4	1580
azules	326	38	241	110	3	718
castaños	343	84	909	412	26	1774
oscuros	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Conceptos asociados al análisis de correspondencias

Antes de presentar varios conceptos asociados al análisis de correspondencias, presentemos la forma general de una tabla de contingencia similar a la del ejemplo. Esta tabla tiene la forma

	Y: Variable columna				
X: Variable renglón	y_1	y_2	$\cdots$	y_J	Total
x_1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet J}$	$n_{\bullet\bullet} = n$

con n_{ij} el número de observaciones en el cruce de la categoría i de la variable $\mathbf{X}$ y la categoría j de la variable $\mathbf{Y}$. Además

$$n_{i\bullet} = \sum_{j=1}^{J} n_{ij}, \qquad n_{\bullet j} = \sum_{i=1}^{I} n_{ij} \quad y \quad n_{\bullet\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$$

Perfiles

El concepto de *perfil* (un vector de frecuencias relativas) es de suma importancia en *AC*. Estos vectores de frecuencias relativas tienen características especiales, debido al hecho de que la suma de sus elementos es *uno o el 100%*. En el análisis de estas tablas de frecuencias, consideraremos los vectores de frecuencias relativas *por renglón* y los vectores de frecuencias relativas *por columna*, que llamaremos *perfil renglón* y *perfil columna*, respectivamente. Cuyas definiciones, basándonos en la forma de la tabla general, son

Perfil Renglón

$$\left(\frac{n_{i1}}{n_{i\bullet}}, \frac{n_{i2}}{n_{i\bullet}}, ..., \frac{n_{iJ}}{n_{i\bullet}}\right) \text{ i=1,2,...,I.}$$

que corresponde a un vector de frecuencias relativas de las columnas, por categoría de renglón.

Con *perfil renglón promedio*, dado por

$$\frac{n_{\bullet j}}{n_{\bullet\bullet}} \text{ j=1,2,...,J}$$

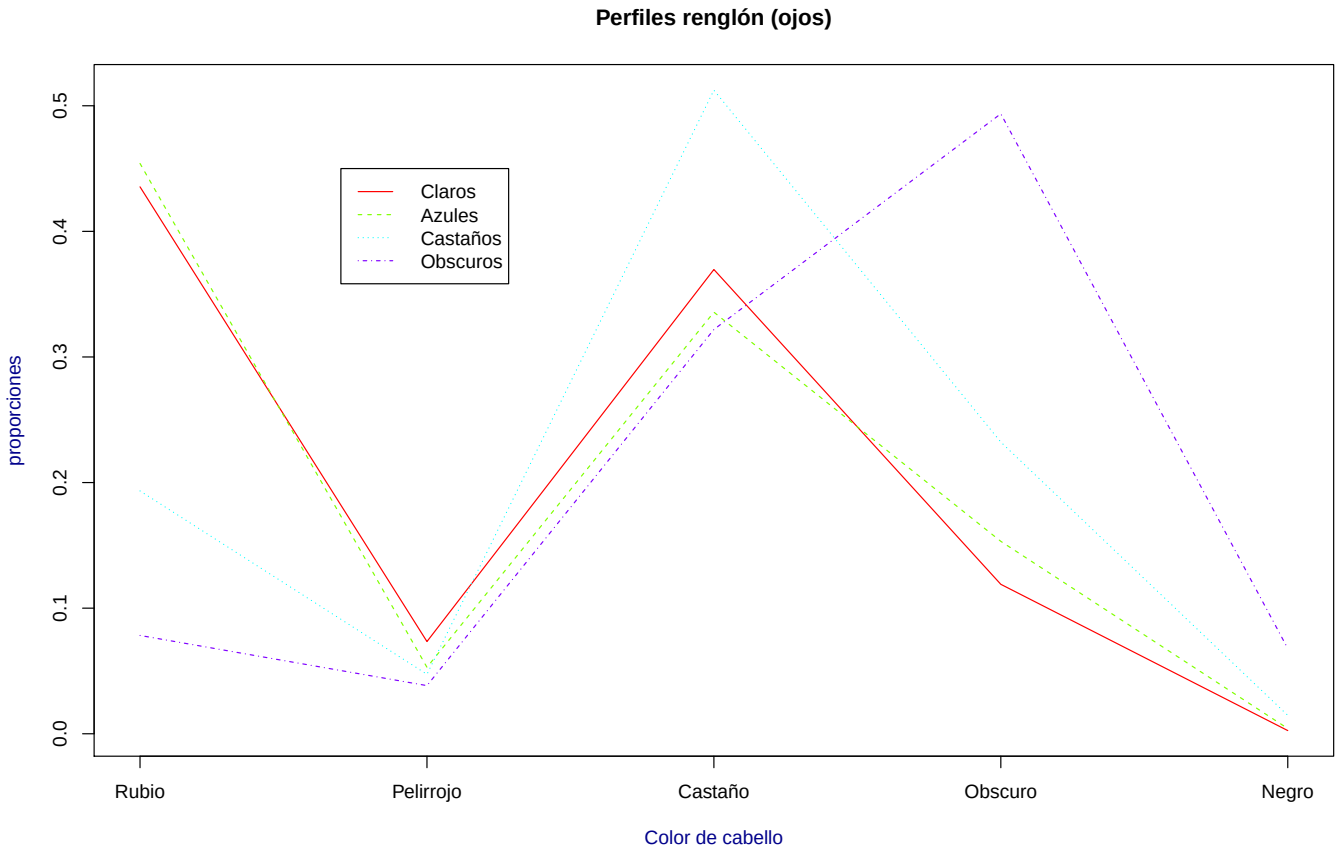
que corresponde al vector de frecuencias relativas del total por columna, entre el total de los datos.

La tabla de perfiles renglón para nuestro ejemplo es:

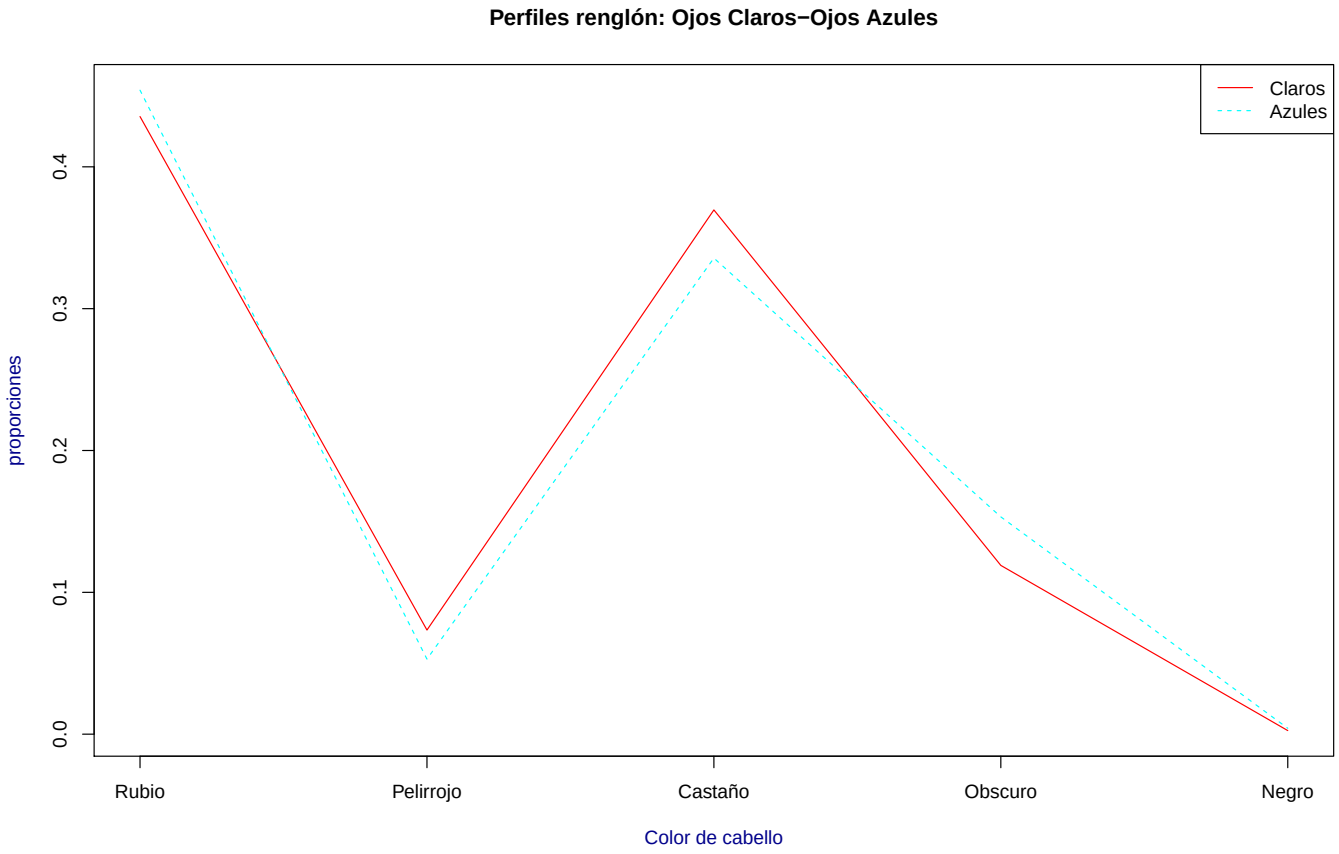
Tabla de Perfiles Renglón

	Color de cabello					
Color de ojos	Rubio	Pelirrojo	Castaño	Obscuro	Negro	Total
claros	0.435	0.073	0.370	0.119	0.003	1
azules	0.454	0.053	0.336	0.153	0.004	1
castaños	0.193	0.047	0.512	0.232	0.015	1
oscuros	0.078	0.038	0.322	0.494	0.068	1
Perfil renglón promedio	0.273	0.054	0.401	0.261	0.022	1
Total	1.161	0.212	1.539	0.998	0.0892	4

Observemos, por ejemplo, que el perfil del renglón: Ojos claros $(0.435, 0.073, 0.370, 0.119, 0.003)$, corresponde a $(688/1580, 116/1580, 584/1580, 188/1580, 4/1580)$, que son los valores observados en ese renglón, entre el total del mismo.



En esta gráfica de los perfiles renglón, podemos observar una gran similitud entre los perfiles de los sujetos de ojos claros y los de ojos azules.



Perfil columna

$$\left(\frac{n_{1j}}{n_{\bullet j}}, \frac{n_{2j}}{n_{\bullet j}}, ..., \frac{n_{2I}}{n_{\bullet j}}\right) \text{ j=1,2,...,J.}$$

que corresponde al vector de frecuencias relativas de los renglones, por categoría de columna.

Con perfil columna promedio

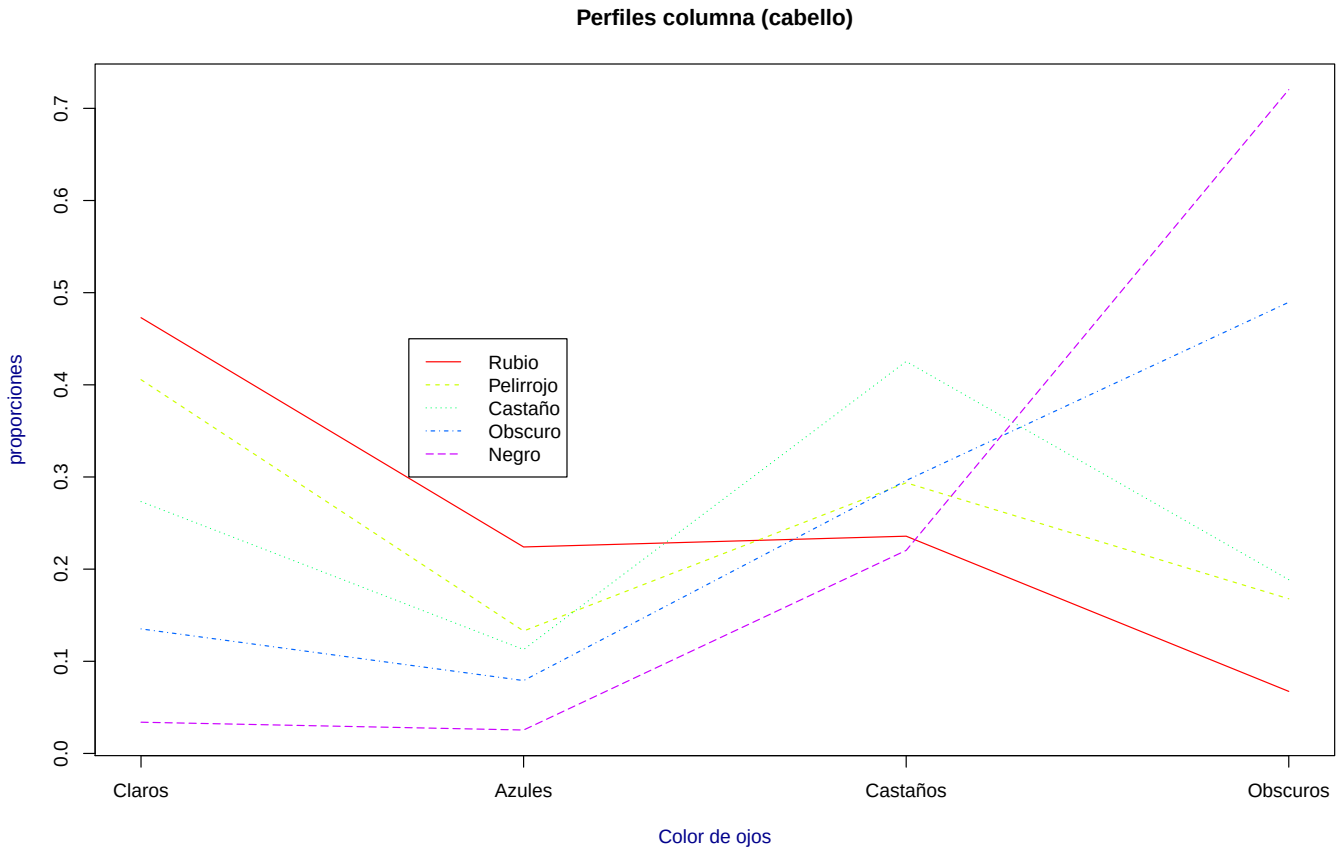
$$\frac{n_{i\bullet}}{n_{\bullet\bullet}} \text{ i=1,2,...,I}$$

que corresponde al vector de frecuencias relativas del total por renglón, entre el total de los datos.

La tabla de perfiles columna para nuestro ejemplo es:

Tabla de Perfiles Columna

	Color de cabello						
Color de ojos	Rubio	Pelirrojo	Castaño	Obscuro	Negro	Perfil columna promedio	Total
Claros	0.473	0.406	0.273	0.142	0.034	1.327	0.2967693
Azules	0.224	0.133	0.113	0.083	0.025	0.578	0.1348610
Castaños	0.236	0.294	0.425	0.310	0.220	1.485	0.3332081
Obscuros	0.067	0.168	0.189	0.465	0.720	1.609	0.2351615
Total	1	1	1	1	1	4	



Masa

En el cálculo usual de la media (no ponderada), todos los puntos tienen la misma masa (o peso). Sin embargo, una media ponderada permite asociar diferentes masas a los diferentes valores (puntos) que la conforman. Cuando ponderamos estos valores de distinta forma, el centriode no se sitúa exactamente en el centro "geográfico" de la nube de puntos, sino que tiende a situarse cerca de los puntos con mayor masa.

Por ejemplo, supongamos que en una clase de 30 estudiantes, la media de calificaciones, calculada sumando sus calificaciones y dividiendo entre 30, es 7.43. Se sabe que tres estudi-

antes obtuvieron 9 de calificación, siete obtuvieron 8 y 20 obtuvieron 7, entonces, podemos calcular la media de manera equivalente, asignando un peso de 3/30 a la calificación de 9, 7/30 a la de 8 y 20/30 a la de 7. Dado que la calificación de 7 tiene mayor peso que las otras, el valor de la media ponderada, 7.43, se encuentra "más cerca" de esta calificación. La media aritmética usual de los valores 7, 8 y 9 es 8.

En el AC, los pesos asignados a los perfiles reciben el nombre de *masas*. Los totales de las columnas, con relación al total de la tabla, son las masas de las columnas que asignaremos a los perfiles columna. En símbolos

$$m_j = \frac{n_{\bullet j}}{n_{\bullet\bullet}} \quad j = 1, 2, ..., J$$

El perfil columna promedio, está constituido por los totales de las columnas, divididos entre el total de la tabla.

De manera semejante, los totales de los renglones, con relación al total de la tabla, son las masas de los renglones que asignaremos a los perfiles renglón.

$$m_i = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \quad i = 1, 2, ..., I$$

El perfil renglón promedio, está constituido por los totales de los renglones, divididos entre el total de la tabla.

Entonces, un perfil con mayor masa se hubicará más cercano a su correspondiente perfil columna promedio o perfil renglón promedio, según corresponda a un perfil columna o renglón.

Distancias entre perfiles

Una forma de saber qué tan parecido es un renglón con respecto a otro, o una columna con respecto a otra, es a través de la distancia entre sus correspondientes perfiles.

La distancia χ^2

En AC la forma de calcular la distancia entre perfiles es un poco complicada, y se realiza a través de la llamada *distancia* χ^2. Existen varias maneras de justificar el uso de esta

distancia, algunas más técnicas que no viene al caso considerar, y otras más intuitivas que utilizaremos aquí.

Recordemos que la estadística χ^2 asociada a estas tablas de contingencia tiene la forma

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \mathbb{E}_{ij})^2}{\mathbb{E}_{ij}}$$

Reflexionemos un poco sobre esta estadística

Mencionamos que el objetivo en este análisis de correspondencias, es determinar las asociaciones entre los distintos elementos de una tabla de contingencia, a saber, entre renglones, entre columnas y entre renglones y columnas. Recordemos que uno de los usos de la estadística χ^2, es probar la asociación *a nivel global* entre dos variables categóricas. Al comparar los valores observados en la tabla contra los esperados, estos últimos se calculan bajo el supuesto de que las variables involucradas *son independientes*; entonces, es claro que la discrepancia entre los valores observados y esperados

$$n_{ij} - \mathbb{E}_{ij}, \;\; i = 1, 2, ..., I \;\; j = 1, 2, ..., J$$

es una cantidad que evidencia el grado de *no inpendencia* entre las celdas correspondientes, i.e., es, en algún sentido, una medida del grado de asoción entre estas celdas, lo que conduce a que el valor de esta χ^2, sea una medida del nivel de asociacion global de las variables. Recordemos también que, bajo el supuesto de indpendencia entre las variables de la tabla, los valores esperados se calculan como

$$\mathbb{E}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}, \;\; i = 1, 2, ..., I \;\; j = 1, 2, ..., J$$

Retomando la expresión de esta estadística, observemos que la podemos escribir como

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{i\bullet}\left(\dfrac{n_{ij}}{n_{i\bullet}} - \dfrac{n_{\bullet j}}{n_{\bullet\bullet}}\right)\right)^2}{\dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{i\bullet}^2\left(\dfrac{n_{ij}}{n_{i\bullet}} - \dfrac{n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{i\bullet}\left(\dfrac{n_{ij}}{n_{i\bullet}} - \dfrac{n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\dfrac{n_{\bullet j}}{n_{\bullet\bullet}}}
\end{aligned}
$$

Ahora, obsérvese que

$\dfrac{n_{ij}}{n_{i\bullet}}$ es el perfil renglón para $i = 1, 2, ..., I$ y

$\dfrac{n_{\bullet j}}{n_{\bullet\bullet}}$ el perfil renglón promedio o perfil esperado.

