Capítulo 1

Introducción al Análisis Cluster. Consideraciones generales.

1.1. El problema de la clasificación.

Una de las actividades más primitivas, comunes y básicas del hombre consiste en clasificar objetos en categorías. Las personas, objetos y sucesos encontrados en un día son demasiado numerosos para procesarlos mentalmente como entidades aisladas.

Clasificación o identificación es el proceso o acto de asignar un nuevo objeto u observación en su lugar correspondiente dentro de un conjunto de categorías establecido. Los atributos esenciales de cada categoría son conocidos, aunque haya algunas incertidumbres a la hora de asignar alguna observación dada. Como ejemplo claro, la clasificación se necesita para el desarrollo del lenguaje, el cual consiste en palabras que nos ayudan a reconocer y discutir los diferentes tipos de sucesos, objetos y gentes que nos encontramos. Cada nombre es una etiqueta usada para describir una clase de objetos que poseen notables características en común. Nombrar es clasificar.

Al igual que es una actividad humana conceptual básica, la clasificación es también fundamental en la mayoría de las ramas de la ciencia. En Biología, por ejemplo, la clasificación de organismos ha sido una preocupación desde las primeras investigaciones biológicas. Aristóteles construyó un elaborado sistema de clasificación de especies del reino animal; él empezó dividiendo los animales en dos grupos principales: los que tenían sangre roja (correspondiente a los vertebrados) y los que no la tienen (invertebrados). Además subdividió esos dos grupos de acuerdo a la forma en la que los descendientes venían al mundo (ovíparos y vivíparos). Tras Aristóteles, Teófrates redactó el primer informe fundamental sobre la estructura y clasificación de las plantas. El resultado fue unos libros ampliamente documentados y profundos, abarcando tantos conceptos en sus temas que nos han provisto de la base de las investigaciones biológicas durante muchos siglos. Fueron sustituidos en los siglos XVII y XVIII cuando los grandes exploradores europeos dieron lugar a la segunda búsqueda y colección, bajo la dirección del naturalista sueco Linnaens. Dicho naturalista publicó su trabajo *Genera Plantarum*, en el cual podemos leer:

> *...Todo el conocimiento real que nosotros poseemos depende de los métodos por los cuales distinguimos lo similar de lo no similar. El gran número de diferencias naturales que este método comprende llega a darnos una idea más clara de las cosas...*

En Biología, la teoría y práctica de la clasificación de los organismos es conocida generamente como Taxonomía. Inicialmente, la taxonomía en un sentido más amplio fue, quizás, más un arte que un método científico, pero, eventualmente, fueron desarrolladas técnicas menos subjetivas por Adanson (1727-1806), quien es avalado por Sokal y Sneath (1963) con la introducción del *polithetic*, tipo de sistemas dentro de la Biología en los cuales las clasificaciones se basan en muchas características de los objetos, siendo estudiados por oposición a los sitemas *monothetic*, los cuales usan una única característica para producir una clasificación.

La clasificación de animales y plantas ha jugado un papel importante en el campo de la Biología y de la Zoología, particularmente como una base para la teoría de la evolución de Darwin. Pero la clasificación ha jugado también un papel central en el desarrollo de teorías en otros campos de la ciencia. La clasificación de los elementos en la tabla periódica, por ejemplo, producida por Mendelejev en 1860 causó un profundo impacto en el entendimiento de la estructura del átomo. En Astronomía, la clasificación de las estrellas en *enanas*

y *gigantes* usando el campo Herstsprung-Russell de temperatura frente a luminosidad, afectó fuertemente a las teorías de la evolución de las estrellas.

Las técnicas numéricas para obtener clasificaciones se originaron en las ciencias naturales como la Biología y la Zoología, en un esfuerzo por librar a la Taxonomía de su subjetivismo tradicional y proporcionar clasificaciones objetivas y estables, objetivas en el sentido de que el análisis del mismo conjunto de organismos por diferentes métodos numéricos proporcionen la misma clasificación y estables en el sentido de que la clasificación permanezca igual bajo la inclusión de una gran variedad de organismos o de nuevos caracteres.

La segunda mitad de este siglo ha visto un gran aumento en el número de técnicas numéricas de clasificación disponibles. Este crecimiento ha ido paralelo con el desarrollo de los ordenadores, que son necesarios para poder realizar el gran número de operaciones aritméticas que se precisan. Asimismo, un desarrollo similar ha tenido lugar en las áreas de aplicación. Actualmente tales técnicas son usadas en campos como la arqueología, psiquiatría, astronomía e investigación de mercados.

Una variedad de nombres han sido aplicados a estos métodos, dependiendo del área de aplicación. *Taxonomía Numérica* se usa en Biología. En Psicología se emplea el término *Q-análisis*. En inteligencia artificial se usa el nombre de *Reconocimiento de Patrones*. En otras áreas se emplea *Agrupación* y *agrupamiento*. Actualmente, no obstante, el término más genérico es *Análisis Cluster*. El problema con el que estas técnicas se encuentran puede ser establecido en general como sigue:

> *Dado un conjunto de m objetos individuales (animales, plantas, etc.), cada uno de los cuales viene descrito por un conjunto de n características o variables, deducir una división útil en un número de clases. Tanto el número de clases como las propiedades de dichas clases deben ser determinadas.*

La solución generalmente buscada es una partición de los m objetos, o sea, un conjunto de grupos donde un objeto pertenezca a un grupo sólo y el conjunto de dichos grupos contenga a todos los objetos. Formalmente hablando, se parte de una muestra Ξ de m individuos, $X_1, \ldots, X_m$, cada uno de los cuales está representado por un vector n-dimensional, $X_j = (x_{j1}, x_{j2}, \ldots, x_{jn})'$, $j = 1, \ldots, m$ y debemos encontrar una partición de la muestra en regiones $\omega_1, \ldots, \omega_c$ de forma que

$$\bigcup_{i=1}^{c} \omega_i = \Xi$$

$$\omega_i \cap \omega_j = \emptyset \quad ; \quad i \neq j$$

El problema de la clasificación puede ser complicado debido a varios factores, como la presencia de clases definidas de forma imperfecta, la existencia de categorías solapadas y posibles variaciones aleatorias en las observaciones. Una forma de tratar estos problemas, desde el punto de vista estadístico, sería encontrar la probabilidad que tiene cada nueva observación de pertenecer a cada categoría. En este sentido, el criterio de clasificación más simple sería elegir la categoría más probable, mientras que pueden necesitarse reglas más sofisticadas si las categorías no son igualmente probables o si los costos de mala clasificación varían entre las categorías.

