Introducción.

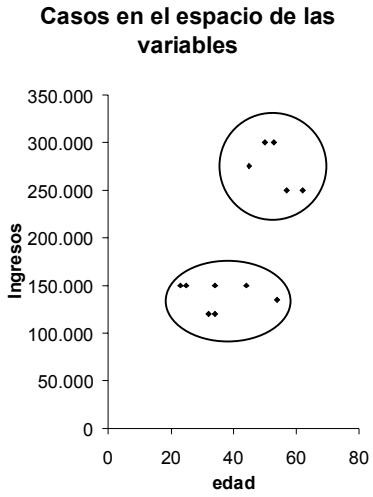
La técnica del análisis de conglomerados tiene como objetivo investigar la estructura de grupos que pudiera existir en un conjunto de datos. Y ya la simple contemplación de cualquier representación de individuos o casos en el espacio de sus variables observadas, nos puede llevar a reconocer, de una forma puramente intuitiva, ciertas "agrupaciones" de casos y a distinguir entre "diferentes grupos", ya que solemos identificar el "concepto de semejanza o parecido entre individuos" con la proximidad física o geométrica de los puntos que los representan; y al "concepto de grupo", con el conjunto de puntos que se encuentran más cercanos entre sí que comparativamente con el resto.

Para confirmar lo que decimos, consideremos a modo de ejemplo el siguiente conjunto de datos, en el que recogemos un conjunto de individuos a los cuales les hemos medido una serie de variables: la edad, el sexo, los ingresos, cómo participan los ingresos de las familias, el nivel de estudios, el número de miembros de las familias y el IRPF que cotizan. :

Caso	Edad	Sexo	Ingresos	Part. Ingre.	Niv. Estud	Miem-bros	IRPF
1	34	1	120.000	100	1	3	22,1
2	45	1	275.000	85	2	3	24,5
3	34	2	150.000	50	1	4	18,0
4	25	1	150.000	35	3	2	23,1
5	62	2	250.000	99	1	2	32,3
6	53	1	300.000	75	1	3	34,1
7	32	2	120.000	100	2	3	22,1
8	54	2	135.000	85	2	3	24,5
9	23	2	150.000	50	3	4	18,0
10	44	1	150.000	35	1	2	23,1
11	57	1	250.000	100	2	2	32,3
12	50	2	300.000	75	1	3	34,1

Para representar estos casos en el espacio de las variables, de forma completa, necesitaríamos recurrir al espacio de las 7 variables observadas; esto es, a un espacio de dimensión 7. Como la representación gráfica de la nube de puntos en tal espacio no resulta fácilmente asimilable por nuestra mente, acostumbrada a utilizar los patrones de la geometría euclídea de no más de tres dimensiones, planteemos la siguiente visión parcial de dicha representación, consistente en observar su proyección sobre el plano de dimensión dos formado, por ejemplo, con las variables edad e ingresos.

Con esta aproximación, y a la vista de su representación gráfica, podemos preguntarnos ¿qué estructura de grupos se observa en los datos? La respuesta admite una amplia gama de interpretaciones resultantes de diferentes planteamientos. Así, en una primera instancia y asimilando el concepto de "proximidad" con la intuitiva distancia euclídea entre los puntos, podríamos pensar que los dos grupos de elementos más "próximos" serían los que se señalan rodeados en el gráfico siguiente.

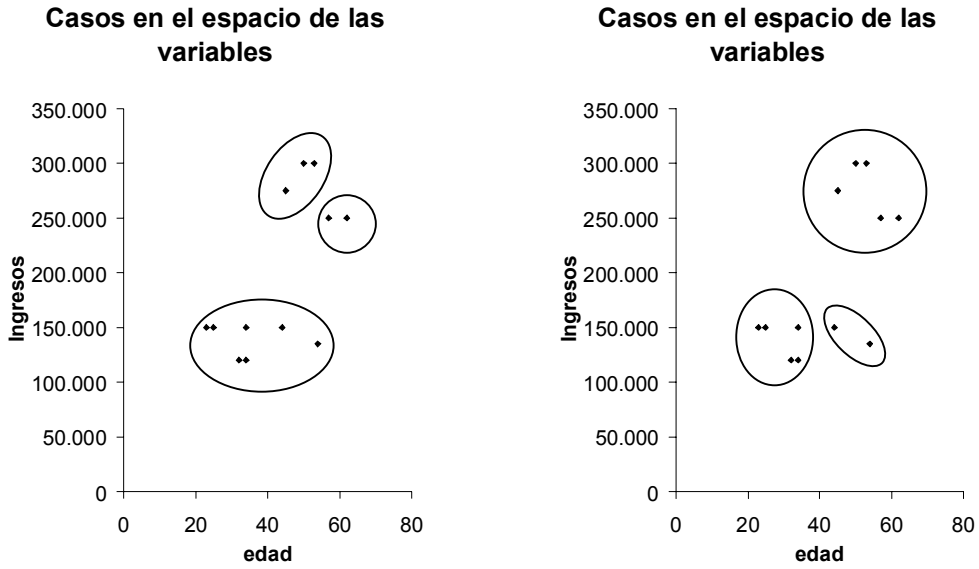


Sin embargo, este concepto intuitivo de similitud entre individuos, que asimilamos a distancias pequeñas entre los mismos, debe entenderse bien y exclusivamente en términos de las variables consideradas y de "sus propiedades". Pensemos en cómo podríamos medir con rigor la distancia entre puntos en espacios mixtos cuantitativos-cualitativos en que se consideran simultáneamente variables (ingresos, edad,...) y atributos (sexo, nivel de estudios,...) como el que tenemos. La respuesta no es trivial y trataremos de dar en el capítulo algunas guías de actuación para abordar este problema.

Además, si introducimos en el estudio una nueva variable, podríamos ver que la estructura de grupos considerada pudiera no ser la más adecuada, debido a la información que introduce la nueva variable. Así, en nuestro ejemplo, fijémonos en dos de los puntos aparentemente más próximos cuando se consideran exclusivamente la edad y los ingresos (casos 1 y 7), los cuales perciben el mismo ingreso y tienen edades que sólo difieren en dos años. ¿Qué ocurriría si plantáramos una dimensión tercera, por ejemplo el sexo? Resultaría que para el caso 1 la variable sexo toma el valor 1 y para el caso 7 la variable toma el valor 2, de tal manera que estos dos datos, que a primera vista en la representación parcial parecían estar tan juntos, resulta que realmente no lo están tanto, ya que al considerar esa nueva dimensión, cada uno de los casos se situaría en lados totalmente opuestos sobre esta tercera dimensión. Lo mismo ocurre para otros casos aparentemente juntos sobre la nube proyectada en el espacio de dimensión dos, por lo que resulta que la nube de puntos globalmente considerada en todas sus dimensiones no es tan homogénea como aparenta en la proyección simplificada.

Pero aún asumiendo que la distancia euclídea represente bien nuestro concepto de proximidad entre los casos y aún admitiendo que las demás variables no produjesen diferencias de comportamiento entre casos que en este espacio podrían parecer muy próximos o incluso idénticos, podría pensarse que esta clasificación en dos grupos que tenemos dibujada en el gráfico pudiera dejar insatisfechos a aquéllos que advirtiesen,

desde su punto de vista, que entre los individuos que se han incluido en ambos grupos haya demasiadas diferencias y pensaran, en consecuencia, que deberíamos buscar un "número de grupos" mayor, de forma que fueran "grupos más homogéneos"; por ejemplo, como los que se reflejan en los siguientes gráficos.



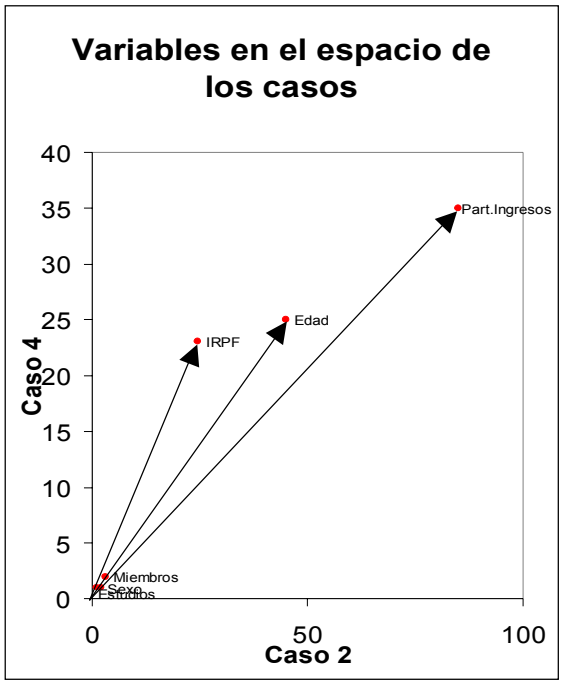
Observemos que, de forma intuitiva pero en definitiva, estamos definiendo una estructura de grupos de casos que pivota sobre los parecidos o diferencias observados entre los individuos (a las que nos referiremos como *proximidad entre los casos*) y del grado de diferencia que estemos dispuestos a admitir entre cada uno de los puntos que se integran en el mismo grupo, lo que nos conducirá a una configuración final de un determinado número de grupos considerados diferentes.

Este será el fundamento intuitivo de la clasificación en grupos que abordaremos en el capítulo dedicado al Análisis de Conglomerados: basándonos en las *proximidades* que presenten los casos entre sí, estructuraremos el espacio en grupos, tanto más homogéneos internamente cuanto mayor sea el número de grupos considerados o, recíprocamente, tanto más heterogéneos internamente cuanto menor sea el número de grupos que construyamos. Cuantos menos grupos consideremos, más heterogéneos serán internamente porque estaremos admitiendo mayores distancias entre los individuos pertenecientes al mismo grupo. Cuanto más grupos permitamos, más homogéneos serán los grupos internamente porque estaremos reduciendo más las distancias admitidas entre los individuos del mismo grupo.