Por lo tanto, esta χ^2 la podemos reescribir, como

$$
\sum_i \text{total renglón } i \times \frac{(\text{perfil observado renglón } i\text{-perfil esperado renglón } i)^2}{\text{perfil esperado renglón } i}
$$

justamente como una distancia entre los perfiles renglón y su perfil promedio.

De manera similar, si en el segundo paso del desarrollo anterior, factorizamos $n_{\bullet j}$, en lugar de $n_{i\bullet}$, obtenemos

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

Que podemos reescribir, como

$$\sum_j \text{total columna } j \times \frac{(\text{perfil observado columna } j\text{-perfil esperado columna } j)^2}{\text{perfil esperado columna } j}$$

justamente como una distancia entre los perfiles columna y su perfil promedio.

Inercia

Otro de los conceptos importantes, de hecho, muy importante, en AC es el de la *inercia*. Desde el punto de vista de la Física, en particular de la mecánica, que es de donde se traslada este concepto, se tiene que cualquier objeto tiene un centro de gravedad (centroide); cualquier partícula en el objeto tiene cierta masa y cierta distancia al centroide; entonces, el momento de inercia está dado por $I = md^2$ sumado sobre todas las partículas que constituyen el objeto. Es decir:

$$\mathbf{I} = \sum \mathbf{md^2}$$

El concepto análogo en AC consiste en considerar a los puntos perfiles, cuya masa suma uno. Estos puntos tienen un centroide (su perfil promedio) y una distancia (distancia ji-cuadrada) entre puntos perfiles. Cada punto en un perfil contribuye a la inercia en la nube total de puntos.

Ahora, hagamos la deducción analítica de este concepto partiendo de la expresión de la distancia χ^2 escrita como distancias entre perfiles. En esta expresión, dividamos ambos lados de la igualdad entre el total de la muestra, $n_{\bullet\bullet} = n$, con lo que obtenemos

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

obsérvese que $\frac{n_{i\bullet}}{n}$ corresponde a la masa de cada perfil renglón; entonces tenemos una expresión semejante a la que define la inercia. En este caso *la masa es la masa de cada perfil*, *el centroide es el perfil promedio o medio* y *la distancia del perfil a este centroide está dada*

por la distancia χ^2. En AC a esta cantidad se le conoce con el nombre de *inercia* o *inercia total*.

Dada esta relación

$$\frac{\chi^2}{n} = \mathbf{I}$$

es claro que la inercia es una medida de la varianza o variabilidad de nuestros datos. De hecho, sabemos que es una medida de la asociación entre los categorías en la tabla. Además, al escribirla como una medida de la discrepancia entre un perfil y su perfil medio, también es una medida de qué tan "lejos" se hallan los perfiles renglón o columna de su perfil medio. Podemos considerar que este pefil medio representa la hipótesis de homogeneidad, en este caso, de homogeneidad entre los perfiles. Entonces, debe ser claro que si los perfiles difieren poco de sus perfiles medios, el valor de la inercia sería bajo, e implicaría una pobre asociación entre las variables, así como entre los renglones, columnas, y renglones columnas de la tabla de contingencia.

Reducción de dimensión

La dimensión natural de una tabla de contingencia de $\mathbf{I} \times \mathbf{J}$ es $min\,(\mathbf{I}-1, \mathbf{J}-1)$, así que si la menor de estas dimensiones es grande, entonces debemos hacer una reducción de dimensión, idealmente a *dos* o *tres* dimensiones para poder representarlas gráficamente, de manera que la varianza explicada, en este caso la *inercia explicada*, por esas pocas dimensiones sea cercana al 100%.

Descomposición en valor singular

En muchas de las técnicas multivariadas, la información se concentra en la matriz de varianza-covarianza o de correlación, ¿existe un concepto semejante en AC?

Matriz asociada a la χ^2. Ya vimos que la información relevante para las asociaciones de las categorías de la tabla, la proporcionan las diferencias entre los perfiles observados y los perfiles medios, estos últimos asumiendo que los perfiles medios representan la homogeneidad entre los perfiles correspondientes. Equivalentemente, esta matriz también representa las asociaciones de las categorías de la tabla, medidas a través de la diferencia entre los valores esperados, que se obtienen bajo el supuesto de independencia de esta tabla, y los valores observados. Entonces, consideremos la matriz

$$\mathbf{S} = (s_{ij}) \qquad \text{con } s_{ij} = \frac{n_{ij} - \mathbb{E}_{ij}}{\sqrt{\mathbb{E}_{ij}}} = \frac{n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}}{\sqrt{\frac{n_{i\bullet} n_{\bullet j}}{n}}} \;\; i = 1, 2, ..., I \;\; J = 1, 2, ..., J$$

que constituye nuestra "matriz de correlación".

Por lo que, nuestro objetivo de representar gráficamente las asociaciones, se convierte en el de encontrar un número reducido de dimensiones (idealmente 2 ó 3) donde se puedan representar estas desviaciones expresadas en "nuestra matriz de correlación". En este sentido, este objetivo es similar al de componentes principales, factores, etc. Para lograrlo, debemos hacer la *descomposición en valor singular*, de esta matriz.

$$\mathbf{S} = \mathbf{ULA}^{'} \text{ con } \mathbf{G} = \mathbf{UL} \;\; \mathbf{y} \;\; \mathbf{H} = \mathbf{A}'$$

entonces, los elementos de $\mathbf{S}$ se pueden escribir como

$$s_{ij} = \sum_{k=1}^{R} \lambda_k^{1/2} g_{ik} h_{ik}, \quad i = 1, 2, ..., I \;\; j = 1, 2, ..., J$$

con $R = min(I-1, J-1)$, el número máximo de dimensiones que se pueden tener, y que corresponde al rango de $\mathbf{S}$. Los valores g_{ik} y h_{ik} son los elementos de la k-ésima columna de $\mathbf{G}$ y $\mathbf{H}$, respectivamente. $\lambda_1, \lambda_2, ..., \lambda_R$ son los eigenvalores de $\mathbf{S}$, que constituyen la matriz $\mathbf{L}$.

Obsérvese que

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} s_{ij}^2 = \mathbf{SS}' = \mathbf{ULA}' \left(\mathbf{ULA}'\right)' = \mathbf{ULA}'\mathbf{AL}'\mathbf{U}' = \mathbf{ULL}'\mathbf{U}' = \mathbf{L^2UU}' = \mathbf{L^2}$$

$$\Rightarrow \; traza\left(\mathbf{SS}'\right) = traza(\mathbf{LL}') = traza(\mathbf{L^2}) = \sum_{k=1}^{R} \lambda_k$$

por lo tanto, la variabilidad de la matriz asociada a la tabla de contingencia, es igual a la suma de sus eigenvalores. Con lo que tenemos una analogía completa con el proceso de componentes principales.

Entonces, lo que deseamos es poder representar estos elementos de $\mathbf{S}$, en pocas dimensiones, pero asegurándonos de que esta representación es buena, en algún sentido. Si queremos una representación bidimensional, entonces

$$s_{ij} \approx \sum_{k=1}^{2} \lambda_k^{1/2} g_{ik} h_{ik} \quad i = 1, 2, ..., I \;\; j = 1, 2, ..., J$$

¿Y cómo determinamos si, en este caso, dos dimensiones proporcionan una buena aproximación de los elementos de esta matriz?

Similar al criterio de componentes principales, consideramos el porcentaje de inercia que explican estas dos dimensiones

$$\frac{\sum_{k=1}^{2} \lambda_k}{\sum_{k=1}^{R} \lambda_k} \times 100\%$$

debe ser grande, idealmente cercano al 100%.

Representación bidimensional

En la práctica, es común representar el AC en una gráfica bidimensional. Aunque se pueden proyectar los datos en cualquier subespacio de dimensión menor, las proyecciones bidimensionales son particularmente atractivas, ya que representan nuestra forma habitual de representar una gráfica.
En este caso, cada categoría de un renglón estará representada por un par coordenado (g_{i1}, g_{i2}) $i = 1, 2, ..., I$. Y cada categoría de una columna, por el par coordenado (h_{j1}, h_{j2}) $j = 1, 2, ..., J$. Por lo que, para representarlos hay que desplegarlos en una gráfica bidimensional.

Representaciones bidimensionales

• Representación bidimensional de los renglones

• Representación bidimensional de las columnas

• Representación bidimensional de renglones y columnas (generalmente la más usual. *Biplot*).

Interpretación del AC

Dimensiones: Algunas veces es posible interpretar o "dar nombre" a las dimensiones que se obtienen a través del **AC**. Podemos examinar la posición de las categorías renglones/columnas en cada dimensión y analizar qué tienen en común las categorías de estos renglones/columnas que aparecen juntas, y qué distingue a aquéllas que aparecen separadas. Sin embargo, cuando se interpreta una dimensión, es importante prestar particular atención a aquellos puntos que contribuyen más a la inercia de cada dimensión.

Podemos particionar la contribución de cada punto a la inercia total, en su contribución a la inercia de cada dimensión. La cantidad de inercia de la k-ésima dimensión explicada por renglón i es

$$\frac{(masa\ del\ \text{renglón}\ i) * g_{ik}^2}{\sqrt{\lambda_k}} \quad \text{equivalentemente} \quad \frac{(masa\ de\ la\ columna\ j) * h_{ik}^2}{\sqrt{\lambda_k}}$$

entonces, puntos correspondientes a renglones con una gran masa y grandes coordenadas en

la k-ésima dimensión, contribuirán más a la inercia de esta dimensión. Puntos con una relativa mayor contribución a la inercia de una dimensión, son más importantes para la misma y proporcionan la clave para su correspondiente interpretación.

Puntos

Como debe ser obvio, categorías de renglones que tienen un perfil similar deben aparecer cercanos en la representación bidimensional. Misma situación para categorías de columnas.

Asociación entre renglones y columnas: En este caso, la distancia entre una categoría renglón y otra columna, no representan ninguna similitud de los mismos. Para interpretar puntos de distintas naturalezas, se recurre a su llamada representación en *biplot*.

Como sabemos, el *biplot* se basa en el producto escalar entre los vectores columna y renglón, por lo que depende más de las longitudes y ángulos formados por estos vectores que de la distancia entre los puntos.

Geométricamente, el producto escalar entre vectores es igual al producto de las longitudes de los vectores multiplicado por el coseno del ángulo formado entre ellos, es decir

$$\mathbf{x}'\mathbf{y} = \sum_i x_i y_i = \|\mathbf{x}\| \, \|\mathbf{y}\| \, cos\,(\theta)$$

Recordemos también que la proyección perpendicular de un vector $\mathbf{x}$ sobre la dirección definida por otro vector $\mathbf{y}$, tiene una longitud igual al producto de la longitud de $\mathbf{x}$, multiplicada por el ángulo que forman estos vectores. En concreto

$$Proy\; x_y = \|x\| \, cos\,(\theta)$$

Entonces, en el AC, la idea es que a través de los productos escalares del biplot, podamos recuperar de manera aproximada, los elementos de la matriz dada por la tabla de contingencia o clasificación cruzada.

En la palabra biplot, el prefijo *bi* indica que en el mapa se representan conjuntamente renglones y columnas, pero no indica que el mapa sea bidimensional, ya que los biplots pueden tener cualquier dimensionalidad. No obstante, lo más frecuente es una representación en dos

dimensiones.

Recordemos que $g_{i2} = (g^*_{i1}, g^*_{i2})$ es un punto en dimensión dos que representa el i-ésimo renglón (i=1,2,...,I), y $h_{i2} = (h^*_{i1}, h^*_{i2})$ es un punto bidimensional que representa a la j-ésima columna (j=1,2,...,J). Utilizando los conceptos asociados al biplot, el producto $g^*_{ik} \times h^*_{ik}$ representa la contribución conjunta del renglón i y la columna j al residuo (que dijimos que era una medida de asociación) en la dimensión k, es decir, la "asociación" entre el renglón i y la columna j. O, de forma más precisa, la contribución del renglón i y la columna j, a la asociación global medida por la χ^2.

En este sentido, un valor grande y positivo de $g^*_{ik} \times h^*_{ik}$ indica una asociación positiva entre el renglón i y la columna j en la dimensión k, misma que se obtiene si ambos son grandes y positivos, o grandes y negativos.

Si $g^*_{ik} \times h^*_{ik}$ proporciona un valor grande y negativo, implica una asociación negativa entre el rengón i y la columna j en la dimensión k, misma que se obtiene si ambos son grandes y uno es positivo y el otro negativo.

Un valor cercano a cero de $g^*_{ik} \times h^*_{ik}$ indica que no hay asociación entre el renglón i y la columna j en la dimensión k, que se obtiene si alguno o ambos están cercanos a cero en esa dimensión.

Calidad de la representación de un punto. Dado que se han elegido un número reducido de dimensiones para representar un punto, una pregunta de interés es saber qué calidad de representación tiene cada punto en estas pocas dimensiones. La calidad de esta representación se mide a través del cociente entre la distancia al origen del punto en las dimensiones elegidas, y su distancia al origen en el máximo de dimensiones posibles ($min(I-1, J-1)$). Si un punto tiene baja calidad, implica que su representación en este espacio reducido, no es adecuada.

Correspondencias múltiples

El *análisis de correspondencias múltiples (MCA)*, es una extensión del *AC* que permite analizar las asociaciones entre más de dos variables categóricas.

Un punto importante en *MCA* es la forma en que se debe manejar la información de una tabla de frecuencias multidimensional. La manera de hacerlo es a través de la llamada *matriz indicadora* o *matriz disjunta* que no es mas que una matriz cuyos valores son sólo "0" ó "1". Entonces, el *MCA* consiste en analizar una serie de observaciones descritas por un conjunto de *variables nominales* o *variables dummy's*.

Ejemplo: Enfermedad de Hodgkin. En este caso sólo tenemos una tabla bidimensional, pero servirá para ilustrar la construcción de la matriz indicadora o disjunta.

Enfermedad de Hodgkin

	Tipo de respuesta			
Tipo histológico	Positiva	Parcial	Nula	Total
LP	74	18	12	104
NS	68	16	12	96
MC	154	54	58	266
LD	18	10	44	72
Total	314	98	126	538

La matriz indicadora correspondiente

Matriz indicadora: Enfermedad de Hodgkin

	Tipo histológico				Respuesta		
Sujeto	LP	NS	MC	LD	Positiva	Parcial	Nula
1	1	0	0	0	1	0	0
2	1	0	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
74	1	0	0	0	1	0	0
75	1	0	0	0	0	1	0
76	1	0	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
92	1	0	0	0	0	1	0
93	1	0	0	0	0	0	1
94	1	0	0	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Extensión a tres variables categóricas. Si aumentamos un criterio de clasificación referente, por ejemplo, al género, la tabla se ampliaría de la siguiente manera

Matriz indicadora: Enfermedad de Hodgkin con tres variables categóricas

	Tipo histológico				Respuesta			Género	
Sujeto	LP	NS	MC	LD	Positiva	Parcial	Nula	F	M
1	1	0	0	0	1	0	0	1	0
2	1	0	0	0	1	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
74	1	0	0	0	1	0	0	0	1
75	1	0	0	0	0	1	0	1	0
76	1	0	0	0	0	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
92	1	0	0	0	0	1	0	0	1
93	1	0	0	0	0	0	1	1	0
94	1	0	0	0	0	0	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Entonces, el punto inicial del análisis de correspondencias múltiples es construir la matriz indicadora, $\mathbf{Z}$. Cada renglón de esta matriz tiene k *"unos"* y *C-k "ceros"*, donde k es el número de variables categóricas en cuestión, y C es el total de categorías de estas k variables, esto es

$$C = \sum_{i=1}^{k} c_i$$

con c_i el número de categorías de la *i-ésima variable*. Por lo tanto, la matriz indicadora, $\mathbf{Z}$, es de tamaño *(n,k)*, donde k es el número total de variables. Las suma de cada una de sus filas es igual a k, el número de variables, y la suma de cada columna es el número de individuos que que tiene la característica en cuestión.

Para una tabla de contingencia con k variables, la matriz indicadora puede escribirse como:

$$\mathbf{Z} = [\mathbf{Z}_1 \ \mathbf{Z}_2 \ \cdots \ \mathbf{Z}_k]$$

con $\mathbf{Z}_i$ es la matriz de $n_i \times c_i$ de la i-ésima tabla de contingencia.

La matriz de Burt

La matriz

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z}$$

se conoce como la *matriz de Burt*, y contiene las submatrices $\mathbf{Z}_i'\mathbf{Z}_j$ de la tabla de contingencia bidimensional, basada en las variables i y j. Esto es

$$\mathbf{B} = \begin{bmatrix} \mathbf{Z}_1'\mathbf{Z}_1 & \mathbf{Z}_1'\mathbf{Z}_2 & \cdots & \mathbf{Z}_1'\mathbf{Z}_k \\ \mathbf{Z}_2'\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{Z}_2 & \cdots & \mathbf{Z}_2'\mathbf{Z}_k \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{Z}_k'\mathbf{Z}_1 & \mathbf{Z}_k'\mathbf{Z}_2 & \cdots & \mathbf{Z}_k'\mathbf{Z}_k \end{bmatrix}$$

Entonces, el análisis de correspondencias múltiple consiste, esencialmente, en aplicar todos los procesos de correspondencias simples, ya sea a la matriz indicadora o a la matriz de Burt.

La matriz o tabla **B** es simétrica y está conformada por $k \times k$ subtablas. Las k subtablas diagonales son a su vez diagonales y contienen las frecuencias marginales de cada una de las variables. Las subtablas fuera de esta diagonal, son las tablas de contingencia entre parejas de variables.

Ejemplo de una matriz de Burt con tres variables categóricas

Matriz de Burt

	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
A1	119	0	0	0	0	27	28	30	22	12	49	40	18	7	5
A2	0	322	0	0	0	38	74	84	96	30	67	142	60	41	12
A3	0	0	204	0	0	3	48	63	73	17	18	75	70	34	7
A4	0	0	0	178	0	3	21	23	79	52	16	50	40	56	16
A5	0	0	0	0	48	0	3	5	11	29	2	9	9	16	12
B1	27	38	3	3	0	71	0	0	0	0	43	19	4	3	2
B2	28	74	48	21	3	0	174	0	0	0	36	88	34	15	1
B3	30	84	63	23	5	0	0	205	0	0	37	90	57	19	2
B4	22	96	73	79	11	0	0	0	281	0	27	88	75	74	17
B5	12	30	17	52	29	0	0	0	0	140	9	31	27	43	30
C1	49	67	18	16	2	43	36	37	27	9	152	0	0	0	0
C2	40	142	75	50	9	19	88	90	88	31	0	316	0	0	0
C3	18	60	70	40	9	4	34	57	75	27	0	0	197	0	0
C4	7	41	34	56	16	3	15	19	74	43	0	0	0	154	0
C5	5	12	7	16	12	2	1	2	17	30	0	0	0	0	52

Algunas características de estas matrices

Entonces, las características de estas matrices son:

- La matriz es $\mathbf{Z} = (z_{ij})$ con $z_{ij} = \begin{cases} 1 \\ 0 \end{cases}$
- $\sum_i \sum_j z_{ij} = nk.$
- n: número de individuos.
- k: número de variables categóricas.

- c_i: número de categorías de la variable i, i=1,2,...,k.

- C=$\sum_{i=1}^{k} c_i$: total de categorías.

- *Marginales*: $z_{i\bullet} = k$ (puesto que hay k variables, y por lo tanto, k *uno's* por renglón).

- *Marginales*: $z_{\bullet j} =$ individuos que tienen la característica j.

- *Perfil renglón*: $\frac{z_{ij}}{z_{i\bullet}} = \frac{z_{ij}}{k}$.