1.2. El Análisis Cluster.

Análisis Cluster es el nombre genérico de una amplia variedad de procedimientos que pueden ser usados para crear una clasificación. Más concretamente, un método cluster es un procedimiento estadístico multivariante que comienza con un conjunto de datos conteniendo información sobre una muestra de entidades e intenta reorganizarlas en grupos relativamente homogéneos a los que llamaremos *clusters*.

En Análisis Cluster poca o ninguna información es conocida sobre la estructura de las categorías, lo cual lo diferencia de los métodos multivariantes de asignación y discriminación. De todo lo que se dispone es de una colección de observaciones, siendo el objetivo operacional en este caso, descubrir la estructura de las categorías en la que se encajan las observaciones. Más concretamente, el objetivo es ordenar las observaciones en grupos tales que el grado de asociación natural es alto entre los miembros del mismo grupo y bajo entre miembros de grupos diferentes.

Aunque poco o nada se conoce sobre la estructura de las categorías a priori, se tiene con frecuencia algunas nociones sobre características deseables e inaceptables a la hora de establecer un determinado esquema de clasificación. En términos operacionales, el analista es informado suficientemente sobre el problema, de tal forma que puede distinguir entre buenas y malas estructuras de categorías cuando se encuentra con ellas. Entonces, ¿por qué no enumerar todas las posibilidades y elegir la más atrayente?

El número de formas en las que se pueden clasificar m observaciones en k grupos es un número de Stirling de segunda especie (Abramowitz y Stegun, 1968).

$$\mathbb{S}_m^{(k)} = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^m$$

El problema se complica aún más por el hecho de que usualmente el número de grupos es desconocido, por lo que el número de posibilidades es suma de números de Stirling; así, por ejemplo, en el caso de m observaciones tendríamos que el número total de posibles clasificaciones sería

$$\sum_{j=1}^{m}\mathbb{S}_m^{(j)}$$

que es un número excesivamente grande, por lo que el número de posibles clasificaciones puede ser enorme (por ejemplo, en el caso de 25 observaciones, se tiene que $\sum_{j=1}^{25}\mathbb{S}_{25}^{(j)} > 4\times10^{18}$). Así es necesario encontrar una solución aceptable considerando sólo un pequeño número de alternativas.

Los métodos cluster han sido desarrollados a lo largo de este siglo, pero la mayor parte de la literatura sobre Análisis Cluster ha sido escrita durante las pasadas tres décadas. El principal estímulo para el desarrollo de estos métodos fue el libro *Principios de Taxonomía Numérica*, publicado en 1963 por dos biólogos, Sokal y Sneath. Dichos autores argumentan que un procedimiento eficiente para la generación de clasificaciones biológicas debe recoger todos los posibles datos sobre un conjunto de organismos de interés, estimar el grado de similaridad entre esos organismos y usar un método cluster para colocar los organismos similares en un mismo grupo. Una vez que los grupos de organismos similares han sido encontrados, los miembros de cada uno de ellos deben ser analizados para determinar si representan especies biológicas diferentes. En efecto, Sokal y Sneath asumen que el proceso de reconocimiento de patrones debe ser usado como base para comprender el proceso evolutivo.

A partir de ese momento, la literatura sobre Análisis Cluster se desarrolla de forma considerable. Hay dos razones para el rápido crecimiento y desarrollo de este tipo de técnicas:

1. **El desarrollo de los ordenadores.**

   Antes del auge de los ordenadores, los métodos clusters resultaban molestos y dificultosos desde el punto de vista computacional cuando eran aplicados a conjuntos grandes de datos. Por ejemplo, clasificar un conjunto de datos con 200 entidades requiere buscar una matriz de similaridad con 19.900 valores, y trabajar con una matriz de ese tamaño es una tarea costosa en tiempo que muchos investigadores debían emprender. Obviamente, con la difusión de los ordenadores, el proceso de manejo de grandes matrices se vuelve mucho más factible.

2. **La importancia fundamental de la clasificación como un procedimiento científico.**

   Todas las ciencias están construidas sobre clasificaciones que estructuran sus dominios de investigación. Una clasificación contiene los mejores conceptos usados en una ciencia. La clasificación de los elementos, por ejemplo, es la base para comprender la química inorgánica y la teoría atómica de la materia; la clasificación de las enfermedades proporciona la base estructural para la medicina.

A pesar de su popularidad, los métodos cluster están todavía poco comprendidos y desarrollados en comparación con otros procedimientos estadísticos multivariantes como el análisis factorial, análisis discriminante o multidimensional scaling. La literatura en las ciencias sociales sobre cluster refleja una serie desconcertante y con frecuencia contradictoria de terminología, métodos y aproximaciones, lo cual ha creado un complejo mundo que es virtualmente impenetrable.

Como hemos notado, los métodos cluster se han diseñado para crear grupos homogéneos de casos o entidades. La mayor parte de los usos del Análisis Cluster pueden ser resumidos bajo cuatro objetivos principales:

1. Desarrollar una tipología o clasificación.

2. Investigar esquemas conceptuales útiles para agrupar entidades.

3. Generar hipótesis a través de la exploración de los datos.

4. Contrastar hipótesis o intentar determinar si tipos definidos por otros procedimientos están de hecho presentes en un conjunto de datos.

De estos objetivos, la creación de clasificaciones, probablemente, resulta el objetivo más frecuente de los métodos cluster, pero en la mayor parte de los casos muchos de estos objetivos se combinan para formar la base de estudio.