Recordemos, además, que los datos multivariantes admiten ser vistos desde una segunda perspectiva: en el espacio de los casos, donde los elementos representados serían las variables observadas.

Cuando estemos trabajando con variables representadas en el espacio de los casos, los grupos serán conjuntos de variables similares o semejantes en algún sentido y, generalmente, las medidas de asociación y correlación van ser útiles para medir la "proximidad" entre dichas las variables y ayudarnos para decidir si son semejantes o no.

Consideremos a modo de ejemplo la representación siguiente, en el espacio de los dos casos (Caso 2, Caso 4), de las variables: Edad, sexo, IRPF, Miembros del hogar, Nivel de estudios y participación de ingresos (se ha obviado la variable Ingresos para visualizar mejor la casuística, ya que la gran diferencia de escalas utilizadas en las variables no nos permitiría ver un cierto detalle para todas ellas).



Nuevamente, insistimos en que es un mero ejemplo y que la visión dada por este gráfico es muy parcial, ya que al estar considerando 12 casos, la representación de la nube de puntos tiene lugar en un espacio de dimensión 12. Aquí, solamente estamos observando el subespacio generado por dos de ellos, en el que podemos proyectar dichas variables, que usualmente se representan como vectores, característicos de cada una de ellas, con origen en el origen de coordenadas y afijo en cada uno de esos puntos.

Desde este enfoque, dos puntos (afijos) muy cercanos mostrarían variables que presentan valores muy parecidos para todos los casos, llegando a ser idénticas en el caso en que los puntos coincidieran perfectamente. Para este concepto de proximidad (relación del tipo Y=X), la distancia euclídea podría seguir siendo válida. Sin embargo, la proximidad entre las variables medida en términos de una dependencia lineal de tipo más general puede evaluarse intuitivamente a partir del grado de correlación que presentan las variables comparadas (correlación lineal del tipo Y=aX+b), o de la existencia de proporcionalidad entre dichos vectores (correlación lineal del tipo Y=aX).

El análisis de conglomerados podrá ser enfocado pues, en esos dos sentidos: en el espacio de las variables (donde la nube de puntos representa a los casos) para realizar grupos de casos que presenten comportamientos similares para el conjunto de variables, o bien en el espacio de los casos (donde la nube de puntos representa a las variables) para realizar grupos de variables que presenten comportamientos similares para el conjunto de los casos. Y para medir el grado de parecido o semejanza entre los objetos (casos o variables) clasificados en ambas situaciones emplearemos las llamadas medidas de proximidad o simplemente proximidades, cuyo estudio vamos a abordar en este capítulo.

El Concepto de Proximidad: distancias, disimilaridades y similaridades.

Llamamos proximidades a ciertos instrumentos matemáticos que pretenden medir el grado de semejanza que presentan dos objetos cualesquiera, que denotamos *i* y *j,* y que, como ya sabemos, pueden ser casos o variables según el espacio en el que estemos trabajando con los datos.

Este grado de semejanza que pretendemos medir, no requerirá normalmente considerar todas las variables observadas, sino que muchas veces bastará con considerar un cierto número de ellas. Por ejemplo, si a partir de los datos del apartado anterior pretendemos ver cómo de parecidas son las familias (casos) observados en relación a su composición demográfica, bastaría con considerar las variables número de miembros, edad y sexo.

Por tanto, estas proximidades o medidas de proximidad medirán, para esas características que estemos considerando, las semejanzas existentes entre los datos (casos o variables) y, en general, utilizaremos como instrumentos matemáticos que nos permitirán alcanzar este objetivo los llamamos *medidas de distancia y disimilaridades*, que en general denotaremos con la letra *d*, y donde, por tanto, $d(i,j)$ significará la distancia o disimilaridad existente entre los elementos *i* y *j*.

La idea que subyace en la definición de estos conceptos, trata de replicar la que intuitivamente tenemos acerca de las distancias físicas: cuanto más separados están los individuos, cuanto mayor es la distancia entre ellos, intuitivamente interpretamos que menos se parecen. Así, a mayor semejanza de los individuos, menos distancia o disimilaridad debe haber. A más distancia, menos semejanza entre los individuos y más diferencia.

Distancias y disimilaridades intuitivamente se interpretan de forma similar, y su característica fundamental es que aumentan a medida que decrece la semejanza. La diferencia existente entre ellas, como vamos a ver, es puramente teórica y se deriva del conjunto de propiedades deseables que son capaces de cumplir en el espacio que estén siendo utilizadas.

Medida de distancia o métrica

La medida de proximidad más extendida por su conocimiento general en el ámbito de la geometría básica y que todos conocemos desde la escuela, es la Distancia Euclídea que se define como la raíz cuadrada de la suma de los cuadrados de las diferencias de las coordenadas de los puntos considerados. Sin embargo, ésta nos es más que un caso particular de Medida de Distancia o Métrica, y que vemos a continuación.

Una **medida de distancia o métrica**, es una función real que a cada par de objetos (i, j), casos o variables, les asocia un número real positivo o nulo:

$$d(i, j) \geq 0 \ , \ \forall i, j$$

verificando, además, las siguientes propiedades:

a) la única posibilidad para que la distancia entre dos elementos comparados, *i* y *j*, sea exactamente cero, es que los dos elementos comparados sean realmente el mismo

$$d(i,j)=0 \quad \Leftrightarrow \quad i=j \quad \forall i,j$$

y por tanto, dos individuos que no sean iguales, tienen alguna distancia no nula.

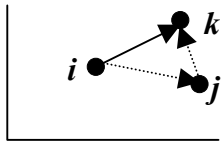
b) La distancia es simétrica, es decir, da lo mismo medir la distancia desde el elemento *i* al *j*, que desde el *j* hasta el *i*; las distancias medirán lo mismo se midan en el sentido que se midan:

$$d(i,j)=d(j,i) \quad , \quad \forall i,j$$

c) Y finalmente, la distancia debe verificar la llamada propiedad triangular que nos viene a decir que, dados 3 objetos *i, j, k*, la distancia entre dos puntos (por ejemplo *i* y *j*) es siempre más corta que la suma de la distancia del primero al tercero *d(i,k)* más la distancia del tercero al segundo *d(k,j)*.

$$d(i,j)\le d(i,k)+d(k,j) \quad , \quad \forall i,j,k$$

Esta propiedad nos permitía decir, en el espacio euclídeo, que la distancia más corta entre dos puntos era la longitud del segmento que los unía en línea recta; lo que gráficamente puede ser representado de la siguiente forma:



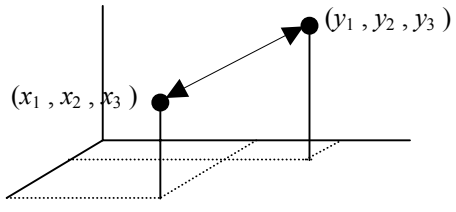
Como consecuencia de estas propiedades anteriores, una distancia o una métrica verifica que la distancia de un individuo consigo mismo es nula:

$$d(i,i)=0 \quad , \quad \forall i$$

Como decíamos, la distancia más conocida y utilizada es la distancia euclídea, que nos conduce a calcular la distancia entre dos puntos cualesquiera del espacio $\Re^m$, con coordenadas $(x_1 , x_2 , \ldots , x_m)$ e $(y_1 , y_2 , \ldots , y_m)$, como la raíz cuadrada de la suma de los cuadrados de las diferencias de sus respectivas componentes:

$$d_2((x_1,x_2,..,x_m),(y_1,y_2,...,y_m))=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_m-y_m)^2}$$

Intuitivamente, esta cantidad representa la distancia que hay en línea recta entre los puntos $(x_1 , x_2 , \ldots , x_m)$ e $(y_1 , y_2 , \ldots , y_m)$, lo que representamos gráficamente, a modo de ilustración, en el caso estar trabajando en el espacio de tres dimensiones.



Sin embargo, no todas las situaciones son representables en espacios euclídeos del tipo $\Re^m$. Pensemos en los individuos que todos los días se desplazan en coche desde una localidad de la corona metropolitana de una gran ciudad a ésta, y definamos la proximidad entre ellos en términos de la diferencia del tiempo que tardan en realizar sus recorridos. En este caso podríamos encontrar individuos diferentes que tardasen el mismo tiempo en sus desplazamientos, presentando una proximidad máxima (diferencia cero) sin ser necesariamente el mismo individuo. O pensemos que nos interesa definir la distancia entre dos localidades, en una franja horaria determinada, en términos del tiempo empleado para desplazarse entre las mismas, de forma que dos puntos (localidades) serían próximos si se emplease poco tiempo para desplazarse entre ellos. En este caso, la definición de proximidad empleada no es ni siquiera simétrica; lo que comprobaríamos fácilmente si pensamos en la posible diferencia de fluidez de tráfico entre las dos localidades en ambos sentidos (por ejemplo, en una autovía que une una localidad de la corona metropolitana con el centro de la metrópolis) y que provoca que los tiempos empleados en una u otra dirección puedan ser muy diferentes.

Así pues, debemos relajar el concepto de distancia al mínimo conceptual exigible por el concepto de proximidad, lo que nos conduce a la definición de disimilaridad.