- *Masa del perfil renglón*: $\frac{z_{i\bullet}}{nk} = \frac{k}{nk} = \frac{1}{n}$

- *Perfil columna*:$\frac{z_{ij}}{z_{\bullet j}}$

- *Masa del perfil columna*: $\frac{z_{\bullet j}}{nk}$

Una manera de definir la distancia χ^2 entre los perfiles es a través de las frecuencias relativas de la tabla, de la siguiente manera

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

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

en este caso, con $f_{ij} = \frac{z_{ij}}{nk}$, $f_{i\bullet} = \frac{z_{i\bullet}}{nk}$ y $f_{j\bullet} = \frac{z_{\bullet j}}{nk}$

Observemos que se pondera las diferencias cuadradas de los perfiles renglón o columna, por el inverso de su frecuencia, lo que hace que perfiles con poca frecuencia, contribuyan de manera similar a los que tienen mayor frecuencia en la tabla. De hecho, lo único que se está

haciendo es dotar de una métrica distinta a estas distancias.

Entonces, en este caso de correspondencias múltiples, estas distancias son

- $d^2\left(i,i^{'}\right)=\sum_j \frac{nk}{z_{\bullet j}}\left(\frac{z_{ij}}{z_{i\bullet}}-\frac{z_{i^{'}j}}{z_{i^{'}\bullet}}\right)^2=\sum_j \frac{nk}{z_{\bullet j}}\left(\frac{z_{ij}}{k}-\frac{z_{i^{'}j}}{k}\right)^2=\frac{n}{k}\sum_j \frac{1}{z_{\bullet j}}\left(z_{ij}-z_{i^{'}j}\right)^2$

- $d^2\left(i,j^{'}\right)=\sum_i \frac{nk}{z_{i\bullet}}\left(\frac{z_{ij}}{z_{\bullet j}}-\frac{z_{ij^{'}}}{z_{\bullet j}}\right)^2=\sum_i \frac{nk}{k}\left(\frac{z_{ij}}{z_{\bullet j}}-\frac{z_{ij^{'}}}{z_{\bullet j}}\right)^2=n\sum_i \left(\frac{z_{ij}}{z_{\bullet j}}-\frac{z_{ij^{'}}}{z_{\bullet j}}\right)^2$

Análisis de estas distancias

Podemos reescribir la distancia entre los perfiles renglón de la siguiente manera:

$$\frac{n}{k}\sum_j \frac{1}{z_{\bullet j}}\left(z_{ij}-z_{i^{'}j}\right)^2=\frac{n}{k}\sum_{j\in M_{ii^{'}}} \frac{1}{z_{\bullet j}}$$

con $M_{ii^{'}}$ modalidades que poseen sólo un individuo i ó $i^{'}$. Entonces, los perfiles serán más parecidos (distancia más pequeña), conforme poseean más modalidades en común.

Perfiles columna

$$n\sum_i \left(\frac{z_{ij}}{z_{\bullet j}}-\frac{z_{ij^{'}}}{z_{\bullet j^{'}}}\right)^2=n\frac{\#\left(ind[j,no\ j^{'}]\right)\#\left(ind[j^{'},no\ j]\right)}{z_{\bullet j}z_{\bullet j^{'}}}$$

Entonces, entre más objetos tengan sólo una de j o $j^{'}$ mayor es la distancia.

Interpretación

- Dos modalidades escogidas por los mismos individuos coinciden
- Dos individuos son cercanos si escogen las mismas modalidades
- Modalidades con poca frecuencia están alejadas del centro de gravedad

Inercia en ACM

- *Centroide*: $G = \left(\frac{1}{n}, ..., \frac{1}{n}\right)$

- *Distancia del perfil columna al centroide*

$$d^2(j, G) = n \sum_i^n \left(\frac{z_{ij}}{z_{\bullet j}} - \frac{1}{n}\right)^2 = n \sum_i^n \left(\frac{z_{ij}}{z_{\bullet j}^2} - 2\frac{z_{ij}}{z_{\bullet j}} + \frac{1}{n^2}\right) = \frac{n}{z_{\bullet j}} - 1 \quad \left(\text{recordando que } \sum_i^n z_{ij} = z_{\bullet j}\right)$$

  Cuya distancia es más grande si $z_{\bullet j}$ es pequeña.

- *Inercia de un perfil columna*

$$\mathbf{I}(j) = \frac{z_{\bullet j}}{nk} d^2(j, G) = \frac{z_{\bullet j}}{nk}\left(\frac{n}{z_{\bullet j}} - 1\right) = \frac{1}{k}\left(1 - \frac{z_{\bullet j}}{n}\right)$$

  Mayor inercia si $z_{\bullet j}$ es pequeña.

- *Inercia de la k-ésima variable*

$$\mathbf{I}_k = \sum_{j=1}^{c_k} I(j) = \sum_{j=1}^{c_k} \frac{1}{k}\left(1 - \frac{z_{\bullet j}}{n}\right) = \frac{1}{k}(c_k - 1)$$

  que crece con el número de categorías.

- *Inercia total*

$$\mathbf{I} = \sum_k \mathbf{I}_k = \sum_k \frac{1}{k}(c_k - 1) = \frac{1}{k}(C - k) = \frac{C}{k} - 1$$

  que no tiene ningún significado estadístico.

Como mencionamos, el *ACM* es una extension del análisis de correspondencias simples (ACS), y se basa en un *ACS* de la matriz indicadora, $\mathbf{Z}$, o de la matriz de Brurt, $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$.

En el *ACM* la identificación de la verdadera dimensión de los datos (de la tabla) es particularmente difícil, pese a que un *MCA* es un *ACS* de una tabla particular, porque la prueba ji-cuadrada no tiene sentido. Es decir, para la matriz de Burt $\mathbf{B}$, se puede calcular la estadística χ^2 como si fuera una tabla de contingencia usual, y esta estadística puede simplificarse a

$$\chi^2_{\mathbf{B}} = 2\sum_{k}^{i=1}\sum_{i-1}^{j=1}\chi^2_{ij} + n(C-k)$$

con χ^2_{ij} la estadística ji-cuadrada de la subtabla $\mathbf{Z}_i^{'}\mathbf{Z}_j$, $i \neq j$. Pero, desafortunadamente, la correspondiente estadística χ^2 de independencia, calculada con la matriz $\mathbf{Z}$ *no se distribuye como una ji-cuadrada.*

Por otro lado, ya que el *ACM* codifica cada variable en varias variables binarias, entonces, este esquema de codificación, crea, *de manera artificial, dimensiones adicionales a la tabla*, ya que una variable categórica se codifia en múltiples columnas. Como consecuencia de esto, la inercia (i.e. la varianza) se infla de manera artificial y por lo tanto el porcentaje de inercia explicado por la primer dimensión (de hecho, por pocas dimensiones) es severamente subestimado.

El término *inflación* que se aplica al alto número de eigenvalores del *MCA*, fue derivado por Benzécri (1979) que lo explica en términos del arbitrario número de niveles en que una característica continua puede discretizarse hacerla cualitativa (discreta o categórica), y el hecho que, si comparamos el *ASC* y el *ACM* aplicado a la misma tabla de contingencia bidimensional, es posible encontrar una relación entre los eigenvalores. Es decir, al particionar una tabla de Burt, $\mathbf{B} = \mathbf{Z}^{'}\mathbf{Z}$, de dos variables discretas en submatrices, se puede demostrar la relación

$$\mu_\alpha = \frac{1 \pm \sqrt{\lambda_\alpha}}{2}$$

que se cumple entre los eigenvalores de $\mathbf{Z}$ (μ_α) y los del *ACS* de la tabla de contingencia entre estas dos variables (λ_α). En este caso, es evidente que los eigenvalores, $\lambda_\alpha = 0$, del

ACS corresponden a eigenvalores, $\mu_\alpha = \frac{1}{2}$, del *ACM* de $\mathbf{Z}$, y $\nu_\alpha = \frac{1}{4}$ de $\mathbf{B}$ [1], mientras que los otros dos eigenvalores son uno mayor y otro menor que $\frac{1}{2}$ y $\frac{1}{4}$, respectivamente. Realizando la generalización de este argumento a $k > 2$ variables categóricas, resulta que hay que limitar la atención en *ACM* únicamente a aquellos eigenvalores mayores que su media, esto es

$$\mu_\alpha \geq \bar{\mu}_\alpha = \frac{1}{k}$$

Este argumento es discutido por Benzécri (1979) y Greenacre (1988, 2006). Ambos autores sugieren, para dar una medida de importancia a cada dimensión, reevaluar los vectores propios más grandes que su media, de la siguiente manera

$$\mu_\alpha \begin{cases} \left[\left(\frac{k}{k-1}\right)\left(\lambda_\alpha - \frac{1}{k}\right)\right]^2, & si\ \ \mu_\alpha \geq \bar{\mu}_\alpha = \frac{1}{k} \\ 0 \qquad\qquad\qquad\qquad\qquad , & si\ \ \mu_\alpha < \frac{1}{k} \end{cases}$$

Entonces, utilizando esta fórmula se tiene una mejor estimación de la inercia extraida por cada eigenvalor. Benzécri sugiere considerar el total de inercia, como la suma de los eigenvalores reevaluados, y tomar como porcentaje de inercia explicado por un de estos eigenvalores, como el cociente

$$\frac{\mu_\alpha}{\sum_\alpha \mu_\alpha}$$

Esto resulta en una *dramática* reevaluación de la importancia relativa del primer eigenvalor.

Análisis de correspondencias conjunto

Greenacre (1988) critica el enfoque *ACM* ya que en su opinión "no es natural generalización de la geometría [...] de la aproximación de mínimos cuadrados del [de SCA] " y propone el *análisis de correspondencias conjunto* (JCA) como su generalización natural en el caso de

[1] Ya que el análisis se puede hacer sobre la matriz $\frac{\mathbf{Z}}{k}$, donde k es el número de variables categóricas en la tabla

los datos nominales, considerándolo como un conjunto de tablas de contingencia obtenidas cruzándolas sobre los mismos individuos. Según él, "en el *ACM* no parece haber justificación para el ajuste de las subtablas en la diagonal de la matriz de Burt, **B**, que contribuyen el término *n(C-k)* en la variación total ", un término que "infla artificialmente la variabilidad total, puede ocasionar que el porcentaje de varianza explicada por los ejes principales pueda ser muy baja, especialmente si *J-Q* es grande. Una medida más natural del total de variación es la suma

$$\sum_i \sum_{i \neq j} \chi^2_{ij}$$

Esto sugiere una alternativa en la generalización del análisis de correspondencias, que ajusta sólo las tablas de contingencia *fuera de la diagonal*, análogo a análisis de factores donde los valores de la diagonal de la matriz de varianza-covarianza o de correlación, no tienen un interés obvio.

En efecto, la redefinición propuesta de la variación total, mediante la eliminación de las matrices *diagonales por bloque*, en la diagonal de la matriz **B**, produciría un sesgo importante debido a la manera como se realiza la aplicación en la Tabla de Burt de las métricas de Ji-cuadrada, ya que la estructura de estas matrices *diagonales por bloque* de la diagonal, representa una gran desviación de los valores esperados, que el *ACM* analiza como si se tratara de una verdadera desviación. Dada esta situación, el uso del *ACM* no es muy adecuado, por lo que (JCA) parece ser una mejor propuesta.

Interpretación de ACM

Al igual que con *ACS*, la interpretación en el *ACM* se basa en las proximidades entre los puntos en el mapa de pocas dimensiones (es decir, dos o tres dimensiones). Así como para *ACS*, las proximidades sólo son significativos entre los puntos del mismo conjunto (es decir, renglones con renglones, columnas con columnas). Específicamente, cuando dos perfiles renglón están cerca uno de otro, implica que tienden a presentar los mismos niveles de las variables nominales.

Para interpretar la proximidad entre los perfiles columna es necesario distinguir dos casos. En primer lugar, la proximidad entre los niveles de diferentes perfiles columna, significa que

estos niveles tienden a aparecer juntos en las observaciones. En segundo lugar, debido a que los niveles de la misma variable nominal no pueden ocurrir al mismo tiempo, necesitamos un tipo diferente de la interpretación para este caso. Aquí la proximidad entre los niveles significa que el grupos de observaciones asociados con estos dos niveles son en sí mismos similares.

ANÁLISIS DISCRIMINANTE

INTRODUCCIÓN

Un problema muy importante en estadística lo constituye el llamado problema de clasificación. En esta sección y en la siguiente discutiremos el problema de clasificación desde dos perspectivas diferentes. Al considerar grupos de objetos en un conjunto de datos multivariados pueden surgir dos situaciones: es algunos casos es de interés determinar si de manera natural las observaciones forman grupos o clases, mientras que en otras ocasiones nos interesa clasificar a los objetos de acuerdo a un conjunto de categorías definidas previamente. En este último caso se trata de un problema de *clasificación supervisada* y será discutido en esta sección.

El problema de discriminación o clasificación es habitual en muchas áreas de la actividad humana, que van desde un diagnóstico médico hasta los sistemas que posibilitan la concesión de un crédito bancario o de reconocimiento de falsas obras de arte (pinturas o escritos).

El problema de *discriminar* aparece en muchas situaciones en que es necesario clasificar elementos con información incompleta. Por ejemplo, los sistemas automáticos de concesión de créditos (*credit scoring*) implementados en muchas instituciones financieras o bancarias, deben utilizar algunas variables de los individuos sujetos al crédito, tales como : nivel de ingresos, historial crediticio, antigüedad en el trabajo, patrimonio, edo. civil, etc., para decidir si el sujeto es o no confiable para otorgarle dicho crédito. En ingeniería este problema se conoce con el nombre de reconocimiento de patrones (pattern recognition), para diseñar máquinas capaces de realizar clasificaciones de manera automática. Por ejemplo, reconocer voces y sonidos, clasificar billetes o menedas, reconocer caracteres escritos en una pantalla de una computadora o clasificar cartas según el distrito postal. Otros ejemplos de aplicaciones del análisis discriminante son: asignar la autoría de un texto escrito de procedencia desconocida a uno de entre varios autores por las frecuencias de uso de palabras; asignar una partitura musical o un cuadro a un artista; determinar una declaración de impuestos como potencialmente fraudulenta o no; determinar una empresa como en riesgo de quiebra o no; un paciente como enfermo de cáncer o no; en Biología se presenta en la llamada taxonomía de especies, que consiste en asignar diversos individuos en *taxones*, etc.

El nombre del análisis discriminante como *técnica de clasificación supervisada*, proviene del

hecho que conocemos una muestra de elementos bien clasificados (nuestra muestra) que sirve de pauta o modelo para la clasificación de futuras observaciones.

Planteamiento estadístico del problema

Desde el punto de vista estadístico, el anális discriminante tiene los siguientes elementos.

• Se dispone de un conjunto de elementos que pueden provenir de dos o más poblaciones distintas.

• En cada elemento se ha observado un vector aleatorio de dimensión p: $\mathbf{X} = (x_1, x_2, ..., x_p)$ de características de los individuos que, suponemos, son potencialmente distintas en las poblaciones, i.e., pueden ayudar a discriminar entre estas poblaciones.

Objetivos del análisis discriminante

• **Disciminación**: Describir las características que diferencian a los distintos grupos conocidos de una población. Para encontrar factores discriminantes cuyos valores numéricos sean tales que separen a los grupos lo más posible.

• **Clasificación**: Asignar nuevos sujetos a un grupo, de entre dos o más. Derivar una regla que pueda usarse para asignar de forma *óptima* un individuo a un grupo de los ya conocidos.

Nota hitórica: La primera aplicación del análisis discriminante consitió en clasificar los restos de un cráneo descubierto en una excavación, como humano, utilizando la distribución de medidas físicas para los cráneos humanos y los de antropoides (Def. dicccionario: Que se parece al ser humano en sus características externas | antropomorfo).

En resumen

¿Qué es el análisis discriminante?

• Es una técnica estadística *de reducción de dimensión*, cuyo objetivo es maximizar la separación entre los datos de $p >> 2$ ó 3 dimensiones, cuando se realice esta reducción de dimensión a 2 ó 3.

¿Para qué?

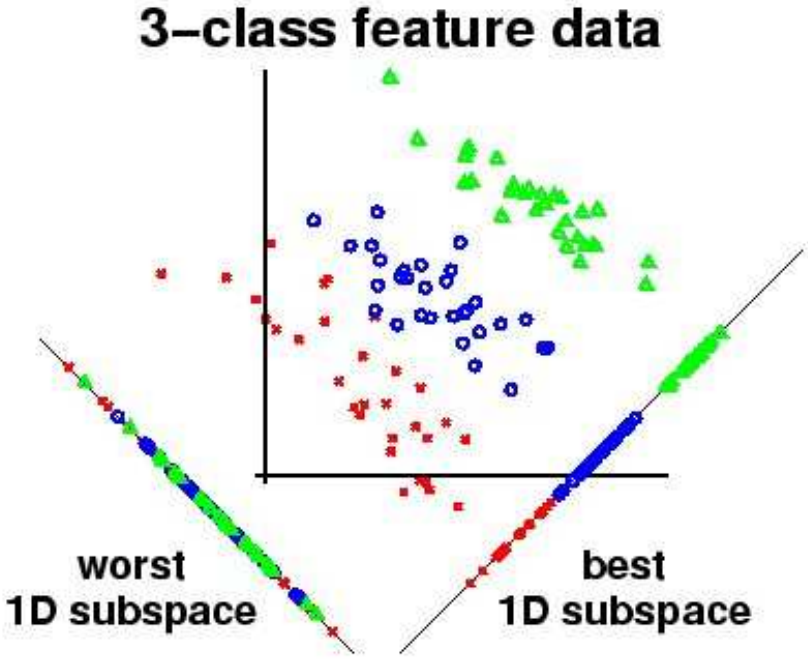
- Identificación de las características de los grupos
- Variables que discriminan entre los grupos
- Selección de las variables discriminantes
- Clasificación de nuevos individuos en los grupos ya existentes

IMPORTANTE: Para implementar esta técnica, *los grupos deben estar definidos de antemano*. Esta agrupación podría ser producto de algunos de los métodos multivariados para este fin, como *cluster o componentes principales*, producto de una agrupación natural o de la experiencia del usuario.

Variables canónicas discriminantes: Dos grupos

Mencionamos que el análisis discriminante es una técnica de reducción de dimension, reducción que sabemos se logra proyectando nuetras observaciones, originalmente en dimensión p, a un espacio de dimensión menor, idealmente *2* ó *3* para poderlas visualizar gráficamente. En este caso, esta proyección debiera ser tal que logre la mayor separación posible de los grupos en el espacio donde se proyectan. El ejemplo de la gráfica siguiente muestra que la elección del plano de proyección (en este caso una recta) no es trivial, por lo que se requieren de elementos técnicos para su determinación.

Podemos observar que la proyección sobre el plano en $\mathbb{R}^1$: la línea recta del lado izquierdo de la gráfica, no posibilita la separación de los tres grupos de observaciones. Por el contrario, una proyección de estos datos sobre el plano en $\mathbb{R}^1$, representado por la línea recta del lado derecho, logra una muy buena separación de los grupos en este espacio reducido. En este caso de dos grupos, el problema se transforma en elegir, de todas las posibles líneas rectas, aquella que maximice la separación de estas proyecciones, que son valores escalares.


3–class feature data
worst
1D subspace
best
1D subspace

Las funciones lineales discriminantes de Fisher

La función lineal discriminante para dos grupos fue deducida por primera vez por Fisher, a través de un razonamiento intuitivo. El criterio propuesto por Fisher es encontrar una variable escalar, que sea tal que maximice la distancia entre los datos proyectados.

$$Y = \mathbf{a}^{'}\mathbf{X}$$

Como tenemos sólo dos poblaciones, entonces necesitamos una única función lineal discriminante

$$Y = \mathbf{a}^{'}\mathbf{X} = a_1 X_1 + ... + a_p X_p$$

Entonces, de manera general, tenemos el siguiente planteamiento.