No obstante, hay que tener algunas precauciones sobre los métodos cluster:

1. La mayor parte de los métodos de Análisis Cluster son procedimientos que, en la mayor parte de los casos, no están soportados por un cuerpo de doctrina estadística teórica. En otras palabras, la mayor parte de los métodos son heurísticos. Esto contrasta con otros procedimientos como el Análisis Factorial, por ejemplo, que está basado sobre una extensa teoría estadística.

2. La mayor parte de los métodos clusters han nacido al amparo de ciertas ramas de la ciencia, por lo que, inevitablemente, están impregnados de un cierto sesgo procedente de esas disciplinas. Esta cuestión es importante puesto que cada disciplina tiene sus propias preferencias tales como los tipos de datos a emplear en la construcción de la clasificación. Así puede haber, por ejemplo, métodos que sean útiles en psicología pero no en biología o viceversa.

3. Distintos procedimientos clusters pueden generar soluciones diferentes sobre el mismo conjunto de datos. Una razón para ello radica en el hecho ya comentado de que los métodos clusters se han desarrollado a partir de fuentes dispares que han dado origen a reglas diferentes de formación de grupos. De esta manera, lógicamente, es necesaria la existencia de técnicas que puedan ser usadas para determinar qué método produce los grupos naturalmente más homogéneos en los datos.

1.3. Cluster por individuos y por variables.

El punto de partida para el Análisis Cluster es, en general, una matriz X que proporciona los valores de las variables para cada uno de los individuos objeto de estudio, o sea

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

La i-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la j-ésima columna muestra los valores pertenecientes a la j-ésima variable a lo largo de todos los individuos de la muestra.

El objetivo de clasificar los datos, como ya se ha comentado, es agrupar individuos u objetos representados por las filas de X. Aparentemente no hay razón para que estos procedimientos no se apliquen a X', obteniéndose así una clasificación de las variables que describen cada individuo. De hecho, muchas de las técnicas cluster existentes (no todas) pueden ser aplicadas para clasificar variables; incluso algunos paquetes estadísticos, como es el caso de BMDP, incluyen implementaciones por separado que permiten realizar análisis cluster por variables (1M) y análisis cluster por individuos (2M).

1.4. Clasificación de las técnicas clusters.

La clasificación que vamos a dar está referida a algunas de las distintas técnicas clusters existentes. Como se podrá comprobar, es bastante extensa, ya que múltiples son los métodos existentes. Asimismo hay que hacer notar que no todos los procedimientos mencionados van a ser tratados con posterioridad, sino que trataremos solamente los más usuales en las aplicaciones prácticas, y por ende sobre los que se posee un mayor grado de experiencia, y que suelen ser los normalmente implementados en los paquetes estadísticos existentes, ya que no se debe perder de vista que sin un potente ordenador y programa informático no es factible el desarrollo práctico de ninguna técnica cluster.

A grandes rasgos se distinguen dos grandes categorías de métodos clusters: métodos jerárquicos y métodos no jerárquicos.

1.4.1. Métodos Jerárquicos.

Estos métodos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen a su vez en aglomerativos y disociativos. Los aglomerativos comienzan el análisis con tantos grupos como individuos haya en el estudio. A partir de ahí se van formando grupos de forma ascendente, hasta que, al final del proceso, todos los casos están englobados en un mismo conglomerado. Los métodos disociativos o divisivos realizan el proceso inverso al anterior. Empiezan con un conglomerado que engloba a todos los individuos. A partir de este grupo inicial se van formando, a través de sucesivas divisiones, grupos cada vez más pequeños. Al final del proceso se tienen tantos grupos como individuos en la muestra estudiada.

Independientemente del proceso de agrupamiento, hay diversos criterios para ir formando los conglomerados; todos estos criterios se basan en una matriz de distancias o similitudes. Por ejemplo, dentro de los métodos aglomerativos destacan:

1. Método del amalgamamiento simple.
2. Método del amalgamamiento completo.
3. Método del promedio entre grupos.
4. Método del centroide.
5. Método de la mediana.
6. Método de Ward.

Dentro de los métodos disociativos, destacan, además de los anteriores, que siguen siendo válidos:

1. El análisis de asociación.
2. El detector automático de interacción.

1.4.2. Métodos no Jerárquicos.

En cuanto a los métodos no jerárquicos, también conocidos como partitivos o de optimización, tienen por objetivo realizar una sola partición de los individuos en K grupos. Ello implica que el investigador debe especificar a priori los grupos que deben ser formados, siendo ésta, posiblemente, la principal diferencia respecto de los métodos jerárquicos, (no obstante hay que señalar que hay diversas versiones de estos procedimientos que flexibilizan un tanto el número final de clusters a obtener). La asignación de individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia de estos métodos respecto a los jerárquicos reside en que trabajan con la matriz de datos original y no precisan su conversión en una matriz de distancias o similitudes. Pedret en 1986 agrupa los métodos no jerárquicos en cuatro familias:

1. **Métodos de Reasignación.**

   Permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior, si ello optimiza el criterio de selección. El proceso acaba cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Dentro de estos métodos están:

   a) El método K-Medias.
   b) El Quick-Cluster análisis.
   c) El método de Forgy.
   d) El método de las nubes dinámicas.

2. **Métodos de búsqueda de la densidad.**

   Dentro de estos métodos están los que proporcionan una aproximación tipológica y una aproximación probabilística.

   En el primer tipo, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre ellos destacan:

a) El análisis modal de Wishart.

b) El método Taxmap.

c) El método de Fortin.