Disimilaridades

Una **Disimilaridad**, es una función real no negativa que mide la diferencia entre dos elementos (i , j), casos o variables, de forma que les asocia un número real positivo o nulo:

$$d(i,j) \geq 0 \ , \ \forall i,j$$

de forma que dos elementos serán tanto más dispares cuanto mayor sea su disimilaridad, y a la que se le exige además que la disimilaridad de un punto consigo mismo sea cero:

$$d(i,i) = 0 \ \ , \ \forall i$$

con lo que podría ocurrir que haya pares de elementos cuya medida de disimilaridad sea cero sin tener que ser precisamente los mismos; lo que sí se exigía a las distancias.

Por tanto la definición que acabamos de dar inicialmente para el concepto de Disimilaridad, es una definición mínima basada en la idea intuitiva de medir la proximidad o similitud entre objetos, en un sentido parecido a como lo hacen las distancias.

Sin embargo, sin ser estrictamente necesario aunque constatable en multitud de situaciones prácticas, con el objeto de facilitar la operatividad de los modelos matemáticos donde se emplea éstas, comúnmente exigimos que la Disimilaridad cumpla además la propiedad de simetría.

$$d(i,j) = d(j,i) \ \ , \ \forall i,j$$

En este capítulo, nos referiremos a las proximidades acordes con esta definición simplemente como **Disimilaridad**, si bien deberíamos referirnos más estrictamente a ellas como **Disimilaridad Simétrica.**

La ventaja que supone el poder definir estas disimilaridades en espacios más generales que los euclídeos, trae emparejada la necesidad de aprender a interpretar en estos nuevos espacios las posiciones relativas de los elementos comparados. Sin embargo, a ninguno se nos escapa la conveniencia de trabajar sobre un espacio euclídeo, en el que sabemos movernos con mucha más soltura.

Tratando de recuperar la capacidad de representación e interpretación que nos ofrece el espacio euclídeo, tratamos de descubrir qué espacios sobre los que se han definido una cierta disimilaridad pueden proyectarse de alguna forma en el espacio euclídeo, permitiéndonos, en consecuencia, seguir trabajando con aquél pero con las comodidades que éste ofrece.

Disimilaridad Euclidizable

Se define **Disimilaridad Euclidizable** como aquella Disimilaridad Simétrica que permite ser puesta en correspondencia con una distancia euclídea sobre un cierto espacio euclídeo, de la siguiente forma:

$$\forall i, \exists I = (x_{i1}, x_{i2}, ..., x_{im}) \in \Re^m \mid d(i,j) = d_2(I,J) = \sqrt{\sum_{h=1}^{m} (x_{ih} - x_{jh})^2}$$

Así pues, una disimilaridad es euclidizable cuando, de alguna manera podemos proyectar los objetos de ese espacio sobre puntos de un espacio euclídeo, asignando a cada objeto unas coordenadas en éste, de forma que se pueden reproducir las disimilaridades entre los objetos de aquél a partir de las distancias euclídeas calculadas con las coordenadas de los correspondientes puntos proyecciones de aquéllos sobre este espacio euclídeo. Es decir, para cualquier elemento del espacio inicial, existe un punto determinado en un cierto espacio $\Re^m$, de tal manera que la distancia euclídea entre cada dos de esos puntos proyectados en el espacio euclideo coincide con la disimilaridad observada entre los objetos de los que provienen. Este tipo de disimilaridad la encontraremos, por ejemplo, cuando abordemos el Análisis Factorial de Correspondencias de Benzecri en un capítulo posterior.

Como consecuencia de esta correspondencia entre la disimilaridad euclidizable en el espacio original y la distancia euclídea en el espacio proyectado, se llega a la conclusión de que ***cualquier Disimilaridad Euclidizables es una Distancia,*** ya que la disimilaridad cumplirá las mismas propiedades adicionales que la distancia euclídea, y por tanto de las distancias en general:

$d(i,j) = 0 \Leftrightarrow i=j \ , \ \forall \ i,j (\Rightarrow d(i,i) = 0 \ , \ \forall \ i)$
$d(i,j) \leq d(i,k) + d(k,j) \ , \ \forall \ i,j,k$ (Propiedad Triangular)

Disimilaridad Ultramétrica

Las necesidades teóricas que impone la construcción de un algoritmo de clasificación jerárquica perfectamente definido, sin ambigüedades, obliga a definir un tipo especial de disimilaridad que llamaremos **disimilaridad ultramétrica.** Se define como una función real que a cada par de elementos (i , j), casos o variables, les asocia un número real positivo o nulo:

$$d(i,j) \geq 0 \ , \ \forall i,j$$

verificando, además, las siguientes propiedades:

a) La disimilaridad de un individuo *i* consigo mismo siempre vale cero.

$$d(i,i) = 0 \qquad \forall i$$

y por tanto, sólo individuos no iguales, pueden presentar disimilaridad no nula.

b) La distancia es simétrica, es decir, da lo mismo medir la distancia desde el elemento *i* al *j*, que desde el *j* hasta el *i*; las distancias medirán lo mismo se midan en el sentido que se midan:

$$d(i,j) = d(j,i) \ \ , \ \ \forall i,j$$

c) Y finalmente, la distancia debe verificar la *propiedad ultramétrica* que nos viene a decir que, dados 3 individuos, *i, j, k*, la disimilaridad entre dos individuos (*i*, *j*) siempre es menor o igual que el máximo de las disimilaridades entre cada uno de esos individuos y un tercero sea cual sea ese tercer individuo (*k*).

$$d(i,j) \leq max_k\,(d(i,k)\,,d(k,j)) \ , \ \forall i,j,k \quad \text{(Propiedad Ultramétrica)}$$

A la vista de la complejidad de esta última propiedad, podemos preguntarnos si existen realmente estas disimilaridades ultramétricas, ya que no parece fácil encontrar situaciones en las que se puedan presentar. Pese a esta lógica duda, la respuesta es sí, para lo que veamos el siguiente ejemplo.

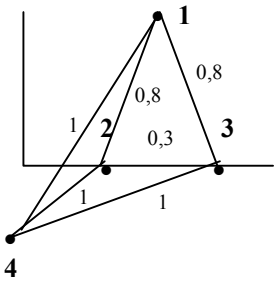
Consideramos cuatro elementos (1,2,3,4) y sus disimilaridades entre cada dos de ellos recogidas en la siguiente tabla.

Disimilaridades	1	2	3	4
1	0	0,8	0,8	1
2	0,8	0	0,3	1
3	0,8	0,3	0	1
4	1	1	1	0

Puede comprobarse que la medida de disimilaridad dada en este ejemplo es realmente una ultramétrica.

Obviamente las disimilaridades son no negativas, nulas cuando se compara un elemento consigo mismo y si comparamos dos elementos cualesquiera, siempre existe un tercero cuya disimilaridad con alguno de los dos anteriores es al menos igual, cuando no mayor, que la disimilaridad presentada entre aquéllos.

Examinemos un poco la estructura de los datos que tenemos en esta tabla, y representemos los cuatro puntos en un espacio de dimensión 3, considerando las disimilaridades como distancias entre los mismos.



Si observamos cuidadosamente, veremos que todos los triángulos formados por cualesquiera tres puntos que elijamos son isósceles.

No es difícil de demostrar analíticamente que, si se cumple la propiedad ultramétrica, esto sigue siendo cierto para un caso general. Así, siendo de complicada interpretación a primera vista, el cumplimiento de la propiedad ultramétrica implica que la disposición de los puntos en el espacio, considerando las disimilaridades como distancias, es tal que cada tres puntos cualesquiera forman un triángulo de tipo isósceles, es decir, con al menos dos lados iguales.

Así que los espacios en los que pueden encontrarse disimilaridades ultramétricas son realmente espacios un tanto "especiales o raros", por lo que pocas veces se dan en la práctica diaria. Y entonces, ¿Qué necesidad hay de considerarlos como algo especial?

Es prematuro entrar en detalles teóricos que justifiquen perfectamente su necesidad, pero como anticipo, digamos que el que las disimilaridades definidas entre los elementos sean ultramétricas será justamente la condición teórica necesaria para generar un algoritmo fundamental para la obtención de clasificaciones jerárquicas indexadas que podamos representar mediante un único esquema de clasificación (dendrograma) sin ambigüedades, como veremos en el capítulo dedicado al Análisis de Conglomerados.

Observemos también que la propiedad ultramétrica es una propiedad bastante más restrictiva que la propiedad triangular que cumplían las distancias. De hecho, ***la propiedad ultramétrica implica, como consecuencia, la propiedad triangular***:

$$\left\{d(i,j) \le max_k\left(d(i,k)\,,d(k,j)\right)\ ,\ \forall i,j,k\right\}\ \Rightarrow\ \left\{d(i,j) \le d(i,k)+d(k,j))\ ,\ \forall i,j,k\right\}$$

Así que, si no fuera por que las disimilaridades admiten valores nulos para elementos diferentes, cosa imposible en las distancias, la disimilaridad ultramétrica sería una distancia métrica. Para obviar este problema, se define **distancia ultramétrica**, o simplemente **ultramétrica**, como una función

$$d : E \times E \to \Re^+$$

que verifica

$$d(i,j) \ge 0\ \ ,\ \ \forall i,j$$
$$d(i,j) = 0 \Leftrightarrow i = j\ \ ,\ \ \forall i,j$$
$$d(i,j) = d(j,i)\ \ ,\ \ \forall i,j$$
$$d(i,j) \le max_k\left(d(i,k)\,,d(k,j)\right)\ ,\ \forall i,j,k$$

en cuyo caso, al ser la propiedad ultramétrica más exigente que la triangular, podemos decir que toda distancia ultramétrica es efectivamente una medida de distancia.