Dos poblaciones: π_1 y π_2, donde cada uno de los individuos que las componenen tiene un vector de p variables medidas $\mathbf{X}^{'} = (X_1, ..., X_p)$, con $\mathbf{X}_1$ y $\mathbf{X}_2$, las matrices de datos de los sujetos en cada uno de los dos grupos, respectivamente.

Entonces, una vez proyectados los datos originales $\mathbf{X}^{'} = (X_1, ..., X_p)$ a través de las funciones lineales (combinaciones lineales). Tenemos

• Todos los puntos (sujetos) $(\mathbf{X}_1, \mathbf{X}_2)$ son proyectados (mapeados) sobre el plano, Y.

• Hay que elegir a Y de tal manera que logremos la mayor separación entre los grupos proyectados.

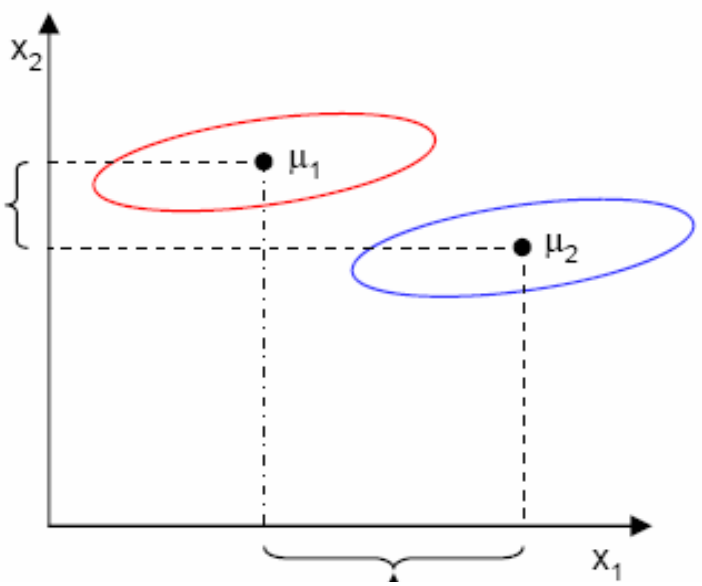
Pero, para encontrar un vector que proporciones una "buena proyección", en el sentido que dijimos, necesitamos definir una medida de separación entre estas proyecciones. Una buena alternativa podría ser elegir la distancia entre las medias proyectadas por estas funciones lineales, como nuestra función objetivo, es decir

$$J(\mathbf{a}) = |\tilde{\mu}_1 - \tilde{\mu}_2| = \left|\mathbf{a}^{'}\mu_1 - \mathbf{a}^{'}\mu_2\right| = \left|\mathbf{a}^{'}(\mu_1 - \mu_2)\right|$$

con

$$\begin{aligned} \mu_1 &= \mathbf{E}\left(\mathbf{X}|\pi_1\right) : \text{media de } \mathbf{X} \text{ en la población 1} \\ \mu_2 &= \mathbf{E}\left(\mathbf{X}|\pi_2\right) : \text{media de } \mathbf{X} \text{ en la segunda población 2} \end{aligned}$$

Sin embargo, la distancia entre las medias proyectadas de cada grupo, no es una muy buena medida, ya que no toma en cuenta la variabilidad (varianza o desviación estándar) dentro de estos grupos. En la gráfica siguiente se muestra que aunque existe mayor separación de las medias proyectando sobre el eje horizontal, se logra una mejor separación de los grupos, proyectando sobre el eje vertical.


x2
μ1
μ2
x1

La solución propuesta por Fisher para salvar esta dificultad, fue maximizar una función que represente esta diferencia de medias, pero normalizada (escalada) por una medida de la variabilidad dentro de los grupos.

Para cada grupo, esta variabilidad es equivalente a la varianza del grupo proyectado

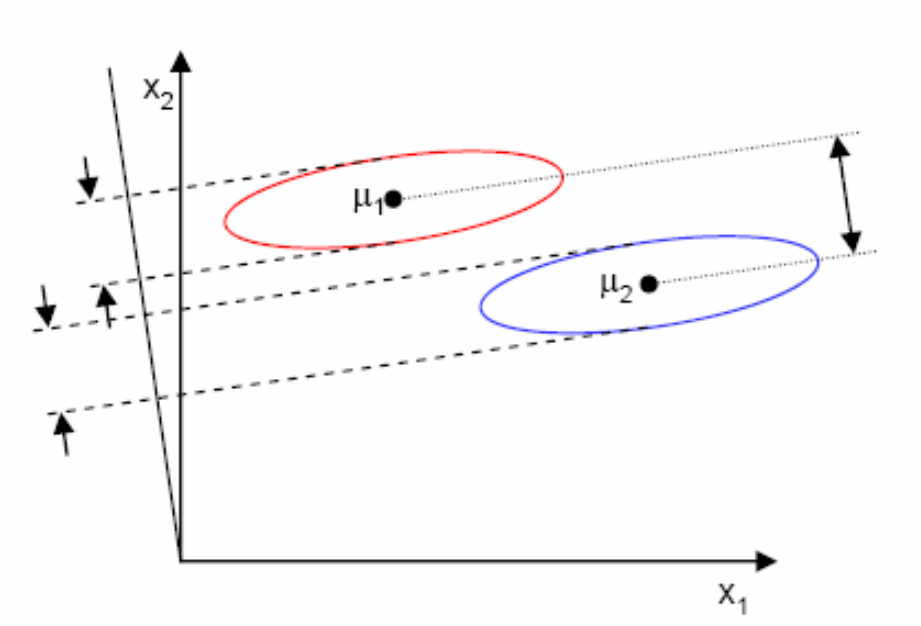
$$\tilde{S}_i^2 = \sum_{Y \in G_i} \left(Y - \tilde{\mu}_i\right)^2, \;\; i = 1, 2$$

Entonces, $\tilde{S}_i$ mide la *variabilidad dentro del grupo* i después de que ha sido proyectado en el plano Y.

Por lo tanto, $\tilde{S}_1^2 + \tilde{S}_2^2$ mide la variabilidad dentro de los dos grupos, una vez realizada la proyección, denominada *variabilidad dentro de grupos* de las muestras proyectadas.

Entonces, la función lineal discriminante de Fisher, se define como la función lineal: $Y = \mathbf{a}'\mathbf{X}$ que maximiza la función objetivo

$$J\left(\mathbf{a}\right) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$


x_2
μ_1
μ_2
x_1

Para encontrar el vector, $\mathbf{a}^*$, que maximice esta expresión, es necesario escribirla de forma explícita como función de $\mathbf{a}$.

Definamos la variabilidad en y dentro de los grupos en el espacio original $\mathbf{X}$, como

$$S_i = \sum_{x \in G_i} (\mathbf{X} - \mu_i)(\mathbf{X} - \mu_i)^{'}, \;\; i = 1, 2 \text{ y}$$

$$S_w = S_1 + S_2$$

Donde S_i es la matriz de varianza-covarianza del grupo i, y S_w la matriz de dispersión dentro de grupos.

Ahora, regresemos a estas mismas definiciones, pero con las observaciones proyectadas en el plano Y. Y tenemos

$$\begin{aligned} \tilde{S}_i^2 = \sum_{Y \in G_i} (Y - \tilde{\mu}_i)^2 &= \sum_{Y \in G_i} \left(\mathbf{a}^{'}\mathbf{X} - \mathbf{a}^{'}\mu_i \right)^2 \\ &= \sum_{x \in G_i} \mathbf{a}^{'} (\mathbf{X} - \mu_i)(\mathbf{X} - \mu_i)^{'} \mathbf{a} \\ &= \mathbf{a}^{'} S_i \mathbf{a} \end{aligned}$$

y

$$\begin{aligned} \tilde{S}_1^2 + \tilde{S}_2^2 &= \mathbf{a}^{'} S_1 \mathbf{a} + \mathbf{a}^{'} S_2 \mathbf{a} \\ &= \mathbf{a}^{'} (S_1 + S_2) \mathbf{a} \\ &= \mathbf{a}^{'} S_w \mathbf{a} \\ &= \tilde{S}_w \end{aligned}$$

Con $\tilde{S}_w$ la matriz de dispersión dentro de grupos proyectados.

De modo similar, las medias proyectadas en el espacio Y, pueden escribirse en términos de las medias en el espacio original, de la siguiente manera

$$
\begin{aligned}
\left(\tilde{\mu}_1 - \tilde{\mu}_2\right)^2 &= \left(\mathbf{a}^{'}\mu_1 - \mathbf{a}^{'}\mu_2\right)^2 \\
&= \mathbf{a}^{'}\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^{'}\mathbf{a} \\
&= \mathbf{a}^{'}S_B\mathbf{a} \\
&= \tilde{S}_B
\end{aligned}
$$

La matriz S_B se conoce como *la matriz de dispersión entre los grupos*, mientras que $\tilde{S}_B$ es la matriz de dispersión entre grupos de las muestras proyectadas.

Ya que $\tilde{S}_B$ es el producto interno entre dos vectores, es de rango a lo más *uno*.

Finalmente, podemos expresar el criterio de Fisher en términos de las dos matrices de dispersion, S_w y S_B, como

$$
J\left(\mathbf{a}\right) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{S}_1^2 + \tilde{S}_2^2} = \frac{\mathbf{a}^{'}S_B\mathbf{a}}{\mathbf{a}^{'}S_w\mathbf{a}}
$$

Una vez planteada la función objetivo, lo que resta es derivarla respecto a $\mathbf{a}$, para encontrar el máximo de ella. Este procedimiento se realiza, por supesto, utilizando técnicas de cálculo vectorial. El desarrollo es el siguiente

Entonces, queremos encontrar el valor de $\mathbf{a}$ que hace máxima la función $J\left(\mathbf{a}\right)$. Diferenciemos esta expresión e igualémosla a cero. Estos es

$$
\begin{aligned}
\frac{d}{d\mathbf{a}}J(\mathbf{a}) &= \frac{d}{d\mathbf{a}}\left(\frac{\mathbf{a}^{'}S_B\mathbf{a}}{\mathbf{a}^{'}S_w\mathbf{a}}\right)=0 \\
&\Rightarrow \mathbf{a}^{'}S_w\mathbf{a}\frac{d}{d\mathbf{a}}\left(\mathbf{a}^{'}S_B\mathbf{a}\right)-\mathbf{a}^{'}S_B\mathbf{a}\frac{d}{d\mathbf{a}}\left(\mathbf{a}^{'}S_w\mathbf{a}\right)=0 \\
&\Rightarrow \left(\mathbf{a}^{'}S_w\mathbf{a}\right)2S_B\mathbf{a}-\left(\mathbf{a}^{'}S_B\mathbf{a}\right)2S_w\mathbf{a}=0, \qquad \cdots(1) \\
&\quad \text{Dividiendo por } 2\mathbf{a}^{'}S_w\mathbf{a} \text{ (que es un escalar)} \\
&\Rightarrow \left(\frac{\mathbf{a}^{'}S_w\mathbf{a}}{\mathbf{a}^{'}S_w\mathbf{a}}\right)S_B\mathbf{a}-\left(\frac{\mathbf{a}^{'}S_B\mathbf{a}}{\mathbf{a}^{'}S_w\mathbf{a}}\right)S_w\mathbf{a}=0 \\
&\Rightarrow S_B\mathbf{a}-J(\mathbf{a})S_w\mathbf{a}=0 \\
&\Rightarrow S_w^{-1}S_B\mathbf{a}-J(\mathbf{a})\mathbf{a}=0 \\
&\Rightarrow S_w^{-1}S_B\mathbf{a}=J(\mathbf{a})\mathbf{a}
\end{aligned}
$$

Obsérvese que $J(\mathbf{a})$ es un escalar, digamos, λ. Entonces tenemos que resolver el problema generalizado de eigenvalores. En concreto, tenemos que encontrar el eigenvalor del sistema

$$S_w^{-1}S_B\mathbf{a}=\lambda\mathbf{a}, \qquad \text{con } \lambda=J(\mathbf{a}) \text{ un escalar}$$

cuya solución es

$$\mathbf{a}^{*}=S_w^{-1}(\mu_1-\mu_2)$$

Una forma alternativa de deducir esta solución es considerar la igualdad (1) de este desarrollo y continuar como sigue

$$\Rightarrow \quad \left(\mathbf{a}^{'} S_w \mathbf{a}\right) 2S_B\mathbf{a} - \left(\mathbf{a}^{'} S_B \mathbf{a}\right) 2S_w\mathbf{a} = 0$$

$$\Rightarrow \quad \left(\mathbf{a}^{'} S_w \mathbf{a}\right) S_B\mathbf{a} = \left(\mathbf{a}^{'} S_B \mathbf{a}\right) S_w\mathbf{a}$$

$$\Rightarrow \quad \left(\mathbf{a}^{'} S_w \mathbf{a}\right) (\mu_1 - \mu_2)(\mu_1 - \mu_2)^{'} \mathbf{a} = \mathbf{a}^{'} (\mu_1 - \mu_2)(\mu_1 - \mu_2)^{'} \mathbf{a} S_w\mathbf{a}$$

$$\Rightarrow \quad (\mu_1 - \mu_2)\left(\mathbf{a}^{'} S_w \mathbf{a}\right) = \mathbf{a}^{'} (\mu_1 - \mu_2) S_w\mathbf{a}$$

$$\Rightarrow \quad (\mu_1 - \mu_2) = \frac{S_w\mathbf{a}\left(\mathbf{a}^{'} (\mu_1 - \mu_2)\right)}{\mathbf{a}^{'} S_w\mathbf{a}}$$

$$\Rightarrow \quad \mathbf{a} = \frac{(\mu_1 - \mu_2)\,\mathbf{a}^{'} S_w\mathbf{a}}{S_w\left(\mathbf{a}^{'} (\mu_1 - \mu_2)\right)}$$

$$\Rightarrow \quad \mathbf{a} = \lambda S_w^{-1} (\mu_1 - \mu_2)$$

con $\lambda = \dfrac{\mathbf{a}^{'} S_w\mathbf{a}}{\mathbf{a}^{'} (\mu_1 - \mu_2)}$ un escalar. Ahora bien, como la función a maximizar es invariante ante multiplicaciones por constantes, y λ lo es, entonces, podemos normalizar $\mathbf{a}$, de tal manera que $\lambda = 1$, de donde obtenemos

$$\mathbf{a}^* = S_w^{-1} (\mu_1 - \mu_2)$$

Función lineal discriminante estimada

Para utilizar el discriminante con datos muestrales, es necesario estimar esta función lineal discriminante a través de los datos observados, recordando que estos datos son los que se generan una vez proyectados a través de la función discriminante. Entonces, las matrices que necesitamos son

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 = \sum_{i=1}^{n}\left(\mathbf{a}'\left(y_i-\overline{y}\right)\right)^2=\mathbf{a}'\mathbf{T}\mathbf{a}: \text{ Suma de cuadrados totales o varianza total, y}$$

$$\sum_{g=1}^{G}n_g\left(\overline{X}_g-\overline{X}\right)^2 = \sum_{g=1}^{G}n_g\left(\mathbf{a}'\left(\overline{y}_g-\overline{y}\right)\right)^2=\mathbf{a}'\mathbf{E}\mathbf{a}: \text{ Suma de cuadrados entre grupos o varianza entre g}$$

$$\widehat{\mu}_1 = \overline{\mathbf{X}}_1,\ \widehat{\mu}_2=\overline{\mathbf{X}}_2 \text{ Las medias en los grupos 1 y 2, respectivamente, y}$$

$$\mathbf{S}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\overline{\mathbf{X}}_1\right)\left(X_{1j}-\overline{\mathbf{X}}_1\right)',\ \ \mathbf{S}_2=\frac{1}{n_2}\sum_{j=1}^{n_2}\left(X_{2j}-\overline{\mathbf{X}}_2\right)\left(X_{2j}-\overline{\mathbf{X}}_2\right), \text{ las respectivas varianza}$$

$$\mathbf{S} = \mathbf{S}_{pool}=\frac{\left(n_1-1\right)\mathbf{S}_1+\left(n_2-1\right)\mathbf{S}_2}{n_1+n_2-2}: \text{ La varianza conjunta de los grupos}$$

El supuesto de matrices de varianza-covarianza iguales dentro de los dos grupos, es fundamental y hace que la matriz de varianza-covarianza total se estime como un *pool* de las correspondientes matrices de cada grupo.

Entonces, la función lineal estimada queda como

$$\widehat{y}=\widehat{\mathbf{a}}'\mathbf{X}=\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)'\mathbf{S}_{pool}^{-1}\mathbf{X}$$

Urgente: Un ejemplo

49 adultos mayores del sexo masculino participaron en un estudio interdisciplinario sobre su condición humana, y fueron clasificados en dos grupos: "factor senil presente" y "factor senil ausente", basados en una intesiva evaluación psicológica. Los siguientes resultados son el resultado de cuatro pruebas realizadas a estos sujetos.

	No Senil (n=37)		*Senil* ($n = 12$)	
Pr*ueba*	$\overline{X}$	*S.D.*	$\overline{X}$	*S.D.*
Información	12.566	3.387	8.750	3.251
Similaridades	9.486	3.380	5.333	4.271
Aritmética	11.514	3.363	8.500	3.631
Pintura	7.973	1.922	4.750	3.571

$$\mathbf{S}_{Senil} = \begin{pmatrix} 11.47 & 8.55 & 6.39 & 2.07 \\ 8.55 & 11.42 & 5.49 & 0.29 \\ 6.39 & 5.49 & 11.31 & 1.82 \\ 2.07 & 0.29 & 1.82 & 3.69 \end{pmatrix}, \quad \mathbf{S}_{No\text{Senil}} = \begin{pmatrix} 10.57 & 10.45 & 9.68 & 7.66 \\ 10.45 & 18.24 & 12.09 & 8.91 \\ 9.68 & 12.09 & 13.18 & 5.32 \\ 7.66 & 8.91 & 5.32 & 12.75 \end{pmatrix},$$

$$\mathbf{S}_{poool} = \begin{pmatrix} 11.26 & 9.00 & 7.16 & 3.38 \\ 9.00 & 13.02 & 7.04 & 2.31 \\ 7.16 & 7.04 & 11.75 & 2.64 \\ 3.38 & 2.31 & 2.64 & 5.81 \end{pmatrix}$$

$$\begin{aligned} \left(\overline{X}_{No\text{Senil}} - \overline{X}_{\text{Senil}}\right)' &= (3.82, 4.15, 3.01, 3.23) \\ \mathbf{a}' &= \left(\overline{X}_{No\text{Senil}} - \overline{X}_{\text{Senil}}\right)' \mathbf{S}_{poool}^{-1} \\ &= (3.82, 4.15, 3.01, 3.23) \begin{pmatrix} 0.249 & -0.127 & -0.060 & 0.066 \\ -0.127 & 0.180 & -0.034 & 0.0182 \\ -0.060 & -0.034 & 0.146 & -0.017 \\ 0.066 & 0.0182 & -0.017 & 0.211 \end{pmatrix} \\ &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) \end{aligned}$$

Las medias de los grupos proyectados son

$$\overline{y}_{No\text{Senil}} = (0.02453159, 0.2162928, 0.01043125, 0.4510016) \begin{pmatrix} 12.566 \\ 9.486 \\ 11.514 \\ 7.973 \end{pmatrix} = 6.07$$

$$\overline{y}_{\text{Senil}} = (0.02453159, 0.2162928, 0.01043125, 0.4510016) \begin{pmatrix} 8.750 \\ 5.333 \\ 8.500 \\ 4.750 \end{pmatrix} = 3.59$$

La función lineal disciminante para cada sujeto es

$$\begin{aligned} y_j = \mathbf{a}^{'} X_j &= (0.02453159, 0.2162928, 0.01043125, 0.4510016)\left(X_{1j}, X_{2j}, X_{3j}, X_{4j}\right)^{'} \\ &= 0.02X_{1j} + 0.22X_{2j} + 0.01X_{3j} + 0.45X_{4j} \end{aligned}$$

Clasificación

Un manera muy simple de utilizar esta función lineal, Y, para clasificar una nueva observación, X_0, a alguno de los grupos es

1.- Calcular la proyección en el plano Y, de esta observación

$$y_0 = \left(\mu_2 - \mu_1\right)^{'} S_w^{-1} X_0$$

2.- Encontrar el punto medio de las medias de los grupos proyectadas $\tilde{\mu}_1$ y $\tilde{\mu}_2$.

$$\begin{aligned} m &= \frac{1}{2}\left(\tilde{\mu}_1 + \tilde{\mu}_2\right) \\ &= \frac{1}{2}\left(\mathbf{a}^{'}\mu_1 + \mathbf{a}^{'}\mu_2\right) \\ &= \frac{1}{2}\left(\mu_2 - \mu_1\right)^{'} S_w^{-1}\left(\mu_2 + \mu_1\right) \end{aligned}$$

3.- Regla de clasificación

$$\begin{aligned} \text{Asignar } \mathrm{X}_0 \text{ al grupo 1 } (\pi_1) \text{ si } y_0 &= \left(\mu_2 - \mu_1\right)^{'} S_w^{-1} X_0 \geq m, \text{ y} \\ \text{Asignar } \mathrm{x}_0 \text{ al grupo 2 } (\pi_2) \text{ si } y_0 &= \left(\mu_2 - \mu_1\right)^{'} S_w^{-1} X_0 < m \end{aligned}$$

o bien si

$$y_0 - m = \left(\mu_2 - \mu_1\right)^{'} S_w^{-1} X_0 \geq 0 \text{ ó } < 0$$

Estimación de la regla de clasificación

La regla de clasificación estimada queda como

$$\begin{aligned} \text{Asignar } \mathrm{X}_0 \text{ al grupo 1 } (\pi_1) \text{ si } y_0 &= \left(\bar{X}_2 - \bar{X}_1\right)' S_{pool}^{-1} X_0 \geq m, \text{ y} \\ \text{Asignar } \mathrm{x}_0 \text{ al grupo 2 } (\pi_2) \text{ si } y_0 &= \left(\bar{X}_2 - \bar{X}_1\right)' S_{pool}^{-1} X_0 < m \end{aligned}$$

En nuestro caso

$$m = \frac{1}{2}(6.07 + 3.59) = 4.83$$

Entonces, un nuevo individuo se asignaría al grupo: *No senil* si y_0, su puntaje dado por la proyección de sus valores en el plano Y, es mayor que 4.83, y se asignaría al grupo *Senil* si es menor a 4.83.