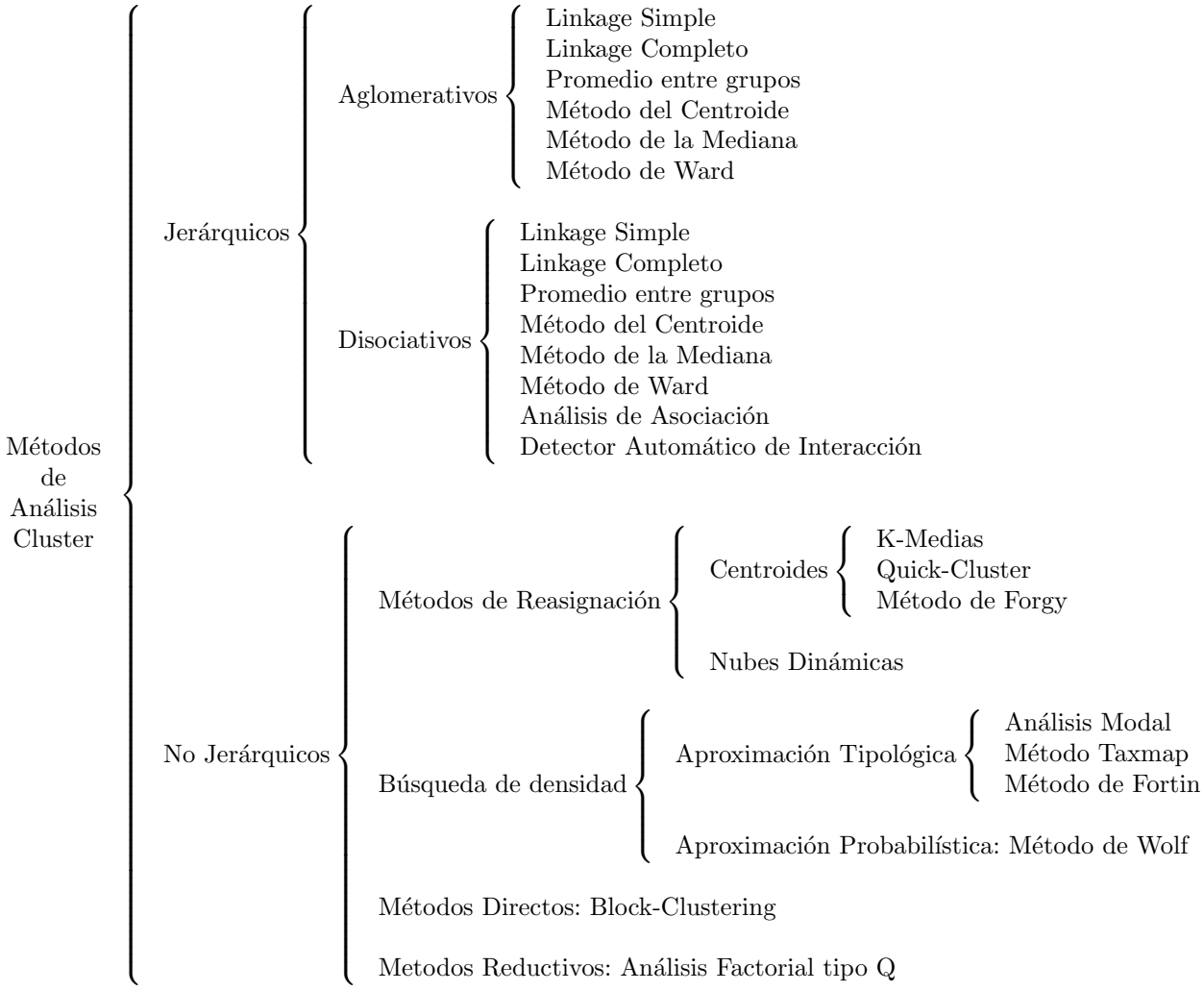
En el segundo tipo se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Entre los métodos de este tipo destaca el método de las combinaciones de Wolf.

3. **Métodos directos.**

   Permiten clasificar simultáneamente a los individuos y a las variables. El algoritmo más conocido dentro de este grupo es el Block-Clustering.

4. **Métodos de reducción de dimensiones.**

   Estos métodos consisten en la búsqueda de unos factores en el espacio de los individuos; cada factor corresponde a un grupo. Se les conoce como Análisis Factorial tipo Q.



1.5. Etapas en Análisis Cluster.

Las etapas a seguir en el empleo de una técnica cluster pueden ser resumidas en los siguientes puntos:

1. **Elección de las variables.**

   La elección inicial del conjunto concreto de características usadas para describir a cada individuo constituye un marco de referencia para establecer las agrupaciones o clusters; dicha elección, posiblemente, refleje la opinión del investigador acerca de su propósito de clasificación. Consecuentemente, la primera cuestión a responder sobre la elección de variables es si son relevantes para el tipo de clasificación que

se va buscando. Es importante tener en cuenta que la elección inicial de variables es, en sí misma, una categorización de los datos, para lo cual sólo hay limitadas directrices matemáticas y estadísticas.

La siguiente cuestión que debe considerarse es el número de variables a emplear. En muchas aplicaciones es probable que el investigador se equivoque tomando demasiadas medidas, lo cual puede dar origen a diversos problemas, bien sea a nivel computacional o bien porque dichas variables adicionales oscurezcan la estructura de los grupos.

En muchas aplicaciones las variables que describen los objetos a clasificar no están medidas en las mismas unidades. En efecto, puede haber variables de tipos completamente diferentes, algunas categóricas, otras ordinales e incluso otras que tengan una escala de tipo intervalo.

Es claro que no sería correcto tratar como equivalentes en algún sentido, por ejemplo, el peso medido en kilos, la altura en milímetros y valorar la ansiedad en una escala de cuatro puntos.

Para variables de tipo intervalo, la solución general consiste en tipificar las variables antes del análisis, calculando las desviaciones típicas a partir de todos los individuos. Algunos autores, por ejemplo Fleiss y Zubin (1969), consideran que esta técnica puede tener serias desventajas al diluir las diferencias entre grupos sobre las variables que más discriminen; como alternativa sugieren emplear la desviación estándar entre grupos para tipificar.

Cuando las variables son de tipos diferentes se suele convertir todas las variables en binarias antes de calcular las similaridades. Esta técnica tiene la ventaja de ser muy clarificadora, pero la desventaja de sacrificar información. Una alternativa más atractiva es usar un coeficiente de similaridad que pueda incorporar información de diferentes tipos de variables de una forma sensible, como el propuesto por Gower en 1971 y que después trataremos. Asimismo, para variables mixtas existe la posibilidad de hacer un análisis por separado e intentar sintetizar los resultados a partir de los diferentes estudios.