Similaridades

Conviene observar que todos los distintos tipos de proximidad definidos hasta ahora — distancias, disimilaridades, disimilaridades simétricas, disimilaridades euclidizables, disimilaridades ultramétricas, ultramétricas — miden la semejanza o diferencia entre los elementos estudiados en el mismo sentido: mayores valores de la medida, expresan menor semejanza entre los elementos comparados; menores valores de la medida, significan mayor semejanza observaremos entre los elementos comparados.

Alternativamente, podríamos medir el parecido de los elementos comparados en el sentido contrario. Esta forma alternativa de medir semejanza es habitual en ciertas herramientas estadísticas básicas como las medidas de correlación y de asociación y que pueden medir proximidad entre variables. Así, decir que dos variables están muy asociadas o muy correlacionadas, supone admitir que poseen mucha información en común, por lo que pueden ser consideradas en este sentido muy semejantes. Y es cuando estas medidas toman valores absolutos altos, cuando decimos que las variables son próximas o semejantes. Justamente estamos midiendo la proximidad en el sentido contrario al que lo hacen las disimilaridades; pero, obviamente estamos midiendo proximidad entre variables.

Así pues, de forma paralela a como se hizo para las disimilaridades, se pueden definir las **similaridades** para medir cercanía o similitud entre los elementos comparados, pero aumentando su valor al crecer la semejanza entre estos: cuánto más semejantes sean los elementos o individuos, mayor valor presentará la correspondiente medida de similaridad. Cuánto menos semejantes sean, menos valor presentarán las medidas. Responden, por tanto, al concepto intuitivo, al significado de similitud: mayor valor, más similitud; menor valor, menos similitud.

Sin embargo, la naturaleza de esta medida exige que estén acotadas por un valor máximo que represente la máxima semejanza entre dos individuos y que lógicamente debe darse, al menos, para el caso en que comparamos un elemento consigo mismo. Lógicamente no puede haber un valor de similitud o semejanza (similaridad) mayor que el que se aprecie en este caso. Por tanto, las medidas de similaridad siempre están acotadas y normalmente se presentan estandarizan entre 0 y 1.

Así, la definición más general de **similaridad** es una función real que mide la semejanza entre dos elementos (*i,j*), casos o variables, de forma que les asocia un número real

$$\delta(i,j) \leq M \ , \ \forall \text{i, j} \quad \text{(M máximo valor)}$$

y de forma que dos elementos serán tanto más semejantes cuanto mayor sea su similaridad, y a la que se le exige además que la similaridad de un punto consigo mismo sea máxima:

$$\delta(i,i) = M \ , \ \forall \text{i} \quad \text{(M máximo valor)}$$

con lo que podría ocurrir que haya pares de elementos cuya medida de similaridad sea máxima también sin tener que ser precisamente idénticos.

Para que la fuese una **similaridad simétrica**, además debería cumplir la siguiente propiedad:

$$\delta(i,j) = \delta(j,i) \ , \ \forall \text{i, j} \qquad \text{(simetría)}$$

En cualquier caso, si tenemos una medida de disimilaridad y es *M* un valor real tal que $0 \le d(i,j) \le M$, (*M* cota superior que puede conseguirse generalmente en investigaciones socioeconómicas ya que en ellas trabajaremos sobre un número finito de casos), *M* representando la máxima semejanza entre los individuos comparados, entonces podemos construir una medida de similaridad a partir de ésta, y que podemos llamar $\delta(i,j)$, simplemente restando de ese valor o cota superior *M*, la medida de disimilaridad:

$$\delta(i,j) = M - d(i,j)$$

cumpliendo

$$\delta(i,i) = M \ , \ \forall \text{i} \quad \text{(M máximo valor)}$$

Con lo cual, si la disimilaridad *d* daba valores grandes para puntos poco semejantes, entonces, δ dará valores pequeños para puntos poco semejantes. Y si *d* daba valores pequeños para puntos muy semejantes, δ dará valores grandes para puntos semejantes. Lo único que estamos haciendo es invertir el sentido de la medida y, de esta forma, construir una medida de similaridad a partir de una de disimilaridad.

Análogamente, de forma inversa, si tenemos una similaridad acotada por un valor M, $\delta(i,j) \le M$, automáticamente podríamos construir una medida de disimilaridad, que podemos llamar $d(i,j)$, simplemente restando de ese máximo, de esa cota superior *M*, la medida de similaridad:

$$d(i,j) = M - \delta(i,j) \ge 0$$

cumpliendo

$$d(i,i) = 0 \qquad \forall i$$

Principales medidas de proximidad.

En este apartado vamos a examinar las más comunes medidas de disimilaridad, distancia y similaridad utilizadas, en función de la naturaleza de las situaciones que se nos planteen, para medir en la práctica la proximidad o semejanza entre los individuos o elementos comparados.

Dada la doble posibilidad de representación de los datos en los dos espacios consabidos, todas las medidas que vamos a ver podrían ser utilizadas tanto en el espacio de los casos, como en el de las variables. Sin embargo, suelen utilizarse preferentemente en uno de ellos; en el que presentan una interpretación o significado más claro. Por ello, indicaremos entre paréntesis la situación en la que son más comúnmente utilizadas, sin que perjuicio para que pueda ser utilizado en el otro cuando convenga.

Medidas de Disimilaridad para escalas de Intervalo

Son probablemente las más conocidas por ser las escalas de intervalo y de razón las que ha permitido históricamente un mayor desarrollo cuantitativo de la ciencia en general.

Para comenzar, observemos que el valor absoluto de la tipificación de un valor puede interpretarse como una medida de disimilaridad del elemento considerado y el centroide o elemento promedio del conjunto, $d(i,\bar{x})$, cuando trabajemos con una sola variable. Esto induce lógicamente una medida de disimilaridad similar para comparar dos elementos, $d(i,j)$, en la misma situación, como se describe a continuación.

Valores Tipificados: (suele usarse para Casos)

$$d(i,\bar{x}) = \frac{|x_i - \bar{x}|}{S} \qquad d(i,j) = \frac{|x_i - x_j|}{S}$$

En cualquier caso, estamos especialmente interesados en las situaciones multivariantes, por lo que presentamos a continuación la más utilizada, sin duda, de las medidas de disimilaridad en esta situación, la distancia euclídea, cuya interpretación geométrica ya fue vista en el apartado anterior para el caso de tres dimensiones.

Distancia Euclídea: (suele usarse para Casos)

$$d_2(i,j) = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2}$$

Sin embargo, podemos definir una distancia más general que contiene como caso particular a la distancia euclídea: la llamada distancia de Minkowski.

Distancia de Minkowski: (suele usarse para Casos)

$$d_m(i,j) = \left(\sum_{h=1}^{p} |x_{ih} - x_{jh}|^m \right)^{1/m}$$

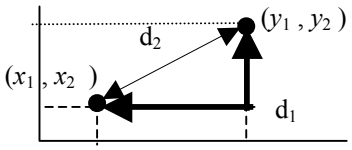
Es evidente que cuando m=2, la distancia de Minkowski coincide exactamente con la distancia Euclídea, teniendo por tanto su misma interpretación. Para m=1 y $m \to \infty$, se obtienen los casos particulares de las distancias llamadas de "city-block" o de "Manhattan" y de "Chebychev" o del "Máximo", respectivamente.

Distancia "City-Block" o de "Manhattan" : (suele usarse para Casos)

$$d_1(i,j) = \sum_{h=1}^{p} |x_{ih} - x_{jh}|$$

Es el caso particular de la distancia de Minkowski cuando m=1, su interpretación geométrica nos lleva a considerar la distancia entre dos puntos como la longitud del

camino que lleva de un punto a otro moviéndonos siempre paralelamente a los ejes. Gráficamente, en el siguiente gráfico, podemos comparar la distancia euclídea, d_2 (línea continua fina) con la distancia City Block, d_1 (línea continua más gruesa), para el caso de dos dimensiones.

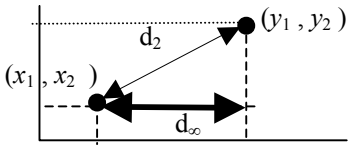


Obsérvese que la distancia de City- Block reproduce el camino que realizamos para desplazarnos entre dos puntos de una ciudad en la que las calles son paralelas que transcurren entre los bloques de edificios. Obviamente en tal situación, para medir la proximidad física entre dos puntos para un peatón, la distancia euclídea no sería adecuada ya que lógicamente el peatón no podría atravesar paredes para ir por el camino teóricamente más corto, sino que deberá andar por las calles realizando el recorrido que mide la distancia de City-Block.

Distancia de Tchevichev: (suele usarse para Casos)

$$d_{\infty}(i,j) = \max_{k=1\ldots p} \left| x_{ih} - x_{jh} \right|$$

Es el caso particular de la distancia de Minkowski cuando $m \to \infty$, y cuya interpretación geométrica nos lleva a considerar la distancia entre dos puntos como la separación máxima que presentan las proyecciones de los dos puntos sobre los ejes del espacio. Gráficamente, en el siguiente gráfico, podemos comparar la distancia euclídea, d_2 (línea continua fina) con la distancia deTchevichev, d_∞ (línea continua más gruesa), para el caso de dos dimensiones.



Esta distancia, por tanto, nos da una idea de cómo se parece o difiere la característica medida en la dimensión en que más se diferencian los dos elementos comparados.

Podemos aún definir una distancia más general que contiene como caso particular a la distancia euclídea e incluso a la distancia de Minkowski:

Distancia de Minkowski Generalizada: (suele usarse para Casos)

$$d_{m,q}(i,j) = \left(\sum_{h=1}^{p} \left| x_{ih} - x_{jh} \right|^{m} \right)^{1/q}$$

Es evidente que cuando m=q, la distancia coincide con la de Minkowski.