Por ejemplo, a qué grupo asignaríamos a un individuo que tiene el siguiente vector de observaciones: $X_0 = (8.150, 6.001, 9.050, 4.510)$?

Notemos primeramente que este vector está cercano a las medias del grupo *senil*: (8.750, 5.333, 8.500, 4.750), entonces, debería de clasificarse en ese grupo. Calculemos su proyección al plano Y, i.e., calculemos

$$\begin{aligned} y_0 = \mathbf{a}' X_0 &= (0.02453159, 0.2162928, 0.01043125, 0.4510016)\left(X_{10}, X_{20}, X_{30}, X_{40}\right)' \\ &= 0.02X_{10} + 0.22X_{20} + 0.01X_{30} + 0.45X_{40} \\ &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) * (8.150, 6.001, 9.050, 4.510) = 3.626326 \end{aligned}$$

Si consideramos ahora un sujeto con valores más cercanos a las medias del grupo *No senil*: (12.566 , 9.486, 11.514, 7.973), digamos, $X_0 = (11.950, 10.00, 10.73, 8.103)$, debería clasificarse como No senil. Su proyección es: 6.222474; que corrobora nuestra especulación.

Discriminante clásico

Se denomina discriminante clásico al discriminante que asume poblaciones normales multivariadas para cada uno de los grupos. Es decir, se supone que cada población tiene función de densidad de probabilidad, dada por

$$f_i(\mathbf{X}) = \frac{(2\pi)^{-p/2}}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)\right\}, \quad \text{i=1,2,...,G}$$

Discriminante clásico con dos grupos

En el caso de que tengamos dos grupos con probabilidades a priori de pertencia a cada uno de ellos π_1 y π_2, respectivamente ($\pi_1 + \pi_2 = 1$). Entonces, para clasificar a un nuevo individuo, $\mathbf{x}_0$, por ejemplo, al grupo *2*, sólo debemos comparar sus densidades, y lo asignamos al grupo *2* síi

$$\pi_2 f_2(\mathbf{x}_0) > \pi_1 f_1(\mathbf{x}_0)$$

si las probabilidades iniciales son iguales, entonces lo asignamos a dicho grupo sii

$$f_2(\mathbf{x}_0) > f_1(\mathbf{x}_0)$$

Bajo el supuesto de que las densidades sean normales de dimensión *p,* tenemos

$$\begin{aligned}
\frac{\pi_2}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_2)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2)\right\} &> \frac{\pi_1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_1)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1)\right\} \Longrightarrow \\
\log(\pi_2) - \frac{1}{2}(\mathbf{x}-\mu_2)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2) &> \log(\pi_1) - \frac{1}{2}(\mathbf{x}-\mu_1)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1) \\
(\mathbf{x}-\mu_1)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1) &> (\mathbf{x}-\mu_2)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2) - 2\log\left(\frac{\pi_2}{\pi 1}\right)
\end{aligned}$$

si denotamos como D_i^2 el cuadrado de la distancia de Mahalanobis entre el punto observado, $\mathbf{x}$, y la media de la población i=1,2, tenemos

$$D_i^2 = (\mathbf{x}-\mu_i)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)$$

si suponemos probabilidades iniciales iguales, entonces la regla que se obtiene para clasificar $\mathbf{x}$ en el grupo *2* es: Clasificar esta observación en el grupo *2* sii

$$D_1^2 > D_2^2$$

es decir, clasificar la observación en el grupo cuyas medias estén más próximas, según la distancia de Mahalanobis cuadrada.

Interpretación de la regla anterior

Desarrollemos las siguientes expresiones

$$\begin{aligned}
(\mathbf{x}-\mu_1)^{'}\,\mathbf{\Sigma}^{-1}\,(\mathbf{x}-\mu_1) &= \mathbf{x}^{'}\mathbf{\Sigma}^{-1}\mathbf{x}-2\mu_1^{'}\mathbf{\Sigma}^{-1}x+\mu_1^{'}\mathbf{\Sigma}^{-1}\mu_1,\ \text{y} \\
(\mathbf{x}-\mu_2)^{'}\,\mathbf{\Sigma}^{-1}\,(\mathbf{x}-\mu_2) &= \mathbf{x}^{'}\mathbf{\Sigma}^{-1}\mathbf{x}-2\mu_2^{'}\mathbf{\Sigma}^{-1}x+\mu_2^{'}\mathbf{\Sigma}^{-1}\mu_2
\end{aligned}$$

entonces, la regla divide al conjunto de valores posibles de $\mathbf{x}$, en dos regiones cuya frontera es

$$-2\mu_1^{'}\mathbf{\Sigma}^{-1}x+\mu_1^{'}\mathbf{\Sigma}^{-1}\mu_1=-2\mu_2^{'}\mathbf{\Sigma}^{-1}x+\mu_2^{'}\mathbf{\Sigma}^{-1}\mu_2$$

que es equivalente, como función de $\mathbf{x}$ a

$$(\mu_2-\mu_1)^{'}\,\mathbf{\Sigma}^{-1}\mathbf{x}=(\mu_2-\mu_1)^{'}\,\mathbf{\Sigma}^{-1}\left(\frac{\mu_2+\mu_1}{2}\right)$$

Observemos que el hecho de suponer matriz de varianzas covarianzas iguales entre los grupos, permite el agrupamiento de los términos de esta manera. Si denotamos por

$$\mathbf{a}^{'}=(\mu_2-\mu_1)^{'}\,\mathbf{\Sigma}^{-1}$$

entonces, la frontera entre las dos regiones de clasificación para π_1 y π_2 puede escribirse como

$$\mathbf{a}^{'}\mathbf{x}=\mathbf{a}^{'}\left(\frac{\mu_2+\mu_1}{2}\right)$$

que es la ecuación de un hiperplano. También equivalente a

$$2\mathbf{a}^{'}\mathbf{x} = \mathbf{a}^{'}(\mu_1 + \mu_2)$$

$$\mathbf{a}^{'}\mathbf{x} - \mathbf{a}^{'}\mu_1 = \mathbf{a}^{'}\mu_2 - \mathbf{a}^{'}\mathbf{x}\,(*)$$

Se puede demostrar que esta regla equivale a proyectar el punto $\mathbf{x}$ que queremos clasificar y las medias de ambas poblaciones sobre la función lineal discriminante, y después asignar el punto a aquella población de cuya media se encuentre más próxima en la proyección. Situación que habíamos visto anteriormente.

Esta última ecuación indica que el procedimiento para clasificar un elemento X_0 puede resumirse como sigue:

• Calcular el vector $\mathbf{a}^{'}$, mediante la expresión correspondiente

• Construir la función lineal discriminante

$$Y = \mathbf{a}^{'}\mathbf{X} = a_1X_1 + ... + a_pX_p$$

• Calcular la proyección en el plano Y, $Y_0 = \mathbf{a}^{'}X_0$, del individuo $X_0 = (X_{10}, ..., X_{p0})$, y el valor de las medias proyectadas de las poblaciones, $\tilde{\mu}_i = \mathbf{a}^{'}\mu_i$. Clasificar esta observación en aquella población donde la distancia, $|Y_0 - \tilde{\mu}_i|$, sea mínima.

Obsérvese que

$$\mathbb{E}(Y|\pi_i) = \tilde{\mu}_i = \mathbf{a}^{'}\mu_i,\ \ i = 1, 2$$

Entonces, la regla de decisión que se desprende de $(*)$, equivale a clasificar la observación en el grupo π_2, síi

$$|Y - \tilde{\mu}_1| > |Y - \tilde{\mu}_2|$$

Esta variable aleatoria Y tiene varianza dada por

$$\mathbb{V}(Y) = \mathbb{V}\left(\mathbf{a}'\mathbf{X}\right) = \mathbf{a}'\mathbb{V}(\mathbf{X})\,\mathbf{a} = \mathbf{a}'\Sigma\mathbf{a} = (\mu_2 - \mu_1)' \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\mu_2 - \mu_1)$$
$$= (\mu_2 - \mu_1)' \boldsymbol{\Sigma}^{-1}(\mu_2 - \mu_1) = D^2$$

y el cuadrado de la distancia que es un escalar, entre las medias proyectadas es la distancia de Mahalanobis entre los vectores de medias originales:

$$(\tilde{\mu}_2 - \tilde{\mu}_1)^2 = \left(\mathbf{a}'(\mu_2 - \mu_1)\right)^2 = (\mu_2 - \mu_1)' \boldsymbol{\Sigma}^{-1}(\mu_2 - \mu_1) = D^2$$

Cálculo de probabilidades de clasificación errónea

Una vez que hemos obtenido la regla de clasificación, en este caso, para dos poblaciones, debemos estimar o calcular las probabilidades de clasificación errónea o estimar el error de clasificación. Bajo el supuesto de que las poblaciones son normales multivariadas con igual matriz de varianza-covarianza, obtuvimos que

$$Y = \mathbf{a}'\mathbf{X} \sim N\left(\mathbf{a}'\mu_i, D^2\right)$$

Bajo esta situación, podemos calcular la probabilidad de clasificar de manera errónea una observación. En concreto, la probabilidad de clasificar erróneamente una observación $\mathbf{X}$ cuando $\mathbf{X} \in \pi_1$, es

$$\mathbb{P}(\pi_2|\pi_1) = \mathbb{P}\left\{y \geq \frac{\tilde{\mu}_1 + \tilde{\mu}_2}{2} | y \sim N\left(\tilde{\mu}_1, D^2\right)\right\}$$

Si construimos la variable estandarizada $z = \dfrac{y - \tilde{\mu}_1}{D} \sim N(0,1)$, entonces esta probabilidad es

$$\mathbb{P}(\pi_2|\pi_1) = \mathbb{P}\left\{z \geq \frac{\dfrac{\tilde{\mu}_1 + \tilde{\mu}_2}{2} - \tilde{\mu}_1}{D}\right\} = 1 - \Phi\left(\frac{D}{2}\right)$$

De forma análoga, la probabilidad de clasificar de manera errónea una observación $\mathbf{X}$ cuando $\mathbf{X} \in \pi_2$, es

$$\mathbb{P}\left(\pi_1|\pi_2\right) = \mathbb{P}\left\{ z < \frac{\frac{\tilde{\mu}_1 + \tilde{\mu}_2}{2} - \tilde{\mu}_2}{D} \right\} = \Phi\left(-\frac{D}{2}\right)$$

Por la simetría de la distribución normal, estas probabilidades son iguales, además de que la regla de clasificación obtenida hace mínimas estas probabilidades de error, y los errores de clasificación sólo dependen de las distancias de Mahalanobis entre las medias.

Probabilidades a posteriori

El grado de certeza de la regla de clasificación, depende de la probabilidad de acertar (clasificar correctamente) mediante la misma. La *probabilidad a posteriori* de que la observación, $\mathbf{X}$, sea clasificada o asignada a la población, π_1, se calcula como

$$\begin{aligned}
\mathbb{P}\left(\pi_1|\mathbf{X}\right) &= \frac{\pi_1 f_1\left(\mathbf{X}\right)}{\pi_1 f_1\left(\mathbf{X}\right) + \pi_2 f_2\left(\mathbf{X}\right)} \\
&= \frac{\pi_1 exp\left\{-\frac{1}{2}\left(\mathbf{X}-\mu_1\right)^{'}\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu_1\right)\right\}}{\pi_1 exp\left\{-\frac{1}{2}\left(\mathbf{X}-\mu_1\right)^{'}\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu_1\right)\right\} + \pi_2 exp\left\{-\frac{1}{2}\left(\mathbf{X}-\mu_2\right)^{'}\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu_2\right)\right\}}
\end{aligned}$$

Que puede escribirse en términos de las distancias cuadradas de Mahalanobis entre la observación y cada una de las dos medias, D_1^2 y D_2^2, como:

$$\mathbb{P}\left(\pi_1|\mathbf{X}\right) = \frac{1}{1 + \frac{\pi_1}{\pi_2} exp\left\{-\frac{1}{2}\left(D_1^2 - D_2^2\right)\right\}}$$

Discriminante logístico

En el problema de clasificación a través del análisis discriminante que hemos tratado en esta sección, vimos que si la distribución conjunta de las observaciones es normal multivariada, utilizar las distancias de Mahalanobis estimadas suele dar buenos resultados y resulta óptimo con muestras grandes. Sin embargo, frecuentemente los datos recabados para realizar esta clasificación no son normales. Por ejemplo, en muchos problemas de clasificación se utilizan variables discretas, lo que haría cuestionable la distribución normal multivariada de los datos, y la condición óptima de los resultados basados en ésta.

El modelo Logit

Consideremos el problema de la discriminación únicamente entre dos poblaciones. Una forma de abordar el problema es definir una variable de clasificación, y, que tome el *valor cero* cuando la observación pertenezca a la primera población, π_1, *y uno* cuando pertenece a la segunda, π_2. Entonces, la muestra consistirá en n elementos del tipo $(y_i, \mathbf{X}_i)$, donde y_i determina el valor de la variable de clasificación, y, y $\mathbf{X}_i$ es un vector de variables explicativas o predictoras. A continuación, construiremos un modelo para pronosticar el valor de la variable de respuesta o de clasificación, y, de una nueva observación, cuando se conocen las variables predictoras, $\mathbf{X}$. El primer enfoque simple es formular el modelo de regresión lineal correspondentiene. Basta analizar este enfoque a través del modelo de regresión simple, aunque el vector de variables explicativas puede ser de dimensión mayor a uno. El modelo es

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n$$

Entonces, sabemos que

$$\mathbb{E}\left(y_i | x_i\right) = \beta_0 + \beta_1 x_i$$

Ya que nuestra variable de respuesta, y, es una indicadora de pertenencia a los grupos, es claro que esta esperanza es la probabilidad de que un sujeto sea clasificado en la población, π_2. Llamenos p_i a la probabilidad de que esta variable tome el *valor uno*, i.e., que la observación pertenezca al grupo, *dos*. Entonces

$$p_i = \mathbb{P}(y = 1|x_i)$$

Entonces, la variable de clasificación, y, es binomial y toma los valores posibles *uno* y *cero* con probabilidades p_i y $1 - p_i$. Por lo que su esperanza es:

$$\mathbb{E}(y|x_i) = p_i \times 1 + (1 - p_i) \times 0 = p_i$$

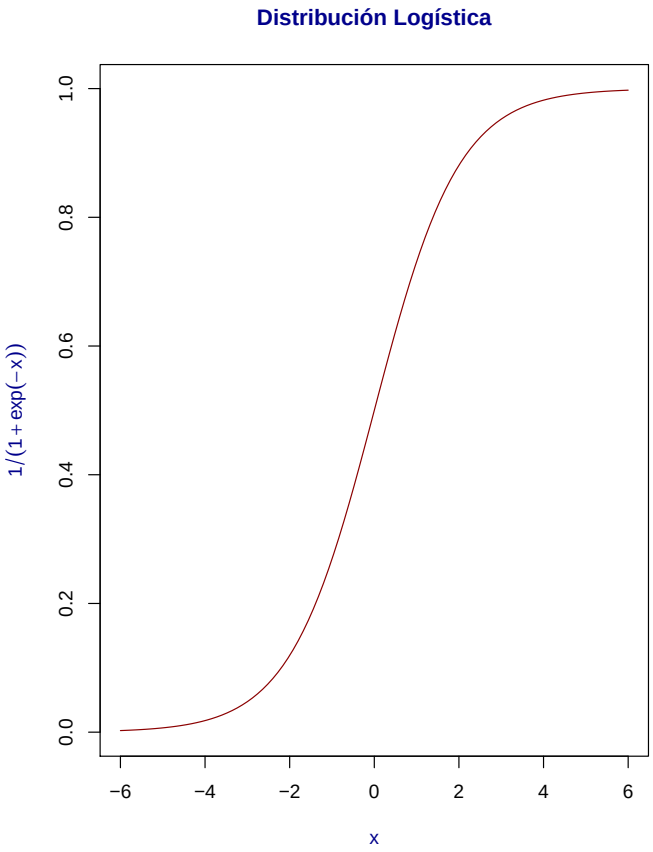
por lo tanto, concluimos que

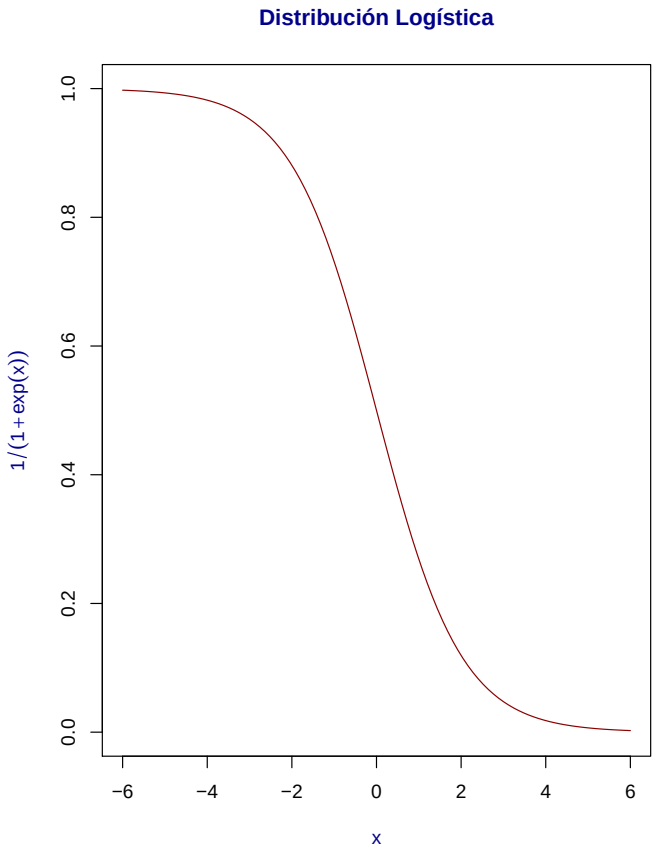
$$p_i = \beta_0 + \beta_1 x_i$$

Si, aparentemente, este modelo funciona bien, entonces...