2. **Elección de la medida de asociación.**

   La mayor parte de los métodos cluster requieren establecer una medida de asociación que permita medir la proximidad de los objetos en estudio. Cuando se realiza un Análisis Cluster de individuos, la proximidad suele venir expresada en términos de distancias, mientras que el Análisis Cluster por variables involucra generalmente medidas del tipo coeficiente de correlación, algunas de las cuales tienen interpretaciones en distintos sentidos mientras que otras son difíciles de describir, dado el carácter subjetivo de las mismas.

   En el capítulo 2 se hace un breve repaso a las medidas de asociación más usuales que suelen emplearse. Destacamos el hecho de estar clasificadas en medidas para variables y para individuos, si bien algunas de ellas pueden considerarse de uso común. La clasificación se ha establecido sobre todo atendiendo a que las prácticas en ordenador se realizarán con el paquete estadístico BMDP, donde existen dos capítulos específicos, uno para Análisis Cluster por variables y otro por individuos, cada uno de los cuales proporciona un conjunto de medidas a poder usar.

   Hay que tener en cuenta, asimismo, la importancia que tienen los tipos de datos a emplear, bien sean éstos categóricos o no. En el capítulo 2 se muestra toda una serie de posibles medidas que abarcan diversas posibilidades según el tipo de datos a utilizar.

3. **Elección de la técnica cluster a emplear en el estudio.**

   Los métodos cluster que se han propuesto y desarrollado en los últimos años son bastante numerosos y muy diversos en cuanto a su concepción, clasificándose, en un primer estado, en jerárquicos y no jerárquicos, distinguiéndose los primeros de los segundos en que las asignaciones de los individuos, hechas por los métodos jerárquicos a los clusters que se van creando permanecen estables durante todo el proceso, no permitiendo reasignaciones posteriores a clusters distintos si hubiera lugar a ello, cuestión que sí es factible en los métodos no jerárquicos. Además, en los métodos jerárquicos, el investigador deberá sacar sus propias conclusiones mientras que en los procedimientos no jerárquicos el número final de clusters está, por lo general, impuesto de antemano, si bien se han desarrollado, dentro de este tipo de métodos, técnicas que permiten una cierta flexibilidad en el número final de clusters, con el fin de evitar posibles perturbaciones en los resultados definitivos.

   Así pues, en algunos problemas prácticos, la elección del método a emplear será relativamente natural, dependiendo, sobre todo, de la naturaleza de los datos usados y de los objetivos finales perseguidos, si bien en otros la elección no será tan clara. Lo que sí es conveniente siempre, a la hora de las aplicaciones prácticas, es no elegir un sólo procedimiento, sino abarcar un amplio abanico de posibilidades y contrastar los resultados obtenidos con cada una de ellas. De este modo, si los resultados finales son parecidos,

podremos obtener unas conclusiones mucho más válidas sobre la estructura natural de los datos. En caso contrario no obtendremos mucha información, si bien grandes diferencias en los resultados obtenidos pueden llevar a plantearnos el hecho de que tal vez los datos con los que se está trabajando no obedezcan a una estructura bien definida.

En los capítulos 3 y 4 desarrollamos los principales métodos cluster existentes, tanto jerárquicos como no jerárquicos.

4. **Validación de los resultados e interpretación de los mismos.**

   Ésta es la última etapa en la secuencia lógica en la que se desarrolla una investigación a través de un método cluster. Sin duda alguna es la más importante, ya que es en ella donde se van a obtener las conclusiones definitivas del estudio.

   Son diversos los métodos propuestos para validar un procedimiento cluster. Por ejemplo, cuando se está trabajando con métodos jerárquicos se plantean dos problemas:

   a) ¿En qué medida representa la estructura final obtenida las similitudes o diferencias entre los objetos de estudio?

   b) ¿Cuál es el número idóneo de clusters que mejor representa la estructura natural de los datos?

   El argumento más empleado para responder a la primera pregunta es el empleo del coeficiente de correlación cofenético, propuesto por Sokal y Rohlf en 1962. Dicho coeficiente mide la correlación entre las distancias iniciales, tomadas a partir de los datos originales, y las distancias finales con las cuales los individuos se han unido durante el desarrollo del método. Altos valores de tal coeficiente mostrarán que durante el proceso no ha ocurrido una gran perturbación en lo que concierne a la estructura original de los datos. En cuanto a la segunda pregunta, muchas son las técnicas existentes, algunas de las cuales, las más empleadas a nivel práctico, están recogidas en el capítulo 3.

   En cuanto a los métodos no jerárquicos, las cuestiones anteriores van perdiendo sentido, mientras que los procedimientos empleados para validar los resultados van encaminados al estudio de la homogeneidad de los grupos encontrados durante el desarrollo del método. Algunos autores han propuesto el empleo de técnicas multivariantes como el análisis multivariante de la varianza (MANOVA), o bien (como BMDP incluye) desarrollar múltiples análisis de la varianza (ANOVA) sobre cada variable en cada cluster. Estos procedimientos, evidentemente, plantean serios problemas y no deben ser considerados como definitivos. Una técnica usualmente empleada, de tipo remuestreo, es la de tomar varias submuestras de la muestra original y repetir el análisis sobre cada una. Si tras repetir el análisis sobre ellas se consiguen soluciones aproximadamente iguales, y parecidas a la obtenida con la muestra principal, se puede *intuir* que la solución obtenida puede ser válida, si bien ésto no sería argumento suficiente para adoptar tal decisión. No obstante, este método es más útil empleado de forma inversa, en el sentido de que si las soluciones obtenidas en las diversas submuestras no guardan una cierta similitud, entonces parece evidente que se debiera dudar de la estructura obtenida con la totalidad de la muestra.

Capítulo 2

Medidas de Asociación.