Cuadrado de la Distancia Euclídea: (suele usarse para Casos)

Obviamente, los cuadrados de todas estas medidas son también medidas de disimilaridad (no necesariamente distancias) y, en particular, suele utilizarse a veces como medida el cuadrado de la distancia euclídea, por varios motivos como son evitar la raiz cuadrada haciéndola operativamente más sencilla, su relación con el coeficiente de correlación lineal cuando se utiliza para medir disimilaridad entre variables, y su relación con la D^2 de Mahalanobis, como veremos posteriormente.

$$d_2^2(i,j) = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2$$

D^2 de Mahalanobis entre 2 individuos: (suele usarse para Casos)

Todas estas medidas, como hemos visto, consideran igualmente importantes las diferencias apreciadas en cada una de las variables; lo cual parece lógico si las variables son incorrelacionadas y se mueven en rangos de valores similares. Sin embargo, la realidad nos dice que suele existir un grado de correlación más o menos importante entre ellas.

Supongamos 3 individuos cuyos pesos y estaturas fuesen 1,70cm y 70kg para el individuo A, 1,80cm y 80kg para el individuo B, y 1,60cm y 80kg para el individuo C. Comparados el individuo A con el B y A con el C, ambas parejas presentan la misma distancia euclidea ya que difieren en 10 cmts de estatura y 10 kgs de peso. Sin embargo, podemos estar de acuerdo que los individuos A y B son más parecidos que los A y C ya que la diferencia de peso se debe a la diferencia de estatura manteniendo ambos individuos una constitución corporal similar, mientras que no ocurre esto al comparar A y C, en cuyo caso la diferencia de peso se debe a que C está bastante más grueso que A modificando su constitución a más obeso..

Para corregir este efecto, podemos utilizar la distancia D^2 de Mahalanobis, que tiene en consideración, como factor de corrección, la matriz $\sum$ de varianzas y covarianzas de los datos, y que al ser realmente una "distancia", cumplirá todas las propiedades de ésta.

$$D^2(i,j) = (x_i - x_j)' \Sigma^{-1} (x_i - x_j)$$

Además de ser una alternativa para resolver el problema de la posible correlación entre las variables, la distancia D^2 de Mahalanobis así definida es una generalización de la distancia euclídea. De hecho, cuando las variables sean independientes y tipificadas, entonces el resultado de D^2 será justamente el mismo que el de d_2 al cuadrado.

Efectivamente, si las variables fueran independientes, entonces sus covarianzas serían nulas; y si estuvieran tipificadas, entonces sus varianzas serían todas iguales a la unidad. En este caso, la matriz $\sum$ de varianzas y covarianzas quedaría reducida a la matriz identidad, y la distancia D^2 de Mahalanobis, al producto escalar del vector diferencia de los dos elementos comparados, por sí mismo, que no sería más que el cuadrado de la distancia euclídea entre los dos puntos.

$$D^2(i,j) = (x_i - x_j)' \Sigma^{-1} (x_i - x_j) = (x_i - x_j)' \cdot I \cdot (x_i - x_j) = (x_i - x_j)' \cdot (x_i - x_j) = d_2^2(i,j)$$

siendo éste el motivo por el que la distancia de Mahalanobis se denota por D^2 , en recuerdo de que es una generalización del cuadrado de la distancia euclídea.

Esta distancia, en ésta, su versión básica, se aplica para comparar dos individuos sin más que considerar para ello las coordenadas de esos dos individuos. Sin embargo, si en vez de tomar las coordenadas de un elemento cualquiera, tomamos las del centroide de un conjunto de elementos (grupo), podremos aplicarla para medir la proximidad de un individuo al centro de su grupo, o incluso la proximidad entre los centros de dos grupos.

D^2 de Mahalanobis de un individuo al centroide de un grupo: (suele usarse para Casos)

$$d(i,\bar{x}) = (x_i - \bar{x})' \Sigma^{-1} (x_i - \bar{x})$$

D^2 de Mahalanobis entre los centroides de 2 grupos: (suele usarse para Casos)

$$d(\bar{x}_i,\bar{x}_j) = (\bar{x}_i - \bar{x}_j)' \Sigma^{-1} (\bar{x}_i - \bar{x}_j)$$

A modo de recordatorio, y para fijar notaciones, puede verse en el anexo a este tema cómo podemos calcular esta matriz Σ de varianzas y covarianzas de los datos.

Medidas de Similaridad para escalas de Intervalo

En este caso, por (X_i, X_j) representaremos las dos variables comparadas, representadas en el espacio de los casos, y que por tanto tendrán coordenadas $X_i \equiv (x_{1i}, x_{2i}, ..., x_{ni})$ y $X_j \equiv (x_{1j}, x_{2j}, ..., x_{nj})$ respectivamente.

Las siguientes medidas expuestas no son otras que el conocido coeficiente de correlación de Pearson y el valor del coseno del ángulo que forman los dos vectores de las variables consideradas.

Coeficiente de Correlación de Pearson: (suele usarse para Variables)

$$d(X_i, X_j) = r = \frac{1}{n}\sum_{h=1}^{n} z_{hi} \cdot z_{hj} \quad ; \quad siendo \quad z_{hk} = \frac{x_{hk} - \bar{x}_k}{S_k},$$

Es decir, z_{hi} y z_{hj} son aquí los valores tipificados que presentó el caso h para las variables i y j respectivamente; y recordemos que la covarianza de variables tipificadas coincide con el coeficiente de correlación de las variables, de donde deducimos que, efectivamente, la fórmula expuesta es, como hemos titulado, el coeficiente de correlación de Pearson de las variables (X_i, X_j).

Para utilizar esta medida como medida de similitud entre las variables comparadas, la similitud debe entenderse en el sentido que marca la correlación entre las variables: las variables serán tanto más parecidas cuanto mayor sea el coeficiente de correlación de Pearson que presenten; es decir, cuanto más información común comparta cada una acerca de la otra, siempre que las dos varíen en el mismo sentido.

Puede demostrarse que cuando se utiliza el cuadrado de la distancia euclídea como medida de la disimilaridad entre variables tipificadas, (por ejemplo, X e Y con un coeficiente de correlación de Pearson r_{xy}), la relación es la siguiente:

$$d_2^2(X,Y) = 2 \cdot n \cdot (1 - r_{xy})$$

Coeficiente de Determinación: (suele usarse para Variables)

$$d(X_i, X_j) = R^2 = r^2 = \left(\frac{1}{n} \sum_{h=1}^{n} z_{hi} \cdot z_{hj} \right)^2,$$

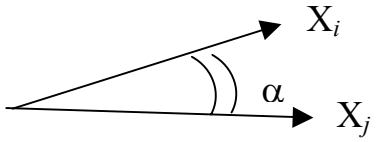
Obsérvese que aparentemente esta medida y la anterior son muy parecidas. De hecho, ambas son medidas de similaridad. Pero en este caso la definición de similitud es bastante diferente. Aquí las variables serán tanto más parecidas cuanto mayor sea el coeficiente de Determinación, es decir cuanto más se acerque este coeficiente a 1; o lo que es igual, cuanto más información común comparta cada una acerca de la otra, varíen las dos variables en el sentido en que lo hagan.

Así, dos variables que presenten un coeficiente de correlación -1, serán muy semejantes según la medida del coeficiente de Determinación (ya que R^2=+1), mientras que serán muy distintos para la medida del Coeficiente de correlación de Pearson ya que tomaría el menor valor posible (r=-1).

Coseno del ángulo formado por las variables: (suele usarse para Variables)

$$d(X_i, X_j) = \cos(\alpha) = \frac{\sum_{h=1}^{n} x_{hi} \cdot x_{hj}}{\sqrt{\sum_{h=1}^{n} x_{hi}^2 \cdot \sum_{h=1}^{n} x_{hj}^2}}$$

siendo α el ángulo formado por los vectores representantes de las variables comparadas, X_i y X_j , en el espacio de los casos.



Cuando estas variables sean muy parecidas, el ángulo α tendería a cero y, por tanto, su coseno sería tendente a +1; mientras que se considerarán máximamente diferentes cuando el coseno valga –1; es decir, cuando sean diametralmente opuestas.

Es, por tanto, una medida muy relacionada con la proporcionalidad de las variables. Cuando el coseno valga 1 significará que las variables se sitúan sobre una misma dirección trazada desde el origen de coordenadas y en el mismo cuadrante. Cuando el coseno valga –1, significará que las variables son diametralmente opuestas y se sitúan sobre una misma dirección trazada desde el origen de coordenadas, pero en cuadrantes

opuestos. Obsérvese, además, que para datos centrados, coincide con el coeficiente de correlación lineal, r.

Cuadrado del Coseno del ángulo formado por las variables: (suele usarse para Variables)

Similarmente a como hemos definido e interpretado el coeficiente de determinación con relación al coeficiente de correlación, podemos definir e interpretar este Cuadrado del Coseno del ángulo formado por las variables con el que prescindimos del sentido de la relación de proporcionalidad, quedándonos sólo con la dirección de la misma.

$$d(X_i, X_j) = \cos^2(\alpha) = \frac{\left(\sum_{h=1}^{n} x_{hi} \cdot x_{hj}\right)^2}{\sum_{h=1}^{n} x_{hi}^2 \cdot \sum_{h=1}^{n} x_{hj}^2}$$

Obsérvese, además, que para datos centrados, coincide con el coeficiente de determinación, R^2.