Porqué no usar el modelo lineal?

Reflexionemos un poco sobre el modelo propuesto. El rango de variación de p_i está entre *cero y uno*, porque es una probabilidad; sin embargo, el rango de variación de $\beta_0 + \beta_1 x_i$ no necesariamente está dentro de este rango, de hecho, **¡podría ser negativo!**, lo que es, por supuesto, una incongruencia. Un inconveniente más es que nuestra variable de respuesta se distribuye $Bernoulli(p_i)$ y no tiene varianza constante, ya que su varianza es $p_i(1 - p_i)$. Por lo tanto, esta propuesta de modelo no parece ser adecuada para el tipo de respuesta que queremos modelar. Necesitamos un modelo cuyos valores para la respuesta estén contenidos en el intervalo (0,1) y que si $\beta_1 > 0$ y x es "grande" entonces $p_i \to 1$ o si $x \to -\infty$ $p_i \to 0$. Por otro lado, si $\beta_1 < 0$ y $x \to \infty$, entonces $p_i \to 0$, o si $x \to -\infty$, entonces $p_i \to 1$ de hecho, necesitamos una función cuya gráfica sea de la forma





La función que permite ajustar este tipo de curvas es *la función logística* y el modelo de regresión asociado es el *modelo de regresión logística*. Este modelo se escribe de la siguiente manera

$$\pi(x) = Pr(Y = 1|x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$$

este es un modelo *no lineal* para nuestra respuesta, $\pi(x)$. Para lograr un modelo lineal a partir del modelo anterior, primero, construyamos el momio correspondiente

$$\frac{\pi(x)}{1 - \pi(x)} = \frac{Pr(Y = 1|x)}{1 - Pr(Y = 1|x)} = exp\left(\beta_0 + \beta_1 x\right)$$

si aplicamos la función logaritmo en ambos lados de la igualdad tenemos

$$logit(\pi(x)) = log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x$$

que es un modelo lineal, para el logaritmo del momio de la respuesta, conocido como *Logit*. Obviamente, este modelo se puede generalizar para más de una variable explicativa. El modelo con *p regresores* es

$$logit(\pi(x)) = log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

la escala de medición de los regresores puede ser cualquiera.

Los parámetros de este modelo se estiman por *máxima verosimilitud*. El modelo predice la probabilidad de cada individuo de presentar la respuesta *y=1*, por lo que, generalmente, se asigna a una observación a la población asociada a este valor de respuesta, si esta probabilidad es *mayor que 0.5* y se asigna a la otra población, en caso contrario.

Funciones lineales discriminantes para varios grupos

El enfoque de Fisher puede generalizarse para encontrar las funciones lineales que tengan máximo poder discriminante para clasificar nuevos elementos entre $G > 2$ poblaciones. La manera de hacerlo es semejante al caso de dos grupos, sólo que ahora se tienen *k=min(G-1,p)* funciones discriminantes. Es decir

$$\begin{aligned} Y_1 &= \mathbf{a}_1^{'} X \\ Y_2 &= \mathbf{a}_2^{'} X \\ &\vdots \\ Y_k &= \mathbf{a}_k^{'} X \end{aligned}$$

Entonces, en este caso, el proceso de clasificación es como sigue:

• Proyectamos las medias de cada grupo. Esto es, obtenemos

$$\tilde{\mu}_i = \mathbf{E}\left(Y|\pi_i\right) = \mathbf{E}\left(\mathbf{a}^{'}\mathbf{X}|\pi_i\right) = \mathbf{a}^{'}\mathbf{E}\left(\mathbf{X}|\pi_i\right) = \mathbf{a}^{'}\mu_i \quad \text{i=1,2,...,G}$$

• Proyectamos el vector de covariables del sujeto a clasificar, X_0, y obtenemos y_0 su proyección sobre el espacio *Y*.

• Clasificamos el punto en aquella población de cuya media se encuentre más cercana.

Las distancias se miden con la distancia euclideana en el espacio de las variables canónicas, *y*. Es decir, clasificaremos al sujeto , X_0, en la población i si:

$$\left(y_0 - \tilde{\mu}_i\right)^{'}\left(y_0 - \tilde{\mu}_i\right) = \min_g \left(y_0 - \tilde{\mu}_g\right)^{'}\left(y_0 - \tilde{\mu}_g\right)$$

Como tenemos varios grupos, la separación entre las medias la mediremos por el cociente entre la variabilidad entre grupos, y la variabilidad dentro de los grupos. Este es el criterio habitual para comparar varias medias en el análisis de la varianza y genera el estadístico *F de Fisher*. De hecho, lo que estamos haciendo es plantear un análisis de varianza en el

espacio de proyección, Y.

Nuevamente, para obtener las variables lineales discriminantes, comenzamos buscando un vector de proyección, $\mathbf{a}$, de norma uno, tal que los grupos de observaciones proyectados sobre él tengan separación relativa máxima. La proyección de la media de las observaciones del grupo g en esta dirección corresponde al escalar:

$$\bar{\tilde{\mu}}_g = \mathbf{a}'\bar{\mathbf{X}}_g$$

Con la correspondiente proyección para la media de todos los datos, dada por

$$\bar{\tilde{\mu}} = \mathbf{a}'\bar{\mathbf{X}}$$

ambas medias proyectadas son vectores de dimensión $p \times 1$, sólo que la primera es para los individuos en el grupo *g, g=1,2,...,k*, y la segunda es para todos los datos, sin importar la pertenencia a algún grupo.

Entonces, tomando como medida de la distancia entre las medias de los grupos proyectadas: $\tilde{\mu}_1, \tilde{\mu}_2, ..., \tilde{\mu}_k$, la varianza total dentro de grupos es

$$\sum_{g=1}^{k} n_g \left(\bar{\tilde{\mu}}_g - \bar{\tilde{\mu}}\right)^2$$

que debemos comparar contra la varianza dentro de grupos o variabilidad total, dada por

$$\sum_i \sum_g \left(y_{ig} - \bar{\tilde{\mu}}_g\right)^2$$

El proceso para encontrar las funciones lineales se realiza mediante el cociente de las varianzas entre grupos y total (idéntico al procedimiento ANOVA, sólo que aquí estas varianzas se obtienen con los elementos proyectados).

$$\frac{\sum_{g=1}^{k} n_g \left(\bar{\tilde{\mu}}_g - \bar{\tilde{\mu}}\right)^2}{\sum_i \sum_g \left(y_{ig} - \bar{\tilde{\mu}}_g\right)^2}$$

Ahora, expresemos este criterio en función de los datos originales. La suma de cuadrados *dentro de grupos*, para los puntos proyectados, es:

$$\sum_{i=1}^{n_g}\sum_{g=1}^{k}\left(y_{ig}-\bar{\tilde{\mu}}_g\right)^2=\sum_{i=1}^{n_g}\sum_{g=1}^{k}\mathbf{a}^{'}\left(\mathbf{X}_{ig}-\bar{X}_g\right)\left(\mathbf{X}_{ig}-\bar{X}_g\right)^{'}\mathbf{a}=\mathbf{a}^{'}\mathbf{W}\mathbf{a}$$

con $\mathbf{W}$, dada por

$$\sum_{i=1}^{n_g}\sum_{g=1}^{k}\left(\mathbf{X}_{ig}-\bar{X}_g\right)\left(\mathbf{X}_{ig}-\bar{X}_g\right)^{'}$$

Esta matriz tiene dimensiones $p\times p$ y, en general, es de rango p, asumiendo que $n-k\geq p$. Estima la variabilidad de los datos respecto a las medias de su grupo.

Por otro lado, la suma de cuadrados *entre grupos*, para los puntos proyectados está dada por

$$\sum_{g=1}^{k}n_g\left(\bar{\tilde{\mu}}_g-\bar{\tilde{\mu}}\right)^2=\sum_{g=1}^{k}n_g\mathbf{a}^{'}\left(\bar{X}_g-\bar{X}\right)\left(\bar{X}_g-\bar{X}\right)^{'}\mathbf{a}=\mathbf{a}^{'}\mathbf{B}\mathbf{a}$$

Es decir, la matriz $\mathbf{W}$ corresponde a las diferencias dentro de grupos (withing) y la matriz $\mathbf{B}$ las diferencias entre grupos (between).

Entonces, la cantidad a maximizar para encontrar las funciones lineales discriminantes es

$$\mathbf{J}=\frac{\mathbf{a}^{'}\mathbf{B}\mathbf{a}}{\mathbf{a}^{'}\mathbf{W}\mathbf{a}}$$

Realizando el proceso usual, tenemos

$$\begin{aligned}\frac{2\mathbf{B}\mathbf{a}\left(\mathbf{a}^{'}\mathbf{W}\mathbf{a}\right)-\left(\mathbf{a}^{'}\mathbf{B}\mathbf{a}\right)\mathbf{W}\mathbf{a}}{\left(\mathbf{a}^{'}\mathbf{W}\mathbf{a}\right)^2} &= 0\\ \mathbf{B}\mathbf{a} &= \mathbf{W}\mathbf{a}\frac{\left(\mathbf{a}^{'}\mathbf{B}\mathbf{a}\right)}{\left(\mathbf{a}^{'}\mathbf{W}\mathbf{a}\right)}\\ \mathbf{B}\mathbf{a} &= \mathbf{J}\mathbf{W}\mathbf{a}\end{aligned}$$

Suponiendo que $\mathbf{W}$ tiene inversa, i.e., es no singular, y observando que $\mathbf{J}$ es un escalar, que denotaremos como λ, entonces, obtenemos el sistema

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$$

lo que implica que $\mathbf{a}$ debe ser un vector propio de la matriz $\mathbf{W}^{-1}\mathbf{B}$ y λ su valor propio asociado. Como el objetivo es maximizar $\lambda = \mathbf{J}$, que corresponde a la versión de la ANOVA en el espacio de proyección, Y, entonces $\mathbf{a}$ debe ser el vector propio asociado al valor propio más grande de la matriz $\mathbf{W}^{-1}\mathbf{B}$, que llameros $\mathbf{a}_1$. Con este vector construiríamos la primer función lineal discriminante

$$Y_1 = \mathbf{a}_1^{'}\mathbf{X}$$

Por construcción, esta función disciminante debe tener el mayor poder para discriminar entre los grupos. La segunda de estas funciones debe tener el mayor poder de disciminación restante, una vez construida la primer función discriminante, y debe ser ortogonal a la primera

$$Y_2 = \mathbf{a}_2^{'}\mathbf{X}, \quad Y_1 \perp Y_2 \Rightarrow \mathbf{a}_1 \perp \mathbf{a}_2$$

de forma análoga a la construcción de la primer función discriminante, se puede demostrar que el poder de disciminación de esta segunda función se maximiza si $\mathbf{a}_2$ es el correspondiente vector propoio asociado al segundo valor propio más grande de la matriz $\mathbf{W}^{-1}\mathbf{B}$. En general, se tiene que si $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_k$ son los vectores propios de $\mathbf{W}^{-1}\mathbf{B}$, asociados a los valores propios $\lambda_1, \lambda_2, ..., \lambda_k$, con $\lambda_1 > \lambda_2 > ... > \lambda_k$, entonces las funciones lineales

$$Y_i = \mathbf{a}_i^{'}\mathbf{X}, \quad i = 1, 2, ..., k$$

proporcionan máxima separación entre los G grupos proyectados. Además son ortogonales entre ellas.

Estimación

Supongamos que

- $\mathbf{X}_i$ es una matriz de datos $n_i \times p$ del grupo i=1,...,G

- $\overline{\mathbf{X}}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{X}_{ij}$ que es un estimador de μ_i

- $\mathbf{S}_i = \dfrac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)^{'}$ y

- $\overline{\mathbf{X}} = \left(\dfrac{1}{\sum_i n_i}\right)\sum_{i=1}^{G} n_i\overline{\mathbf{X}}_i = \left(\dfrac{1}{\sum_i n_i}\right)\sum_{i=1}^{G}\sum_{j=1}^{n_i}\mathbf{X}_{ij}$ que estima a $\overline{\mu}$

- $\mathbf{S}_{pool} = \sum_{g=1}^{G}\dfrac{n_g - 1}{n - G}\mathbf{S}_g$ es la matriz de covarianza común a los G grupos.

La estimación de $\mathbf{B}$, la correspondiente versión muestral de la suma de cuadrados entre grupos, es

$$\widehat{\mathbf{B}} = \sum_{i=1}^{G}\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)^{'}$$

con la correspondiente estimación de $\mathbf{W}$, la suma de cuadrados dentro de grupos, dada por

$$\widehat{\mathbf{W}} = \sum_{i=1}^{G}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)^{'}$$

Discriminante clásico para $G > 2$ grupos

Nuevamente, la idea para generalizar el procedimiento a G poblaciones normales es similar al anterior con dos poblaciones. En este caso, asignaremos el sujeto con covariables $\mathbf{X}$ al grupo $g = 1, 2, ..., G$ sii

$$\pi_g f_g\left(\mathbf{x}\right) > \pi_j f_j\left(\mathbf{x}\right) \quad \forall g \neq j \quad g, j = 1, 2, ..., G$$

Si las probabilidades apriori de pertenencia a cada grupo son iguales, y las matrices de varianza y covarianza son iguales entre los grupos, la condición anterior es equivalente a calcular la distancia de Mahalanobis del punto observado, $\mathbf{X}$, al centriode (vector de medias) de cada

población y clasificarlo en la población que haga mínima esta distancia. Al realizar el proceso que es semejante al caso de dos grupos obtenemos

Las funciones lineales discriminantes tienen las siguientes características:

• Y_1 es la combinación lineal que proporciona el mayor poder de discriminación entre los grupos, y está asociada al valor característico más grande de $\mathbf{W^{-1}E}$.

• Y_2 es la combinación lineal que proporciona el mayor poder discriminador entre los grupos, después de Y_1, y es *ortogonal* a Y_1. Esta función está asociada con el segundo valor característico más grande de $\mathbf{W^{-1}E}$.

Y así sucesivamente. El número máximo de funciones que se puede construir es *k=min(G-1,p)*.

En un proceso de análisis multivariado que lleva inmersa una reducción de dimensión, es muy importante determinar qué tan bien se reproducen los datos en las pocas dimensiones que se consideren para realizar su análisis. Una de las medidas más comunes para determinar lo adecuado de esta reducción de dimension, es el total de varianza explicada por las funciones lineales discriminantes. Una buena representación se logra si la varianza retenida por estas pocas dimensiones está cercana al 100%. El total de la variaza explicada por las primeras $m \leq k$ funciones lineales discriminantes es:

$$\sum_{i=1}^{m} \lambda_i \quad \text{y}$$

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{k} \lambda_i} \times 100\%$$ Porcentaje que explican las primeras m funciones discriminantes

Análisis de las funciones lineales discriminantes

Centroides de los grupos. La media de los puntajes que arrojen las evaluaciones de cada individuo en estas funciones lineales discriminantes. Debería ser una medida inicial de qué tan separados están los grupos *proyectados*. Si la discriminación es buena, deberíamos observar centroides muy alejados uno de otro.

$\mathbf{\Lambda}$ de Wilks. Esta estadística sirve para determinar el poder discriminante de cada una de las funciones discriminantes. Se determina de forma secuencial el número de estas funciones que debemos considerar, a través de la estadística

$$\mathbf{\Lambda} = \frac{\text{Suma de cuadrados dentro de grupos}}{\text{Suma de cuadrados totales}} = \frac{|\mathbf{B}|}{|\mathbf{T}|} = \frac{|\mathbf{B}|}{|\mathbf{W}+\mathbf{B}|}$$

Si la discriminación lograda es buena, entonces la varianza dentro de los grupos será pequeña y la varianza entre grupos será grande. Por lo tanto, $\mathbf{\Lambda}$ estará cercana a cero.

Para este fin, es preferible utilizar el estadístico $\mathbf{V}$ de Barlett, que es una función de $\mathbf{\Lambda}$ y tiene distribución asintótica χ^2. El procedimiento es, inicialmente, considerar sólo una función discriminante, realizar la prueba y, si ésta es significativa, querrá decir que es pertinente la incorporación de otra función discriminante, de lo contrario, será indicativo de que con el número actual de funciones se tiene el máximo poder discriminante; equivalentemente, que la inclusión de otra función no aporta nada a la discriminación entre los grupos.

Las hipótesis a probar mediante este procedimiento son

$$\mathbf{H}_0 : k \text{ funciones lineales son suficientes para discriminar vs.}$$
$$\mathbf{H}_a : \text{ son necesarias más de } k \text{ funciones } \ k = 1, 2, ..., min(G-1, p)$$

La manera de determinar la *importancia relativa de las variables dentro de las funciones discriminantes*, es a través de sus coeficientes estandarizados. La razón es que éstos ya están libres de unidades y son comparables. *La variable que posea el coeficiente estandarizado más grande en valor absoluto*, será la que tiene un *poder discriminante mayor*.

Coeficientes de correlación o de estructura

$Corr(X_i, \mathbf{Y}_g)$: Correlación lineal entre cada una de las variables y cada una de las funciones lineales. Si esta correlación es grande (cercana a uno en valor absoluto) indica una relación lineal fuerte entre la variable y la función, por tanto, la variable tiene una contribución importante para discriminar entre los grupos. Si está cercana a cero, no tiene poder discriminatorio entre los grupos.

Tasa de error de clasificación: Un elemento muy importante, que determina qué tan bien clasifica nuestro discriminante a las observaciones en la población, es la *tasa de error de clasificación*. Si las covariables utilizadas realmente discriminan a los grupos en la población, esta tasa debe ser pequeña, de lo contrario, será grande y concluiremos que las variables utilizadas, no tienen poder de discriminación entre los grupos en la población.

Cuando se hace esta clasificación con la misma muestra que se utilizó para construir el discriminante, generalmente se logra una tasa de error de clasificación"artificialmente" baja. Una forma más honesta de calcular esta tasa, es a través de la llamada *clasificación cruzada*, que no es más que eliminar uno por uno a las observbaciones en la muestra, y utilizar el discriminante para asignarlas a algunos de los grupos; por lo regular, este procedimiento genera tasa de error más elevadas, pero más realistas.

Regresión multinomial

En el caso de que existan más de dos grupos en el proceso de análisis discriminante, el discriminante logístico se generaliza a una variable de respuesta con más de dos categorías nominales, dando origen al llamado *modelo de regresión multinomial*. En este caso tenemos

$\eta_{ij} = log\left(\frac{\pi_{ij}}{\pi_{iG}}\right) = \alpha_j + \mathbf{X}_i'\beta_j$ entonces

$$\pi_{ij} = P(G_j|\mathbf{X}_i) = \frac{exp\,(\eta_{ij})}{\sum_{i=1}^{p} exp\,(\eta_{ij})}, \quad j = 1, 2, ..., G$$

denota la probabilidad de que el sujeto i pertenezca al grupo j.