2.1. Introducción.

Una vez considerado que el objetivo del Análisis Cluster consiste en encontrar agrupaciones naturales del conjunto de individuos de la muestra, es necesario definir qué se entiende por agrupaciones naturales y, por lo tanto, con arreglo a qué criterio se puede decir que dos grupos son más o menos similares. Esta cuestión conlleva otras dos, a saber:

1. Cómo se puede medir la similitud entre dos individuos de la muestra.
2. Cómo se puede evaluar cuándo dos clusters pueden ser o no agrupados.

A continuación vamos a centrarnos en las posibles funciones que pueden elegirse para medir la similitud entre los grupos que sucesivamente se van formando, distinguiendo primeramente entre distancias métricas y similaridades.

2.2. Distancias y Similaridades. Definiciones preliminares.

2.2.1. Distancias.

Definición 2.1 *Sea U un conjunto finito o infinito de elementos. Una función $d : U \times U \longrightarrow \mathbb{R}$ se llama una distancia métrica si $\forall x, y \in U$ se tiene:*

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \Leftrightarrow \ x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)\ ,\ \forall z \in U$

Comentario 2.2.1

La definición anterior de distancia métrica puede exponerse sin necesidad de tantos axiomas. En efecto se puede comprobar que una distancia métrica es una función $d : U \times U \longrightarrow \mathbb{R}$ que verifica los siguientes axiomas

1. $d(x, y) = 0 \ \Leftrightarrow \ x = y$
2. $d(y, z) \leq d(x, y) + d(x, z)\ ,\ \forall x, y, z \in U$

Comentario 2.2.2

Ciertos autores realizan una cierta distinción entre lo que es una función distancia y lo que es una distancia métrica. Para ello definen una distancia como aquella función $d : U \times U \longrightarrow \mathbb{R}$ que verifica

1. $d(x, y) \geq 0$
2. $d(x, x) = 0$

3. $d(x,y) = d(y,x)$

y reservan el nombre de distancia métrica a aquellas distancias que además verifican

1. $d(x,y) = 0 \implies x = y$
2. $d(x,z) \leq d(x,y) + d(y,z)\ ,\ \forall z \in U$

Comentario 2.2.3

Extendiendo el concepto clásico de distancia plasmado anteriormente, algunos autores definen distancias métricas que pueden tomar valores negativos. De esta manera una función distancia métrica sería una función $d : U \times U \longrightarrow \mathbb{R}$ *tal que cumple los siguientes axiomas*

1. $d(x,y) \geq d_0$
2. $d(x,y) = d_0 \iff x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)\ ,\ \forall z \in U$

donde d_0 *puede ser menor que cero. Tal definición la realizan amparándose en el hecho de que, dada una tal función distancia métrica d, se puede definir otra* d' *a partir de ella, de la forma* $d'(x,y) = d(x,y) - d_0$*, demostrándose fácilmente que* d' *es una distancia métrica en el sentido expuesto en la definición 2.1*

Comentario 2.2.4

1. *Una función que verifique los tres primeros apartados de la definición 2.1, pero no así la desigualdad triangular, es llamada semimétrica.*
2. *Se llama ultramétrica a toda métrica que verifique adicionalmente la propiedad*

$$d(x,z) \leq \text{Max}\,\{d(x,y), d(y,z)\}$$

2.2.2. Similaridades.

De forma similar a las distancias, tenemos la siguiente definición de similaridad

Definición 2.2 *Sea* U *un conjunto finito o infinito de elementos. Una función* $s : U \times U \longrightarrow R$ *se llama similaridad si cumple las siguientes propiedades:* $\forall x, y \in U$

1. $s(x,y) \leq s_0$
2. $s(x,x) = s_0$
3. $s(x,y) = s(y,x)$

donde s_0 *es un número real finito arbitrario.*

Definición 2.3 *Una función s, verificando las condiciones de la definición 2.2, se llama similaridad métrica si, además, verifica:*

1. $s(x,y) = s_0 \implies x = y$
2. $|s(x,y) + s(y,z)|s(x,z) \geq s(x,y)s(y,z)\ ,\ \forall z \in U$

Notemos que el segundo apartado de la definición anterior corresponde al hecho de que la máxima similaridad sólo la poseen dos elementos idénticos.

En las siguientes secciones expondremos algunas de las distancias y similaridades más usuales en la práctica.

Consideraremos, en general, m individuos sobre los cuales se han medido n variables $X_1, \ldots, X_n$. Con ello tenemos $m \times n$ datos que colocaremos en una matriz $m \times n$ dimensional

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

La i-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la j-ésima columna muestra los valores pertenecientes a la j-ésima variable a lo largo de todos los individuos de la muestra.

Distinguiremos entre medidas de asociación para individuos y para variables, aunque, técnicamente hablando, son válidas tanto para individuos como para variables (basta, para ello, considerar dichas medidas en un espacio n-dimensional o m-dimensional, esto es, trasponer la matriz).

2.3. Medidas de asociación entre variables.

Para poder *unir* variables es necesario tener algunas medidas numéricas que caractericen las relaciones entre variables. La base de trabajo de todas las técnicas cluster es que las medidas numéricas de asociación sean comparables, esto es, si la medida de asociación de una par de variables es 0,72 y el de otro par es 0,59, entonces el primer par está más fuertemente asociado que el segundo. Por supuesto, cada medida refleja asociación en un sentido particular y es necesario elegir una medida apropiada para el problema concreto que se esté tratando.

2.3.1. Coseno del ángulo de vectores.

Consideremos dos variables X_i y X_j, muestreadas sobre m individuos, y sean x_i y x_j los vectores cuyas $k-$ésimas componentes indiquen el valor de la variable correspondiente en el $k-$ésimo individuo:

$$x_i = (x_{1i}, \ldots, x_{mi})' \quad ; \quad x_j = (x_{1j}, \ldots, x_{mj})'$$

Como es conocido, el producto escalar de dos vectores es:

$$x_i' x_j = \sum_{l=1}^{m} x_{li} x_{lj}$$

que en Estadística se conoce como la suma de los productos cruzados entre x_i y x_j, mientras que el producto escalar de un vector por sí mismo, norma al cuadrado del vector, se llama suma de cuadrados.

Así se tiene:

$$x_i' x_j = ||x_i|| ||x_j|| \cos(\beta) \tag{2.1}$$

donde β es el ángulo entre los vectores x_i y x_j.