Tanto el coeficiente de correlación de Pearson, como el coeficiente de determinación, como el coseno del ángulo formado por las variables, como de forma análoga su cuadrado, son medidas de similaridad (no de disimilaridad), aumentando su valor con el parecido de los individuos que estamos comparando (en este caso, las variables).

Si quisiéramos obtener, a partir de ellas, una medida de disimilaridad, bastaría con observar que 1 es una cota superior para ambas medidas, y aplicar la propiedad estudiada en el apartado anterior a tal efecto. Así, serían medidas de disimilaridad:

$$\text{1-Coeficiente de Correlación} = 1 - \frac{1}{n}\sum_{h=1}^{n} z_{hi} z_{hj}$$

o

$$\text{1-Coeficiente de Determinación} = 1 - \left(\frac{1}{n}\sum_{h=1}^{n} z_{hi} z_{hj}\right)^2$$

o

$$\text{1-Cos(ángulo de las variables)} = d(X_i, X_j) = 1 - \frac{\sum_{h=1}^{n} x_{hi} \cdot x_{hj}}{\sqrt{\sum_{h=1}^{n} x_{hi}^2 \cdot \sum_{h=1}^{n} x_{hj}^2}}$$

o

$$\text{1-Cos}^2\text{(ángulo de las variables)} = d(X_i, X_j) = 1 - \frac{\left(\sum_{h=1}^{n} x_{hi} \cdot x_{hj}\right)^2}{\sum_{h=1}^{n} x_{hi}^2 \cdot \sum_{h=1}^{n} x_{hj}^2}$$

Medidas de Similaridad en escalas Nominales y Ordinales basadas en tablas de contingencia h·k

Recordemos que las medidas de asociación, en términos intuitivos podían interpretarse de forma similar a como lo hacíamos con el coeficiente de correlación de Pearson. Eran medidas que, de forma similar al coeficiente de correlación de Pearson, nos proporcionaban un valor tanto mayor cuanto más dependencia existía entre las variables y, recíprocamente, tanto menor cuanto menos dependencia mutua presentaban las mismas. Así que todas ellas nos van a permitir medir, en este sentido, la similaridad existente entre las variables y, por la misma regla ya referida anteriormente varias veces, restándolas de una cota superior, podremos convertirlas en medidas de disimilaridad.

Sin embargo, debemos tener en cuenta que estas medidas generalmente se definen sobre tablas de contingencia y, por tanto, son válidas para variables medidas en escalas de tipo nominal o tipo ordinal; en definitiva, en escalas de tipo cualitativo en general.

A continuación presentamos algunas medidas, generalmente de asociación basadas en tablas de contingencia, que nos proporcionarán, por lo expuesto, sendas medidas de similaridad. Para la interpretación de estas medidas y su notación, se recomienda repasar el capítulo dedicado medidas de asociación y tablas de contingencia.

- *Similaridades Basadas en el estadístico χ^2 (para escalas nominales)*

Cuadrado del Coeficiente de Contingencia χ^2 :
(suele usarse para variables)

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(e_{ij}-n_{ij})^2}{e_{ij}}, \quad e_{ij} = \frac{n_{i\cdot}\cdot n_{\cdot j}}{N}$$

Coeficiente de Contingencia χ :
(suele usarse para variables)

$$\chi = \sqrt{\chi^2} = \sqrt{\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(e_{ij}-n_{ij})^2}{e_{ij}}}$$

Coeficiente de Contingencia Cuadrático Medio:
(suele usarse para variables)

$$\phi^2 = \frac{\chi^2}{N}$$

Coeficiente de Contingencia de Pearson:
(suele usarse para variables)

$$P = \sqrt{\frac{\chi^2}{N+\chi^2}} = \sqrt{\frac{\phi^2}{1+\phi^2}}$$

Coeficiente T de Tschuprov:
(suele usarse para variables)

$$T = \left\{\frac{\frac{\chi^2}{N}}{\sqrt{(h-1)(k-1)}}\right\}^{1/2}$$

Coeficiente V de Cramer:
(suele usarse para variables)

$$V = \left\{\frac{\frac{\chi^2}{N}}{min(\mathrm{h}-1, \mathrm{k}-1)}\right\}^{1/2}$$

- Similaridades Basadas en la reducción del error de predicción (escalas nominales)

λde Kruskal y Goodman: (suele usarse para variables)

$$\lambda_{X|Y} = \frac{\sum_{j=1}^{k} \max_{i=1,\ldots,h}\{n_{ij}\} - \max_{i=1,\ldots,h}\{n_{i.}\}}{n - \max_{i=1,\ldots,h}\{n_{i.}\}}$$

$$\lambda_{Y|X} = \frac{\sum_{i=1}^{h} \max_{j=1,\ldots,k}\{n_{ij}\} - \max_{j=1,\ldots,k}\{n_{.\mathrm{j}}\}}{n - \max_{j=1,\ldots,k}\{n_{.\mathrm{j}}\}}$$

$$\lambda = \frac{\sum_{i=1}^{h} \max_{j=1,\ldots,k}\{n_{ij}\} + \sum_{j=1}^{k} \max_{i=1,\ldots,h}\{n_{ij}\} - \max_{i=1,\ldots,h}\{n_{i.}\} - \max_{j=1,\ldots,k}\{n_{.\mathrm{j}}\}}{2n - \max_{i=1,\ldots,h}\{n_{i.}\} - \max_{j=1,\ldots,k}\{n_{.\mathrm{j}}\}}$$

τ de Kruskal y Goodman: (suele usarse para variables)

$$\tau_{X|Y} = \frac{n\sum_{i=1}^{h}\sum_{j=1}^{k}\left\{\frac{n_{ij}^2}{n_{.\mathrm{j}}}\right\} - \sum_{i=1}^{h} n_{i.}^2}{n^2 - \sum_{i=1}^{h} n_{i.}^2}$$

$$\tau_{Y|X} = \frac{n\sum_{i=1}^{h}\sum_{j=1}^{k}\left\{\frac{n_{ij}^2}{n_{i.}}\right\} - \sum_{j=1}^{k} n_{.\mathrm{j}}^2}{n^2 - \sum_{j=1}^{k} n_{.\mathrm{j}}^2}$$

$$\tau = \frac{n\sum_{i=1}^{h}\sum_{j=1}^{k}\left\{\frac{n_{ij}^2}{n_{i.}}\right\} + n\sum_{i=1}^{h}\sum_{j=1}^{k}\left\{\frac{n_{ij}^2}{n_{.\mathrm{j}}}\right\} - \sum_{i=1}^{h} n_{\mathrm{i}.}^2 - \sum_{j=1}^{k} n_{.\mathrm{j}}^2}{2n^2 - \sum_{j=1}^{k} n_{.\mathrm{j}}^2 - \sum_{i=1}^{h} n_{\mathrm{i}.}^2}$$

- Similaridades Basadas en las concordancias (escalas ordinales)

τ de Kendall: (suele usarse para variables)

$$\tau_A = \frac{2(P-Q)}{n(n-1)}$$

$$\tau_b = \frac{P-Q}{\sqrt{(P+Q+X_0)(P+Q+Y_0)}}$$

$$\tau_C = \frac{2q(P-Q)}{n^2(q-1)}, \quad q = min(h,k)$$

γ de Goodman: (suele usarse para variables)

$$\gamma = \frac{P - Q}{P + Q}$$

d de Sommers: (suele usarse para variables)

$$d_{Y|X} = \frac{P - Q}{P + Q + Y_0}$$
$$d_{X|Y} = \frac{P - Q}{P + Q + X_0}$$

- Similaridades Basadas en correlación de rangos (escalas ordinales)

Coeficiente de correlación de rangos de Spearman: (suele usarse para variables)

$$\rho_s = 1 - 6\frac{\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Medidas de Similaridad en escalas Binarias o Dicotómicas basadas en tablas de contingencia 2·2

Recordemos que las tablas de contingencia 2·2 enfrentaban variables X_i y X_j medidas en escalas binarias o dicotómicas y que, por tanto, sólo podían tomar dos valores diferentes que, por convenio, habíamos notado como 0 ó 1 (ausencia y presencia de una determinada cualidad). Así, en la tabla se recogía ordenadamente el número de individuos que presentaban cada una de las modalidades conjuntas (0,0), (0,1), (1,0) y (1,1) en la forma ya conocida:

X_i \ X_j	**1**	**0**
1	*a*	*b*
0	*c*	*d*

Sobre esta tabla, podríamos aplicar todas las medidas de asociación vistas para las tablas de contingencia de dimensión h·k, caso general. Pero su estructura simple, permite extraer de ella otra gran variedad de medidas de similaridad y asociación de las que, posteriormente, vamos a destacar las más conocidas.

Pero además, esta tabulación también permite comparar el comportamiento de dos casos cuando sus variables se observan sobre escalas de tipo binario o dicotómico, indicando en cada casilla el número de variables que toman simultáneamente el valor 0, simultáneamente el valor 1, cero en una variable y 1 en la otra, y viceversa, según el siguiente planteamiento.