Discriminante cuadrático

Supongamos que las poblaciones son normales, pero que, como ocurre regularmente, no existe igualdad de varianzas; en el caso de dos grupos, $\Sigma_1 \neq \Sigma_2$. Entonces, la regla de clasificación, bajo el supuesto de probabilidades a priori iguales, es:

$$\mathbb{Q}(\mathbf{X}) = \frac{1}{2}\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)\mathbf{X} + \mathbf{X}'\left(\Sigma_2^{-1}\mu_1 - \Sigma_1^{-1}\mu_2\right) + \frac{1}{2}\mu_2'\Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1'\Sigma_1^{-1}\mu_1 + \frac{1}{2}log\,|\Sigma_2| - \frac{1}{2}log\,|\Sigma_1|$$

Observemos que el término $\mu_i'\Sigma_i^{-1}\mu_i,\ i = 1, 2$, no puede cancelarse y origina términos de grado 2, ya sean cuadráticos o cruzados, lo que justifica el nombre de discriminante cuadrático.

Esta regla es equivalente a asignar a un individuo $\mathbf{X}_0$ al grupo donde se minimice la función

$$\min_{j\in(1,2)}\left[\frac{1}{2}log\,|\Sigma_j| + \frac{1}{2}\left(\mathbf{X}_0 - \mu_j\right)'\Sigma_j^{-1}\left(\mathbf{X}_0 - \mu_j\right)\right]$$

Para el caso de $G > 2$ grupos, y suponiendo que las matrices de varianza-covarianza no son iguales, la regla se extiene trivialmente como: asignar a un individuo $\mathbf{X}_0$ al grupo donde se minimice la función

$$\min_{j\in(1,\ldots,G)}\left[\frac{1}{2}log\,|\Sigma_j| + \frac{1}{2}\left(\mathbf{X}_0 - \mu_j\right)'\Sigma_j^{-1}\left(\mathbf{X}_0 - \mu_j\right)\right]$$

Análisis de conglomerados (clusters)

El análisis de conglomerados (clusters) es una técnica multivariada, cuyo objetivo es identificar los grupos que subyacen a un conjunto de observaciones. La idea es "descubrir" grupos de observaciones homogéneas y que estén separados de otros grupos. En mercadotecnia, por ejemplo, puede ocurrir que una muestra de consumidores con distintas características, esté formada por un pequeño número de grupos dentro de cada uno de los cuales dichas características sean similares. Esto podría tener implicaciones importantes para determinar una estrategia de mercado apropiada o para investigar la tipología del consumidor. En un contexto educativo, los grupos pueden ser conjuntos de individuos con distintas capacidades (grupos de excelencia, estándar o de bajo rendimiento) o con diversos intereses, que los pueden ubicar en distintas áreas de estudio (orientación vocacional). En Biología podría tratarse de diversos tipos de individuos que pertenecen a una misma especie. En ecología podrían referirse a distintos tipos de plantas. En seguros, podemos agrupar a los sujetos que representan riesgos diferentes en alguna cobertura sobre, por ejemplo, automóviles. En fin, existe un sinnúmero de situaciones reales donde, en algún sentido, se tiene que trabajar con grupos de observaciones o individuos.

Estos métodos se conocen también con el nombre de métodos de *clasificación automática o no supervisada, o de reconocimiento de patrones sin supervisión.* El nombre de no supervisados se aplica para distinguirlos del análisis discriminante, que estudiamos en la sección anterior. Este nombre se debe a que, a diferencia del análisis discriminante, aquí no conocemos la naturaleza de los grupos, de hecho, ni siquiera sabemos el número de grupos, antes de clasificar las observaciones dentro de los clusters.

Objetivos:

• Identificar los grupos que de manera natural se forman con los datos

Estos grupos se forman con base a las similaridades o disimilaridades entre los sujetos, no entre las variables. En este sentido, esta es una técnica de análisis multivariado determinada por los sujetos (casos) y no por las variables como, por ejemplo, en el análisis de componentes principales o en el análisis de factores.

• Podemos decir que esta técnica tiene más fundamento computacional que estadístico

• Es una técnica descriptiva

• Aunque el objetivo común es agrupar a los sujetos, el análisis de conglomerados también se puede utilizar para agrupar variables, de una forma similar al análisis de factores.

Consideraciones antes de realizar el análisis de conglomerados

Al hacer un análisis de conglomerados con un conjunto de datos, nos enfrentamos a una serie de cuestionamientos que debemos dar respuesta para llevar a cabo nuestro objetivo. A saber

• Una primer pregunta es ¿qué variables debemos elegir para realizar los clusters?. Aunque esta es una elección muy importante, raras veces es considerada como tal, y, en la práctica, involucra una mezcla de intuición y disponibilidad de los datos.

• ¿Qué medida de distancia utilizar entre los casos?

• ¿Qué tipo de liga utilizar para los grupos?

• ¿Qué tipo de técnica de construcción de los conglomerados usar?

Pasos en el análisis de conglomerados

• Si las variables no están medidas en la misma escala, es conveniente hacer el análisis con las variables estandarizadas. El objetivo es que las variables con mayores magnitudes no dominen el análisis (similar a Componentes Principales)

• Selección de variables. Como este proceso no proporciona ninguna medida acerca de la importancia de una variable en el análisis, ésta es una decisión que el usuario debe hacer *CUIDADOSAMENTE*.

• Construir y evaluar el modelo de conglomerados

• Identificar la pertenencia (membresía) de los casos a su correspondiente cluster.

Tipos de distancias para los casos, de acuerdo a su escala de medición

La primera decisión importante que se debe tomar es sobre cómo calcular la distancia entre dos observaciones. Es claro que esta elección dependerá de la escala de medición de las variables involucradas en la misma.

En realidad, es bastante subjetivo el hecho de elegir una medida de similitud ya que depende de las escalas de medida. Para variables nominales, generalmente se utilizan medidas de similaridad, mientras que para variable medidas en escala de intervalo o de razón usualmente se consideran matrices de distancias. Se pueden agrupar observaciones (sujetos) según la similitud expresada en términos de una distancia. Si el objetivo es agrupan variables (tipo análisis de factores), es habitual utilizar como medida de similitud los coeficientes de correlación en valor absoluto. Para variables categóricas existen también criterios basados en la posesión o no de los atributos (tablas de presencia-ausencia).

Distancia

Dados dos vectores $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^p$, la distancia entre ellos es una función d con las siguientes propiedades:

i) d: $\mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+ \cup 0$, i.e., $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$

ii) $d(\mathbf{x}_i, \mathbf{x}_i) = 0$, $\forall\ \mathbf{x}_i$

iii) $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$ (simetría)

iv) $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_k, \mathbf{x}_j)$ (desigualdad del triángulo)

Variables continuas

• *Distancia euclidiana (la más común)* Supongamos que tenemos dos sujetos con p variables, i.e., $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$. Entonces, su distancia euclideana es

$$\begin{aligned} d(\mathbf{x}_1, \mathbf{x}_2) &= \left[\sum_{i=1}^{p} (x_{1i} - x_{2i})^2\right]^{1/2} \\ d^2(\mathbf{x}_1, \mathbf{x}_2) &= (\mathbf{x}_1 - \mathbf{x}_2)^t (\mathbf{x}_1 - \mathbf{x}_2) \text{ (forma vectorial)} \end{aligned}$$

• *Distancia euclideana al cuadrado*

$$d^2(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{p} (x_{1i} - x_{2i})^2$$

• *Distancia de Mahalanobis*

$$d^2(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)^t \, \mathbf{S}^{-1} \, (\mathbf{x}_1 - \mathbf{x}_2)$$

donde $\mathbf{S}$ es la matriz de covarianzas entre las variables. De este modo, las distancias se ponderan según el grado de relación que exista entre las variables, es decir, si están más o menos correlacionadas. Si la correlación es nula, se obtiene la distancia euclideana.

• *Distancia City block*

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{p} |x_{1i} - x_{2i}|$$

• *Distancia de Minkowski*

$$d(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{p} (x_{1i} - x_{2i})^p \right]^{1/p}$$

Si *p=1* tenemos la distancia City block y si *p=2* la distancia euclideana. Si $p = \infty$ se tiene la distancia de Chebychev, dada por

$$D_\infty = \max_{1 \leq i \leq p} |x_{1i} - x_{2i}|$$

• Distancia de Camberra

$$d_{Can}\,(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{p} \frac{|x_{1i} - x_{2i}|}{|x_{1i} + x_{2i}|}$$

Definida como *cero* si $x_{1i} = x_{2i}$.

Variables de conteo (numéricas discretas)

- Ji-cuadrada

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{\sum_i \left(x_{1i} - \mathbb{E}\left(x_{1i}\right)\right)^2}{\mathbb{E}\left(x_{1i}\right)} + \frac{\sum_i \left(x_{2i} - \mathbb{E}\left(x_{2i}\right)\right)^2}{\mathbb{E}\left(x_{2i}\right)}}$$

- phi-cuadrada

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{\frac{\sum_i \left(x_{1i} - \mathbb{E}\left(x_{1i}\right)\right)^2}{\mathbb{E}\left(x_{1i}\right)} + \frac{\sum_i \left(x_{2i} - \mathbb{E}\left(x_{2i}\right)\right)^2}{\mathbb{E}\left(x_{2i}\right)}}{n}}$$

Variables dicotómicas

- Distancia euclidiana

- Distancia euclidiana al cuadrado

Datos binarios (medidas de similaridad)

En este caso se desea medir la similaridad entre los vectores $x_i = (x_{i1}, ..., x_{ip})^{'}$ y $x_j = (x_{j1}, ..., x_{jp})^{'}$, con la característica particular de que $x_{ik},\ x_{jk} \in \{0, 1\}$, $\forall k = 1, 2, ..., p$. Que generan los siguientes casos

$$\begin{aligned} x_{ik} &= x_{jk} = 1, \\ x_{ik} &= x_{jk} = 0, \\ x_{ik} &= 1, x_{jk} = 0, \\ x_{ik} &= 0, x_{jk} = 0 \end{aligned}$$

Si definimos ahora

$$
\begin{aligned}
a_1 &= \sum_{k=1}^{p} \mathbf{I}\left(x_{ik} = x_{jk} = 1\right) \\
a_2 &= \sum_{k=1}^{p} \mathbf{I}\left(x_{ik} = 0, x_{jk} = 1\right) \\
a_3 &= \sum_{k=1}^{p} \mathbf{I}\left(x_{ik} = 1, x_{jk} = 0\right) \\
a_1 &= \sum_{k=1}^{p} \mathbf{I}\left(x_{ik} = x_{jk} = 0\right)
\end{aligned}
$$

en la práctica, es frecuente el uso de la siguiente medida de similiradidad para este tipo de datos

$$
s_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda\left(a_2 + a_3\right)}
$$

Donde los valores de δ y λ representan pesos que definen distintas medidas de similaridad. Las más comunes se presentan en la tabla siguiente.

Nombre de la medida de similaridad	δ	λ	Definición
Jaccard	0	1	$\frac{a_1}{a_1 + a_2 + a_3}$
Tanimoto	1	2	$\frac{a_1 + a_4}{a_1 + 2\left(a_2 + a_3\right) + a_4}$
Simple Matching (M)	1	1	$\frac{a_1 + a_4}{p}$
Russel and Rao (RR)	–	–	$\frac{a_1}{p}$
Dice	0	1/2	$\frac{2a_1}{2a_1 + \left(a_2 + a_3\right)}$
Kulczynski	–	–	$\frac{a_1}{a_2 + a_3}$

El rango de estas medidas de similaridad está entre cero (mínima similaridad) y uno (máxima similaridad). La idea es ponderar de manera distinta el número de *acuerdos* y *desacuerdos* entre los valores de los vectores binarios observados.

Existen muchísimas medidas de similaridad definidas para datos binarios, estás son las principales, pero, por ejemplo, en los manuales del *innombrable* aparecen las siguientes, además de las ya mencionadas: Sokal and Sneath similarity measure 1, Sokal and Sneath similarity measure 2, Sokal and Sneath similarity measure 3, Ochiai similarity measure, Sokal and Sneath similarity measure 5, Fourfold point correlation (similarity), Binary Euclidean distance, Binary squared Euclidean distance, etc.

Variables continuas (medidas de similaridad)

- Correlación de Pearson

$$r(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{i=1}^{p} (x_{1i} - \bar{x}_{1i})(x_{2i} - \bar{x}_{2i})}{\sqrt{\sum_{i=1}^{p} (x_{1i} - \bar{x}_{1i})^2 \sum_{i=1}^{p} (x_{2i} - \bar{x}_{2i})^2}}$$

- Coseno

$$cos(\theta) = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|} = \frac{\sum_{i=1}^{p} x_{i1} x_{2i}}{\sqrt{\sum_{i=1}^{p} x_{1i}^2} \sqrt{\sum_{i=1}^{p} x_{2i}^2}}$$

Entonces, estas funciones de distancia transforman nuestra matriz de datos $\mathbf{X}_{n \times p}$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

en una matriz de distancias o similaridades, $\mathbf{D}_{n\times n}$, entre los n sujetos

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

Distancias entre clusters

Una vez que se ha definido qué distancia conviene usar para los casos, debemos decidir cómo se habrá de calcular la distancia de un individuo a un conglomerado y la distancia entre los conglomerados. Para este fin, se tiene las siguientes medidas, conocidas en la literatura de análisis de conglomerados, como *ligas*. Ilustraremos cada una de estas ligas con la matriz de distancia

Distancias

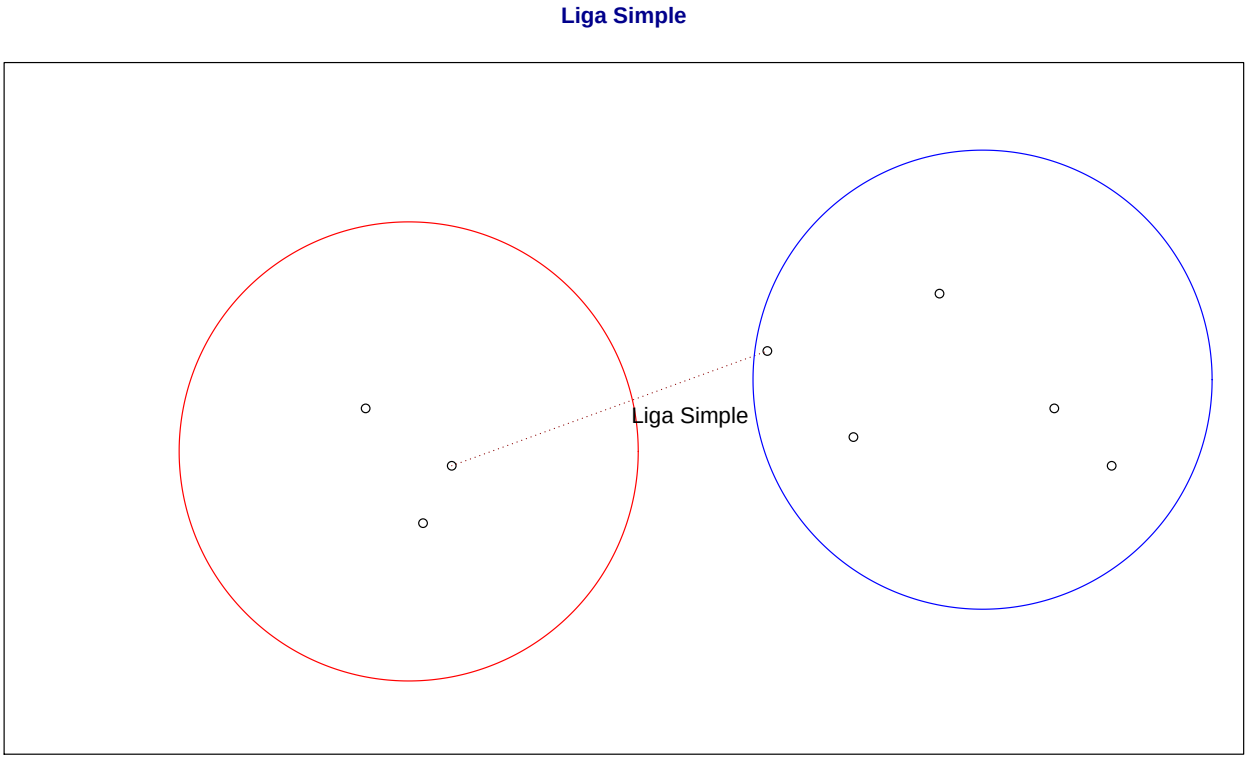
	1	**2**	**3**	**4**	**5**
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

Vecinos cercanos o liga simple

Aquí la distancia entre dos conglomerados es la distancia entre sus sujetos más cercanos.

En términos matemáticos, si tenemos un cluster **R** y otro **S**, entonces la distancia es:

$$d(R,S) = min(d_{ij},\ i \in \mathbf{R}, j \in \mathbf{S})$$



Observando las distancias entre todos los sujetos en nuestra matriz de distancias, $\mathbf{D}$, observamos que la mínima de éstas es *3* y corresponde a los sujetos $(3,5)$, por lo que son los primeros en unirse y lo hacen a una *altura=2*. Ahora bien, esta unión de sujetos ya constituye un grupo, por lo que hay que calcular las distancias de este grupo al resto de los elementos utilizando la liga simple, es decir, hay que calcular

$$\begin{aligned} d\{(3,5),1\} &= min\{d(3,1),d(5,1)\} = min\{3,11\} = 3 \\ d\{(3,5),2\} &= min\{d(3,2),d(5,2)\} = min\{7,10\} = 7 \\ d\{(3,5),4\} &= min\{d(3,4),d(5,4)\} = min\{9,8\} = 8 \end{aligned}$$

La nueva matriz de distancias $\mathbf{D}_1$, se obtiene considerando estas distancias calculadas con este primer grupo formado. En concreto tenemos

$\mathbf{D}_1$

	(3,5)	**1**	**2**	**4**
(3,5)	0	3	7	8
1	3	0	9	6
2	7	9	0	5
4	8	6	5	0

Realizando el mismo proceso inicial, la distancia mínima entre estos grupos es *3* y corresponde a los grupos $(3,5)$ y *1*. Entonces, el siguiente agrupamiento genera al grupo $(1,3,5)$. Nuevamente, debemos calcular la distancia de este grupo a cada uno de los otros elementos, a través de la liga simple

$$
\begin{aligned}
d\{(1,3,5),2\} &= min\{d(1,2),d(3,2),d(5,2)\} = min\{9,7,10\} = 7 \\
d\{(1,3,5),4\} &= min\{d(1,4),d(3,4),d(5,4)\} = min\{6,9,8\} = 6
\end{aligned}
$$

Por lo que nuestra nueva matriz de distancias es

$\mathbf{D}_2$

	(1,3,5)	**2**	**4**
(1,3,5)	0	7	6
2	7	0	5
4	6	5	0

La distancia mínima en esta matriz es *5* y corresponde a la de los sujetos $(2,4)$, que constituyen el siguiente grupo formado y que, como vemos, es diferente al que ya habíamos constituido. Ahora debemos encontrar la distancia entre estos dos grupos