Observando la figura (2.1), la distancia desde el origen (O) a B vale $||x_i|| \cos(\beta)$, siendo esta cantidad la proyección ortogonal de x_i sobre x_j. Así el producto escalar puede interpretarse como el producto de la longitud del vector x_j por la longitud de la proyección de x_i sobre x_j.

A partir de (2.1) se tiene

$$\cos(\beta) = \frac{x_i' x_j}{||x_i|| \; ||x_j||} = \frac{\sum_{l=1}^{m} x_{li} x_{lj}}{\left(\sum_{l=1}^{m} x_{li}^2 \sum_{l=1}^{m} x_{lj}^2 \right)^{\frac{1}{2}}} \tag{2.2}$$

El coseno del ángulo es una medida de similaridad entre x_i y x_j, con valores entre -1 y 1 en virtud de la desigualdad de Schwarz. Además es la mejor medida para establecer el paralelismo entre dos vectores, ya que dos vectores son paralelos cuando el coseno del ángulo que forman es uno en valor absoluto.

Figura 2.1: Ángulo entre vectores

Esta medida es independiente, salvo signo, de la longitud de los vectores considerados. Algebráicamente, sean b y c dos escalares cualesquiera y definamos

$$\widehat{x}_i = bx_i \quad ; \quad \widehat{x}_j = cx_j \quad ; \quad b, c \neq 0$$

Entonces:

$$\cos(\widehat{x}_i, \widehat{x}_j) = \frac{\widehat{x}_i' \widehat{x}_j}{||\widehat{x}_i|| ||\widehat{x}_j||} = \frac{\sum_{l=1}^{m} bx_{li} cx_{lj}}{\left(\sum_{l=1}^{m} b^2 x_{li}^2 \sum_{l=1}^{m} c^2 x_{lj}^2 \right)^{\frac{1}{2}}} =$$

$$= \frac{bc \sum_{l=1}^{m} x_{li} x_{lj}}{|bc| \left(\sum_{l=1}^{m} x_{li}^2 \sum_{l=1}^{m} x_{lj}^2 \right)^{\frac{1}{2}}} \operatorname{Sgn}(bc) \cos(x_i, x_j)$$

con lo cual el coseno entre x_i y x_j es invariante ante homotecias, excepto un eventual cambio de signo.

2.3.2. Coeficiente de correlación.

Consideremos ahora las variables X_i y X_j, anteriores y centrémoslas respecto de sus medias, obteniendo unas nuevas variables cuyos valores para la muestra de los m individuos serán

$$\widehat{x}_i = (x_{1i} - \overline{x}_i, \ldots, x_{mi} - \overline{x}_i)' \quad ; \quad \widehat{x}_j = (x_{1j} - \overline{x}_j, \ldots, x_{mj} - \overline{x}_j)'$$

El producto escalar de las dos variables $\widehat{x}_i$ y $\widehat{x}_j$ se llama dispersión (scatter en la literatura anglosajona) de x_i y x_j. El producto escalar de $\widehat{x}_i$ por sí mismo es llamado la dispersión de x_i o la suma de los cuadrados de las desviaciones respecto a $\overline{x}_i$. Dividiendo por m ambas expresiones obtenemos la covarianza y la varianza, respectivamente.

$$\operatorname{Cov}(x_i, x_j) = \frac{\widehat{x}_i' \widehat{x}_j}{m} = \frac{1}{m} \sum_{l=1}^{m} (x_{li} - \overline{x}_i)(x_{lj} - \overline{x}_j)$$

$$\operatorname{Var}(x_i) = \frac{\widehat{x}_i' \widehat{x}_i}{m} = \frac{1}{m} \sum_{l=1}^{m} (x_{li} - \overline{x}_i)^2$$

La correlación muestral entre x_i y x_j se define como

$$r = \frac{\operatorname{Cov}(x_i, x_j)}{(\operatorname{Var}(x_i) \operatorname{Var}(x_j))^{\frac{1}{2}}} = \frac{\sum_{l=1}^{m} (x_{li} - \overline{x}_i)(x_{lj} - \overline{x}_j)}{\left(\sum_{l=1}^{m} (x_{li} - \overline{x}_i)^2 \sum_{l=1}^{m} (x_{lj} - \overline{x}_j)^2 \right)^{\frac{1}{2}}} \tag{2.3}$$

lo cual muestra que r es el coseno del ángulo entre los vectores centrados $\widehat{x}_i$ y $\widehat{x}_j$.

Alternativamente, si se tipifican las variables anteriores:

$$x^*_{li} = \frac{x_{li} - \overline{x}_i}{(\mathrm{Var}(x_i))^{\frac{1}{2}}} \quad ; \quad l = 1, \ldots, m$$

$$x^*_{lj} = \frac{x_{lj} - \overline{x}_j}{(\mathrm{Var}(x_j))^{\frac{1}{2}}} \quad ; \quad l = 1, \ldots, m$$

entonces la correlación entre x_i y x_j es la covarianza entre x^*_i y x^*_j.

Puesto que el coeficiente de correlación es el coseno del ángulo entre los vectores centrados, posee la propiedad vista con anterioridad, de invarianza, salvo signo, del coseno. Además, es invariante a las adiciones de una constante a cada elemento de x_i y x_j. En efecto, si llamamos $\widetilde{x}_i = x_i + b$, se tiene:

$$\widetilde{x}_i - \overline{\widetilde{x}}_i = (x_i + b) - (\overline{x}_i + b) = x_i - \overline{x}_i$$

Por ello, la correlación es invariante frente a transformaciones lineales, excepto posibles cambios de signo.

Así se observa que el coeficiente de correlación posee una invarianza más restrictiva que el coseno. Sin embargo, esta propiedad indica que el coeficiente de correlación discrimina menos que el coseno a la hora de establecer diferencias entre dos variables, ya que, dadas dos variables X e Y, hay muchos más elementos en la clase de equivalencia de todas las transformaciones lineales de X e Y que en la clase de equivalencia de las homotecias de X e Y.