Cuando tenemos datos en escalas binarias, cada una de las variables, X_1 , X_2 ,..., X_p , puede tomar los valores 0 ó 1 en función de que no tengan o tengan una determinada cualidad. Cuando intentamos comparar dos casos —por ejemplo el caso *i* y el caso *j* —, cada uno de estos casos tendrá unos comportamientos expresados en términos de ceros y unos para cada una de esas variables, porque las variables son binarias, que podemos representar a modo de ejemplo como sigue:

	X_1	X_2	·····	X_p
Caso *i*	0	1	·····	1
Caso *j*	1	1	·····	0

Así, a partir de esta situación podemos construir una tabla de contingencia de dimensión 2·2 donde expresar la comparación de los dos casos de la forma:

Caso j / ***Caso i***	**1**	**0**
1	*a*	*b*
0	*c*	*d*

donde la frecuencia *a*, correspondiente al par (0,0), indicaría el número de variables que toman simultáneamente el valor 0 los casos *i* y *j*; es decir, el número de variables, de entre las *p* que nosotros estamos observando, en las que hay ausencia de la cualidad de referencia simultáneamente en los dos casos comparados; y análogamente, el valor *d*, que corresponde al par (1,1), representaría el número de variables donde su cualidad de referencia está presente simultáneamente en los dos casos. Por tanto, *a* y *d* (la diagonal principal), nos indican el número de variables donde el comportamiento de los dos casos es similar, tanto por presencia como por ausencia simultánea de las cualidades de referencia. Por el contrario, *b* y *c* representarían aquellas situaciones en las que las características se presentan en un caso (1) pero no en el otro caso (0) (variables que para el primer caso tienen un valor 0 y en el otro caso tienen un valor 1, o a la inversa). Dicho de otra manera, *c* y *b* representan a aquellas variables en las que el comportamiento es diferente para los dos casos comparados.

Si hacemos pues esta aproximación, cada Tabla de Contingencia que enfrenta o compara dos casos permite deducir medidas de asociación y similaridad para escalas binarias, que nos informarán sintéticamente sobre la proximidad entre los casos comparados. Presentamos a continuación las más comunmente empleadas.

- *Medidas Basadas en el estadístico χ^2 (escalas binarias o dicotómicas)*

Chi-2 (χ^2): (suele usarse para Variables y Casos)

$$\chi^2_{\exp} = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

$$\textit{corrección de continuidad de Yates}: \chi^2_{\exp} = \frac{N(|ad-bc|-0.5N)^2}{(a+b)(a+c)(c+d)(b+d)}$$

Odds ratio: (suele usarse para Variables y Casos)

$$= \frac{a \cdot d}{b \cdot c} = \frac{a / c}{b / d} = \frac{a / b}{c / d}$$

Q de Yule: (suele usarse para Variables y Casos)

$$Q = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$$

- Medidas Basadas en las concordancias (escalas binarias o dicotómicas)

Russel y Rao: (suele usarse para Variables y Casos)

$$RR(X,Y) = \frac{a}{a+b+c+d}$$

Correspondencia Simple: (suele usarse para Variables y Casos)

$$CS(X,Y) = \frac{a+d}{a+b+c+d}$$

Jaccard: (suele usarse para Variables y Casos)

$$J(X,Y) = \frac{a}{a+b+c}$$

Dice, Czekanowski y Sorenson: (suele usarse para Variables y Casos)

$$DCS(X,Y) = \frac{2a}{2a+b+c}$$

Rogers y Tanimoto: (suele usarse para Variables y Casos)

$$RT(X,Y) = \frac{a+d}{a+d+2(b+c)}$$

Sokal y Sneath 1: (suele usarse para Variables y Casos)

$$SS1(X,Y) = \frac{2(a+d)}{2(a+d)+b+c}$$

Sokal y Sneath 2: (suele usarse para Variables y Casos)

$$SS2(X,Y) = \frac{a}{a+2(b+c)}$$

Sokal y Sneath 3: (suele usarse para Variables y Casos)

$$SS3(X,Y) = \frac{a+d}{b+c}$$

Kulczynski 1: (suele usarse para Variables y Casos)

$$K(X,Y) = \frac{a}{b+c}$$

- Medidas Basadas en probabilidad condicional (escalas binarias o dicotómicas)

Hamann: (suele usarse para Variables y Casos)

$$H(X,Y) = \frac{a+d-(b+c)}{a+b+c+d}$$

Sokal y Sneath 4: (suele usarse para Variables y Casos)

$$SS4(X,Y) = \frac{1}{4}\cdot\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$$

Kulczynski 2: (suele usarse para Variables y Casos)

$$K2(X,Y) = \frac{1}{2}\cdot\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$$

- Medidas Basadas en predictibilidad (escalas binarias o dicotómicas)

D de Anderberg: (suele usarse para Variables y Casos)

$$DA(X,Y) = \frac{max(a,b)+max(c,d)+max(a,c)+max(b,d)-max(a+c,b+d)-max(a+b,c+d)}{2(a+b+c+d)}$$

Y de Yule: (suele usarse para Variables y Casos)

$$Y = \frac{\sqrt{a\cdot d}-\sqrt{b\cdot c}}{\sqrt{a\cdot d}+\sqrt{b\cdot c}}$$

- Otras Medidas (escalas binarias o dicotómicas)

Sokal y Sneath 5: (suele usarse para Variables y Casos)

$$SS5(X,Y) = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Ochiai: (suele usarse para Variables y Casos)

$$O(X,Y) = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

Dispersión: (suele usarse para Variables y Casos)

$$D(X,Y) = \frac{ad - bc}{(a+b+c+d)^2}$$

Coeficiente de correlación de Pearson: (suele usarse para Variables y Casos)

$$r = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Otras medidas de Disimilaridad en escalas Binarias o Dicotómicas basadas en tablas de contingencia 2·2

Distancia Euclídea: (suele usarse para Variables y Casos)

$$d_2 = \sqrt{b+c}$$

Cuadrado de la Distancia Euclídea: (suele usarse para Variables y Casos)

$$d_2^2 = b + c$$

Diferencia de Tamaño: (suele usarse para Variables y Casos)

$$DT(X,Y) = \frac{(b-c)^2}{(a+b+c+d)^2}$$

Diferencia de Configuración: (suele usarse para Variables y Casos)

$$DC(X,Y) = \frac{bc}{(a+b+c+d)^2}$$

Diferencia de forma: (suele usarse para Variables y Casos)

$$DF(X,Y) = \frac{(a+b+c+d)(b+c) - (b-c)^2}{(a+b+c+d)^2}$$

Varianza Disimilar: (suele usarse para Variables y Casos)

$$V(X,Y) = \frac{b+c}{4\cdot(a+b+c+d)}$$

Lance y Williams: (suele usarse para Variables y Casos)

$$LW(X,Y) = \frac{b+c}{2a+b+c}$$

Otras medidas de Disimilaridad válidas cuando se usan varios tipos de escala

Coeficiente de Similaridad de Gower (suele usarse para Casos)

Permite calcularla para cuando las variables vienen en escalas nominales o de intervalo.

$$d_{ij} = \frac{\sum_{k=1}^{p} w_k \delta_{ijk} S_{ijk}}{\sum_{k=1}^{p} w_k \delta_{ikj}}, \quad \text{siendo}: S_{ijk} = \begin{cases} \text{cuando } X_k \text{ es variable}: 1 - \left| \dfrac{x_{ik} - x_{jk}}{\max\limits_{l}\{x_{lk}\} - \min\limits_{l}\{x_{lk}\}} \right| \\ \text{cuando } X_k \text{ es atributo}: \begin{cases} 1 \text{ si } x_{ik} = x_{jk} \\ 0 \text{ si } x_{ik} \neq x_{jk} \end{cases} \end{cases}$$

w_k = factor de ponderación de cada variable k - ésima

$$\delta_{ikj} = \begin{cases} 1, \text{ si la característica k puede compararse para los casos i y j} \\ 0, \text{ si la característica k no puede compararse para los casos i y j} \end{cases}$$

Preparación de datos para el cálculo de proximidades

Hasta aquí, hemos pretendido mostrar un conjunto amplio y útil de indicadores o medidas que nos informan de cómo son de parecidas las variables o los datos, con el objeto de poder clasificar, las unas o los otros, en grupos homogéneos (formados por elementos parecidos) y que se diferencien claramente unos de otros (elementos de grupos distintos poco parecidos).

Ahora bien, todas estas las medidas de similaridad o disimilaridad (con excepción de la de Gower), tanto para casos o para variables, exigen siempre que todas las variables observadas sobre los casos estén evaluadas sobre un mismo tipo de escala. Así, si utilizamos como medida de disimilaridad la distancia euclídea, todas las variables X_i tendrán que estar medidas sobre escalas de intervalo; si utilizamos como medida de smilaridad una medida de asociación basada en concordancias, entonces todas las variables deben estar medidas en escalas ordinales; etc. Todas estas medidas de similaridad o disimilaridad para datos multivariantes se han definido, por tanto, cuando las escalas en las que se miden las variables observadas son todas del mismo tipo; es decir, homogéneas.

Sin embargo, esto no es lo normal cuando nos enfrentamos con un problema real. En la práctica, lo normal es que las variables se presenten medidas en diferentes escalas. Es fácil en la práctica, por ejemplo, considerar variables como el sexo (escala nominal-dicotómica), la edad (escala de razón u ordinal) o el nivel de estudios (escala ordinal) de los individuos estudiados. Es decir, se puede presentar heterogeneidad de las escalas de medida de las variables observadas; así que, normalmente, tendremos que preparar los datos para que esas medidas de similaridad y disimilaridad puedan ser calculadas.

Homogeneización de las escalas

Para homogeneizar las escalas sobre las que se miden las variables observadas, podemos recurrir a una doble vía, como ya se expuso en el apartado dedicado a las diferentes escalas de medida en esta obra.

- Cambiar una escala a un tipo de escala más informativa (de nominal a ordinal, de ordinal a intervalo, o de intervalo a razón) para lo que habría que introducir, de forma subjetiva, la información que nos falta para poder movernos en una escala más detallada. Recordemos que en este caso, los nuevos datos en la nueva escala no sólo contendrán la información de los datos antiguos, sino también la que hayamos introducido subjetivamente para el cambio de escala, por lo que los resultados vendrán afectados también por esta. ¡Cuidado con la información que se introduce!