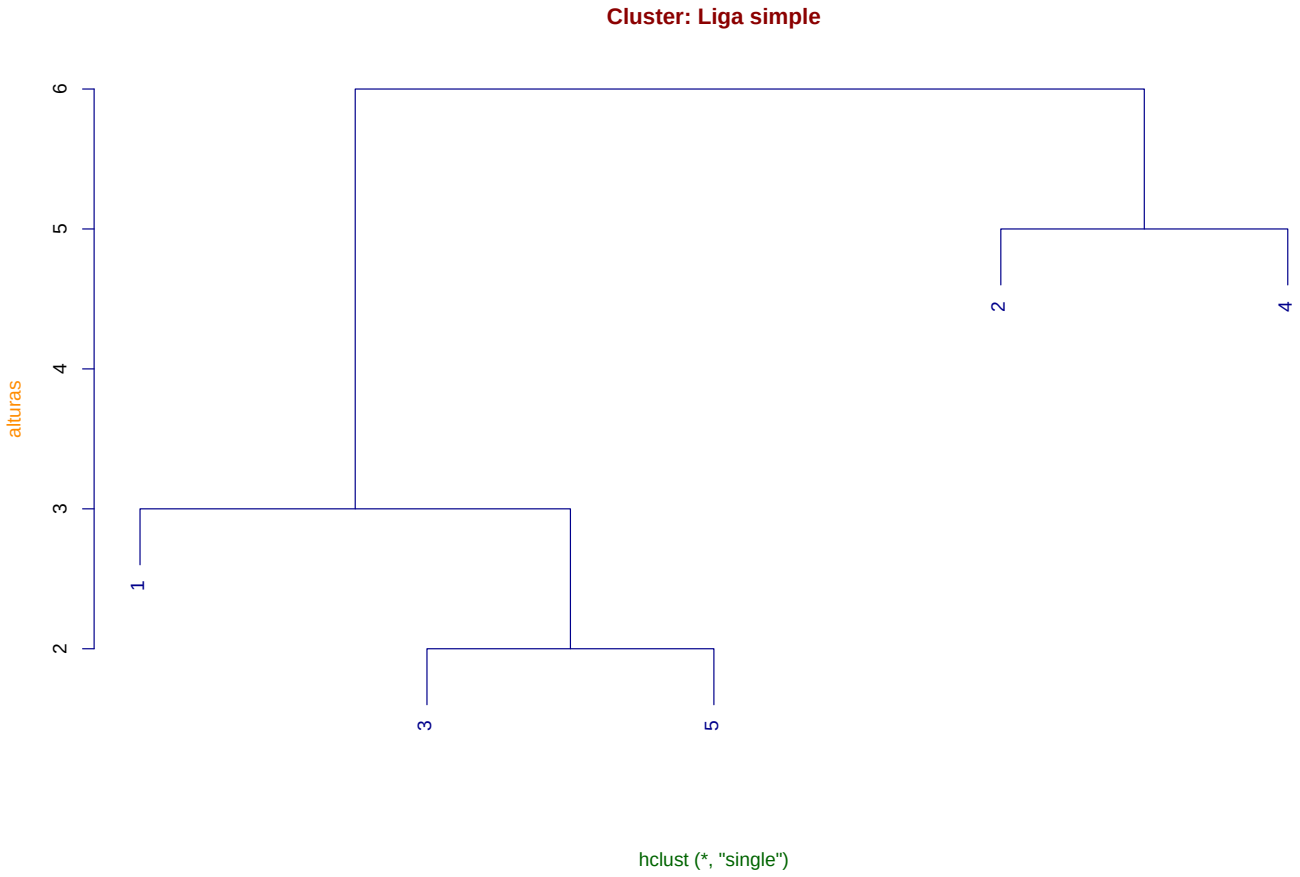
$$
d\{(1,3,5),(2,4)\} = min\{d(1,2),d(3,2),d(5,2),d(1,4),d(3,4),d(5,4)\} = min\{9,7,10,6,9,8\} = 6
$$

Y la última de nuestras matrices es

$\mathbf{D}_3$

	(1,3,5)	**(2,4)**
(1,3,5)	0	6
(2,4)	6	0

Esta distancia a la que se unen los dos grupos formados, constituye la distancia máxima a la que se unen *todas las observaciones* y genera un único grupo, como ya sabemos que debe suceder con un algoritmo *aglomerativo*. Observemos que esta agrupación, así como las distancias (alturas) de unión entre los grupos, coincide totalmente con el *dendrograma* mostrado en la gráfica siguiente.

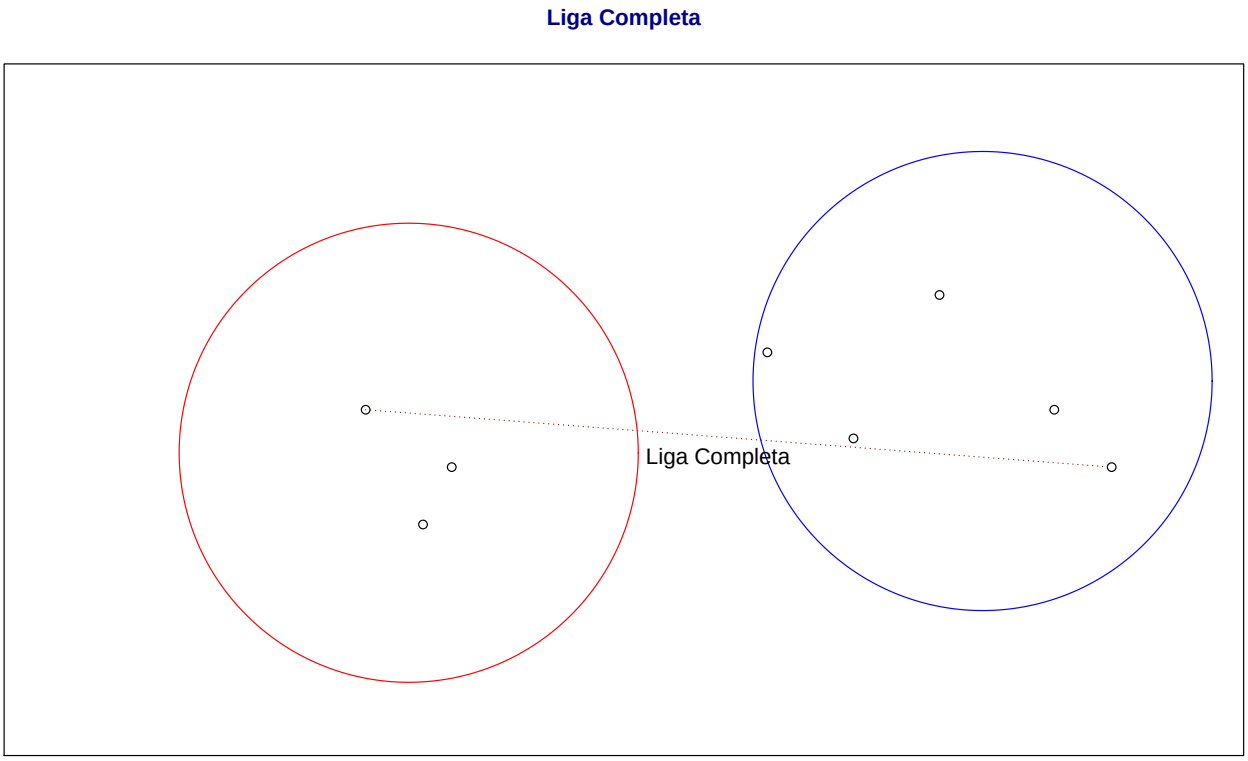


Vecinos lejanos o liga compuesta

Aquí La distancia entre dos conglomerados es la distancia entre sus dos sujetos más lejanos

En términos matemáticos, si tenemos un cluster **R** y otro **S**, entonces la distancia es:

$$d(R,S) = max(d_{ij},\ i \in \mathbf{R}, j \in \mathbf{S})$$



Para esta liga, el primer conglomerado se hace igual que en la liga simple, y está constituido por $(3,5)$, y se unen a altura 2. Una vez establecido este grupo, procedemos a encontrar su distancia al resto de los elementos, utilizando la liga completa, de la siguiente forma

$$\begin{aligned} d\{(3,5),1\} &= max\{d(3,1),d(5,1)\} = max\{3,11\} = 11 \\ d\{(3,5),2\} &= max\{d(3,2),d(5,2)\} = max\{7,10\} = 10 \\ d\{(3,5),4\} &= max\{d(3,4),d(5,4)\} = max\{9,8\} = 9 \end{aligned}$$

La nueva matriz de distancias $\mathbf{D}_1$, se obtiene considerando estas distancias calculadas con

este primer grupo formado. En concreto tenemos

$\mathbf{D_1}$

	(3,5)	**1**	**2**	**4**
(3,5)	0	11	10	9
1	11	0	9	6
2	10	9	0	5
4	9	6	5	0

La distancia mínima en esta matriz corresponde a los individuos $(2, 4)$, con altura=5, que forman el siguiente grupo. El siguiente paso es calcular la distancia entre estos nuevos grupos, con la liga completa.

$$\begin{aligned} d\{(3,5),(2,4)\} &= max\{d(3,2), d(3,4), d(5,2), d(5,4)\} = max\{7,9,10,8\} = 10 \\ d\{(3,5),1\} &= max\{d(3,1), d(5,1)\} = max\{3,11\} = 11 \\ d\{(2,4),1\} &= max\{d(2,1), d(4,1)\} = max\{9,6\} = 9 \end{aligned}$$

con lo que generamos la siguiente matriz

$\mathbf{D_2}$

	(3,5)	**1**	**(2,4)**
(3,5)	0	11	10
1	11	0	9
(2,4)	10	9	0

observamos que la distancia más pequeña es *9*, y corresponde a la unión de los grupos $(2, 4)$ y *1*, que generan el grupo $(1, 2, 4)$. Volvemos a calcular la distancia entre los grupos $(3, 5)$ y $(1, 2, 4)$ a través de la liga completa.

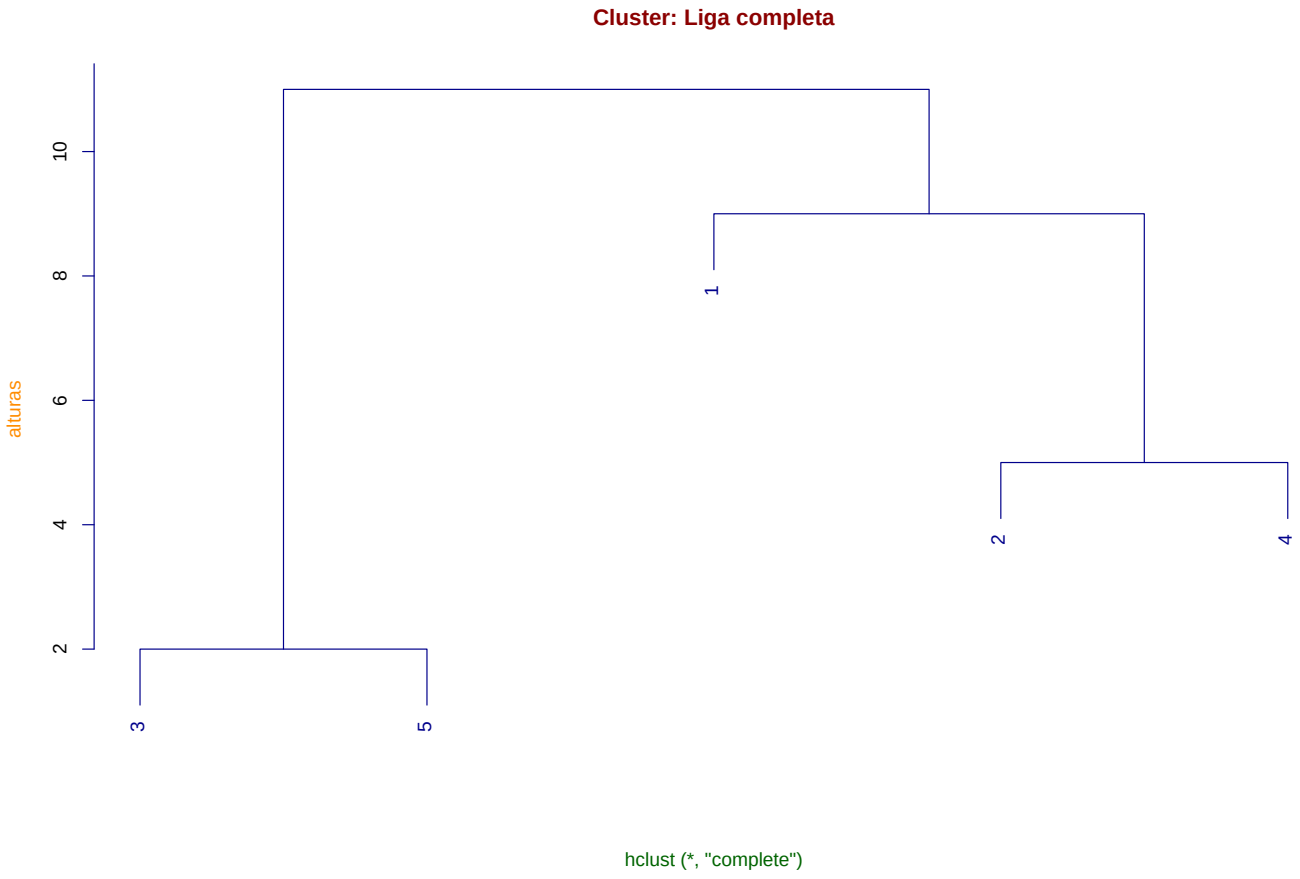
$$\begin{aligned} d\{(3,5),(1,2,4)\} &= max\{d(3,1), d(3,2), d(3,4), d(5,1), d(5,2), d(5,4)\} \\ &= max\{3,7,9,11,10,8\} = 11 \end{aligned}$$

y genera la matriz

$\mathbf{D_3}$

	(3,5)	**(1,2,4)**
(3,5)	0	11
(1,2,4)	11	0

esta es la distancia a la que todas las observaciones se unen en un solo grupo. Nuevamente observamos que los grupos formados y las distancias (alturas) que calculamos, coinciden con la gráfica de esta liga.

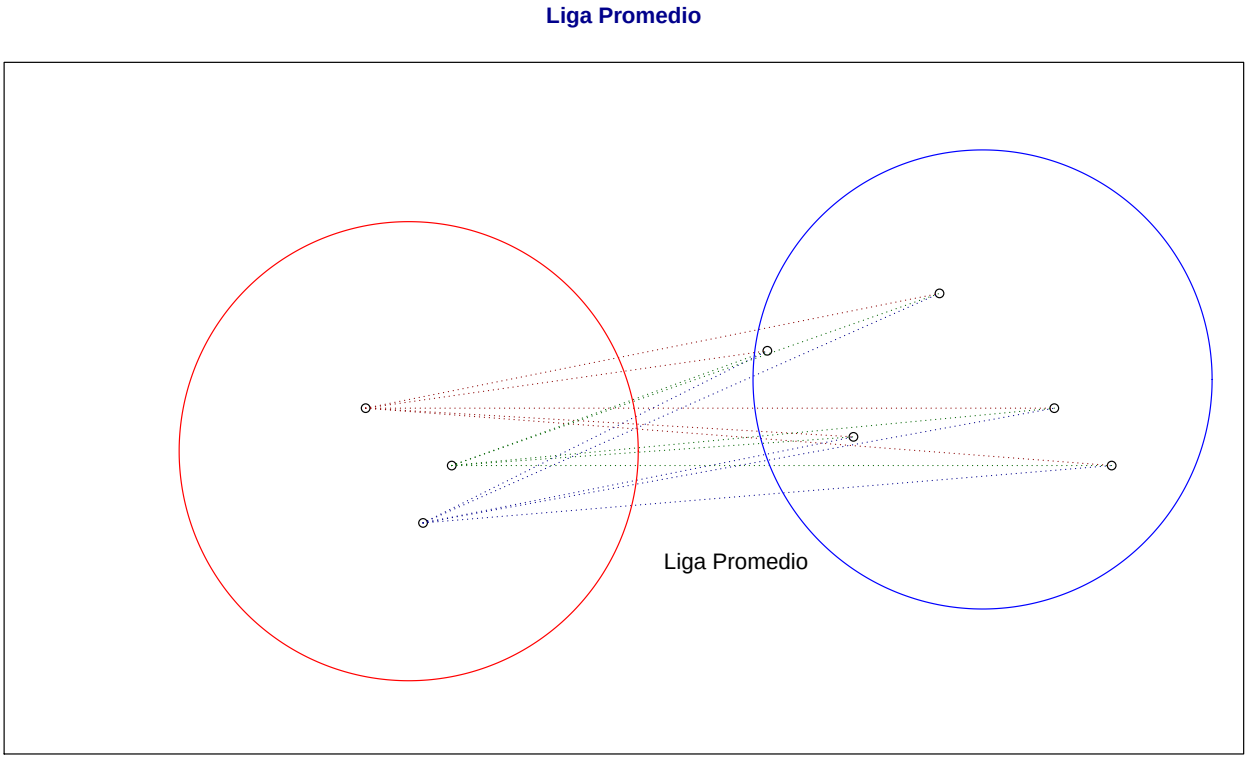


Liga promedio

Es la distancia promedio entre todas las posibles distancias intra o inter clusters. Apropiada, cuando el investigador asume que los grupos son homogéneos.

En símbolos, si tenemos un cluster $\mathbf{R}$ con n_R elementos, y otro $\mathbf{S}$, con n_S elementos, entonces la distancia es:

$$d(R,S) = \frac{1}{n_R}\frac{1}{n_S}\sum_{i\in\mathbf{R}}\sum_{j\in\mathbf{S}} d_{ij}$$



Igual que para los dos casos anteriores, las observaciones que se unen inicialmente, son $(3,5)$, que se unen a altura *2*. Una vez que se obtiene este cluster, hay que calcular sus distancia al resto de los elementos, utilizando la liga promedio. Es decir

$$
\begin{aligned}
d\{(3,5),1\} &= \frac{1}{2}\{d(3,1)+d(5,1)\} = \frac{1}{2}(3+11) = 7 \\
d\{(3,5),2\} &= \frac{1}{2}\{d(3,2)+d(5,2)\} = \frac{1}{2}(7+2) = 4.5 \\
d\{(3,5),4\} &= \frac{1}{2}\{d(3,4)+d(5,4)\} = \frac{1}{2}(9+8) = 8.5
\end{aligned}
$$

y la correspondiente matriz de distancias es ahora

$\mathbf{D}_1$

	(3,5)	**1**	**2**	**4**
(3,5)	0	7	4.5	8.5
1	7	0	9	6
2	4.5	9	0	5
4	8.5	6	5	0

La distancia mínima en esta matriz corresponde a las observaciones $(2,4)$, que forman un nuevo grupo, que se une a *altura=5*. Nuevamente debemos calcular la distancia entre estos clusters mediante la liga promedio.

$$
\begin{aligned}
d\{(3,5),(2,4)\} &= \frac{1}{4}\{d(3,2)+d(5,4)+d(5,2)+d(5,4)\} = \frac{1}{4}(7+9+10+8) = 8.5 \\
d\{(3,5),1\} &= \frac{1}{2}\{d(3,1)+d(5,1)\} = \frac{1}{2}(3+11) = 7 \\
d\{(2,4),1\} &= \frac{1}{2}\{d(2,1)+d(4,1)\} = \frac{1}{2}(9+6) = 7.5
\end{aligned}
$$

que genera la matrz de distancias

$\mathbf{D}_2$

	(3,5)	**1**	**(2,4)**
(3,5)	0	7	8.5
1	7	0	7.5
(2,4)	8.5	7.5	0

cuya distancia mínima es *7* y corresponde a la unión de los grupos $(3,5)$ y *1*, que originan el grupo $(1,3,5)$. La distancia entre estos grupos es

$$
\begin{aligned}
d\{(1,3,5),(2,4)\} &= \frac{1}{6}\{d(1,2)+d(3,2)+d(5,2)+d(1,4)+d(3,4)+d(5,4)\} \\
&= \frac{1}{6}(9+7+9+6+10+8) = 8.166
\end{aligned}
$$

que genera la matriz

$\mathbf{D}_3$

	(1,3,5)	**(2,4)**
(3,5)	0	8.166
(2,4)	8.166	0

entonces, la distancia final a la que se unen todos los grupos es *8.166*. Observe que los grupos y las distancias a las que se unen, coinciden con la gráfica correspondiente a esta distancia promedio.