- Cambiar una escala a un tipo de escala menos informativa (de razón a intervalo, de intervalo a ordinal, o de ordinal a nominal). En este caso, perdemos la información del detalle de las escalas originales, que no se conserva en la nueva.

De cualquiera de estas dos maneras podríamos llegar a tener todas las variables de nuestro trabajo medidas en un mismo tipo de escala, completamente homogeneizadas, pudiendo entonces utilizar sobre los datos las medidas de proximidad pertinentes.

Estandarización de variables (medidas en escalas de intervalo)

Cuando trabajamos finalmente con datos medidos en la escala de intervalo o de razón, y una vez convertidas todas las escalas a este tipo, nos encontramos con que cada variable afecta a las fórmulas de las medidas de similaridad y disimilaridad expuestas, en función de la magnitud de sus valores, así que las diferentes magnitudes de los rangos de las variables intervinientes en las fórmulas de las proximidades pueden hacer que unos valores influyan mucho más en sus resultados que otros. Por ejemplo, si estamos calculando una distancia euclídea e intervienen variables como podría ser un ingreso medido en pesetas y una edad medida en años, es evidente que las diferencias en la dimensión de los ingresos que intervienen en la fórmula de la distancia (probablemente de miles de ptas.) van a influir en la fórmula mucho más que las diferencias de la dimensión edad (probablemente de unas cuantas unidades o decenas a lo sumo).

Habrá veces que los valores de las variables tengan comparativamente un sentido absoluto de interés para el estudio, en cuyo caso emplearemos las variables tal como se

hayan definido. Sin embargo, otras muchas veces nos interesará más considerar los efectos que las variaciones relativas de valor presentan en cada variable, que los efectos absolutos derivados simplemente de las unidades de las escalas utilizadas. En estos casos necesitaremos estandarizar las escalas para llevarlas hacia rangos de variación comparables en magnitud, y exponemos a continuación los más utilizados.

Tipificación:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$$

Este procedimiento es el clásicamente utilizado en estadística para la estandarización de las variables mediante un cambio de origen (la media pasa a ser el cero) y de escala (la desviación típica pasa a ser 1), lo que se consigue restando de cada valor de la variable que se esté estandarizando, su media y dividiendo por su desviación típica. Con este procedimiento, lo que se consigue es que la nube de puntos que forma nuestros datos, se contraiga o expanda en cada dimensión para que finalmente la nube resultante de la transformación, manteniendo su forma, se inscriba, más o menos, en un hipercubo con centro el nuevo origen de coordenadas (0,0,...,0) y con una dispersión media en cada eje (desviación típica) de 1.

Transformación para que la desviación típica =1

Con este procedimiento, se consigue que la nube de puntos se contraiga o expanda en cada dimensión para que finalmente quede inscrita, manteniendo su forma, en un hipercubo en torno al nuevo origen de coordenadas o centroide $\left(\frac{\bar{x}_1}{S_1}, \frac{\bar{x}_2}{S_2}, ..., \frac{\bar{x}_p}{S_p}\right)$ y con una dispersión media en cada eje (desviación típica) de 1.

Transformación para que la media =1

$$z_{ij} = \frac{x_{ij}}{\bar{x}_j}, \quad \text{si } \bar{x}_j \neq 0$$

Transformación para que el máximo =1

$$z_{ij} = \frac{x_{ij}}{\max_i(x_{ij})}, \quad \text{si } \max_i(x_{ij}) \neq 0$$

Transformación de rango al intervalo [-1;+1]

$$z_{ij} = \frac{2 \cdot (x_{ij} - \min_i(x_{ij}))}{\max_i(x_{ij}) - \min_i(x_{ij})} - 1$$

Esta transformación de rango consigue que la nube de puntos se contraiga o expanda en cada dimensión para que finalmente la nube adopte la forma de un cubo con centro en el origen de coordenadas, y lados de longitud 2 (una unidad a cada lado del origen) paralelos a los ejes de coordenadas.

Transformación de rango al intervalo [0;+1]

$$z_{ij} = \frac{x_{ij} - \min_i(x_{ij})}{\max_i(x_{ij}) - \min_i(x_{ij})}$$

Esta transformación de rango consigue que la nube de puntos se contraiga o expanda en cada dimensión para que finalmente la nube adopte la forma de un cubo con lados de longitud la unidad, situado en el primer cuadrante y con uno de sus vértices inferiores en el origen de coordenadas.

Remarquemos aquí que para poder realizar cualquiera de estos tipos de estandarización debemos estar al menos en una escala de intervalo: no podemos calcular una media o una desviación típica si no podemos establecer distancias, y las distancias se establecen en escalas de intervalos. Si el mecanismo de cálculo que estuviésemos empleando (programa de ordenador, por ejemplo) nos permitiese aparentemente realizar una tipificación, u otro tipo de las siguientes estandarizaciones, en una escala ordinal o en una escala nominal, estaría realizando implicitamente, además de la estandarización, una transformación de la escala nominal u ordinal a la necesaria de intervalo considerando las correspondientes "etiquetas numéricas" con las que estamos identificando las modalidades de estas escalas como valores de la escala de intervalo, más informativa. Esta conversión de "etiquetas numéricas" en "valores" sería la información subjetiva que estaríamos introduciendo en el proceso y hemos de tenerlo en cuenta, para evaluar su adecuación y considerar su influencia en los resultados.

Transformación de las proximidades

Cuando evaluamos las disimilaridades (o alternativamente similaridades) entre los n casos o las p variables que estamos considerando, el resultado puede presentarse en una matriz D de dimensiones n·n o p·p, donde cada elemento, d_{ij}, representa la medida de proximidad empleada y evaluada al comparar el elemento(caso o variable) i-ésimo con el j-ésimo. Por tanto esta matriz tiene siempre la diagonal principal nula (o alternativamente máxima), y para cuando las proximidades son simétricas, lo que es la situación más general, la matriz D también es simétrica.

Ello supone que al menos debemos consideran n(n-1)/2 ó p(p-1)/2 coeficientes de proximidad entre los elementos considerados, lo que normalmente suele dar un número bastante elevado. Es por ello, por lo que a la hora de evaluar los resultados, a veces interesa estandarizar también los valores resultantes de forma que sea más sencilla su interpretación comparativa.

Para atacar este problema, análogamente a como se transformaron los rangos de las variables, podemos realizar transformaciones del rango de los resultados (proximidades) a un intervalo, siendo el más comúnmente utilizado el intervalo [0;+1]

Por otro lado, puede interesar extraer de los resultados exclusivamente la información sobre la relación de parecido entre las variables (o casos), sin tener en cuenta el sentido de la relación. Recordemos que éste era el caso cuando considerábamos como medida de similaridad al coeficiente de determinación en lugar del coeficiente de correlación de Pearson. Cuando se quiere generalizar este razonamiento, resulta útil transformar las proximidades resultantes mediante su valor absoluto. En cualquier caso, esta transformación no debe ser empleada de forma indiscriminada y tendremos que recurrir a la definición de proximidad adoptada para garantizar la adecuación de su utilización.

Finalmente, a veces interesará considerar funciones de simililaridad en lugar de disimilaridades. En estos casos puede ser útil la transformación de los resultados mediante el cambio de signo ya que así cualquier disimilaridad se convierte en similaridad de cota superior 0, si bien las similaridades resultantes se vuelven todas negativas, lo que dificulta su manejo y generalmente interesa volver a realizar otra transformación del rango de resultados a otros valores más cómodos.

Anexo:

Notemos por *X* la matriz de los n datos observados en un espacio de p variables. Recordemos que, en la **matriz de datos** *X,* las filas representan los casos y las columnas las variables. Así,

$$X = \begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ x_{21} & \ddots & & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix}$$

Fijémonos que si promediamos cada una de estas columnas, obtenemos las medias de las distintas variables $\bar{x}_1$, $\bar{x}_2$,..., $\bar{x}_p$. A este vector, en forma de columna, lo llamaremos **centroide** de los casos y representa al centro de gravedad de la nube de puntos en el espacio de las variables.

$$\bar{x} = \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_p \end{pmatrix}'$$

Si a partir de esta matriz de datos *X,* restamos a cada variable su media, obtendremos lo que llamamos la **matriz de datos centrados**, X_c. En esta matriz X_c cada columna (variable) presenta datos centrados en torno a cero, ya que al haber restado en cada columna su media, el antiguo valor central (media) se ha convertido en el cero.

$$X_C = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & \ddots & & x_{2p} - \bar{x}_p \\ \vdots & & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

Con esta notación, podemos comprobar fácilmente que la **matriz de varianzas y covarianzas**, S, será:

$$S = \begin{pmatrix} S_1^2 & S_{12} & \cdots & S_{1p} \\ S_{12} & S_2^2 & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1p} & S_{2p} & \cdots & S_p^2 \end{pmatrix} = \frac{1}{n} X_C' X_C$$

La matriz de varianzas y covarianzas, S, presenta en su diagonal principal las varianzas de las variables $(S_1^2,....,S_p^2)$ y en los triángulos superior e inferior, las covarianzas de las variables correspondientes a la fila y a la columna en que se encuentra situada, $S_{12},...,S_{1p},S_{12},...,S_{2p},...,S_{1p},S_{2p},....$ Lógicamente, estos triángulos superior e inferior son simétricos puesto que la covarianza de dos variables es la misma independientemente del orden en que se las considere (S_{xy}=S_{yx}).