La diferencia esencial entre las dos medidas (ángulo entre variables y el coeficiente de correlación) es que el coseno se basa en los datos originales y por ende emplea las desviaciones al origen mientras que el coeficiente de correlación usa los datos centrados y por tanto emplea las desviaciones respecto a la media. Si el origen está bien establecido y tiene sentido, entonces los datos originales tienen sentido de forma absoluta y el coseno es una medida apropiada de asociación; si, por el contrario, el origen es arbitrario o elegido a conveniencia, entonces los datos originales tienen sentido relativo respecto a su media, pero no respecto al origen. En tal caso es más apropiado el coeficiente de correlación.

2.3.3. Medidas para datos binarios o dicotómicos.

En ocasiones encontramos variables que pueden tomar dos valores (blanco-negro, si-no, hombre-mujer, verdadero-falso, etc). En tales casos se emplea, con frecuencia, el convenio de usar los valores dicotómicos 1 y 0 para ambos valores.

Al relacionar dos variables binarias, se forma una tabla de contingencia 2×2, que se puede esquematizar de la forma

$X_i \backslash X_j$	1	0	Totales
1	a	b	$a+b$
0	c	d	$c+d$
Totales	$a+c$	$b+d$	$m = a+b+c+d$

(2.4)

En la anterior tabla se tiene:

1. a representa el número de individuos que toman el valor 1 en cada variable de forma simultánea.
2. b indica el número de individuos de la muestra que toman el valor 1 en la variable X_i y 0 en la X_j.
3. c es el número de individuos de la muestra que toman el valor 0 en la variable X_i y 1 en la X_j.
4. d representa el número de individuos que toman el valor 0 en cada variable, al mismo tiempo.
5. $a + c$ muestra el número de veces que la variable X_j toma el valor 1, independientemente del valor tomado por X_i.
6. $b + d$ es el número de veces que la variable X_j toma el valor 0, independientemente del valor tomado por X_i.
7. $a + b$ es el número de veces que la variable X_i toma el valor 1, independientemente del valor tomado por X_j.
8. $c + d$ es el número de veces que la variable X_i toma el valor 0, independientemente del valor tomado por X_j.

A continuación presentamos la versión binaria de las medidas introducidas anteriormente.

Medida de Ochiai

En el caso particular de variables dicotómicas, se tiene

$$x_i^{'}x_j = \sum_{l=1}^{m} x_{li}x_{lj} = a$$

$$x_i^{'}x_i = ||x_i||^2 = \sum_{l=1}^{m} x_{li}^2 = a+b$$

$$x_j^{'}x_j = ||x_j||^2 = \sum_{l=1}^{m} x_{lj}^2 = a+c$$

con lo cual el coseno del ángulo entre x_i y x_j queda en la forma:

$$\frac{a}{[(a+b)(a+c)]^{\frac{1}{2}}} = \left[\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)\right]^{\frac{1}{2}} \tag{2.5}$$

medida que es atribuida al zoólogo japonés Ochiai.

En el proceso seguido con las variables dicotómicas puede surgir una situación ambigua, como es el hecho de por qué y cómo asignar los valores 1 y 0 a los valores binarios. Puede ocurrir el caso de que intercambiando los papeles de dichos valores se llegue a resultados distintos, lo cual no es deseable. Por ello, en ocasiones, se toma la media geométrica de los cosenos obtenidos tomando ambos criterios y, más concretamente, se toma el cuadrado de dicha media geométrica, obteniéndose:

$$\left[\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)\left(\frac{d}{b+d}\right)\left(\frac{d}{c+d}\right)\right]^{\frac{1}{2}} \tag{2.6}$$

Hagamos notar que cada uno de los términos de la expresión anterior es una probabilidad condicionada. Así

1. $\frac{a}{a+b}$ es la probabilidad condicionada de que un individuo tome el valor 1 en la variable X_j dado que ha tomado el valor 1 en la variable X_i.
2. $\frac{a}{a+c}$ es la probabilidad condicionada de que un individuo tome el valor 1 en la variable X_i dado que ha tomado el valor 1 en la variable X_j.
3. $\frac{d}{b+d}$ es la probabilidad condicionada de que un individuo tome el valor 0 en la variable X_i dado que ha tomado el valor 0 en la variable X_j.
4. $\frac{d}{c+d}$ es la probabilidad condicionada de que un individuo tome el valor 0 en la variable X_j dado que ha tomado el valor 0 en la variable X_i.

De esta forma, la medida de Ochiai es la media geométrica de las probabilidades condicionadas asociadas con la celda con el valor a, mientras que la expresión (2.6) muestra el cuadrado de la media geométrica de las probabilidades condicionadas asociadas con la diagonal de la tabla (2.4).

Medida Φ

Esta medida se obtiene haciendo uso del coeficiente de correlación sobre dos variables dicotómicas.

$$r = \frac{\sum_{l=1}^{m} x_{li}x_{lj} - \frac{1}{m}\sum_{l=1}^{m} x_{li}\sum_{l=1}^{m} x_{lj}}{\left[\left\{\sum_{l=1}^{m} x_{li}^2 - \frac{1}{m}\left(\sum_{l=1}^{m} x_{li}\right)^2\right\}\left\{\sum_{l=1}^{m} x_{lj}^2 - \frac{1}{m}\left(\sum_{l=1}^{m} x_{lj}\right)^2\right\}\right]^{\frac{1}{2}}}$$

y teniendo en cuenta que

$$\sum_{l=1}^{m} x_{li}x_{lj} = a \qquad \sum_{l=1}^{m} x_{li} = a+b \qquad \sum_{l=1}^{m} x_{lj} = a+c$$

$$\sum_{l=1}^{m} x_{li}^2 = a+b \qquad \sum_{l=1}^{m} x_{lj}^2 = a+c$$

se tiene

$$r = \frac{a - \frac{(a+b)(a+c)}{m}}{\left[\left\{(a+b) - \frac{(a+b)^2}{m}\right\}\left\{(a+c) - \frac{(a+c)^2}{m}\right\}\right]^{\frac{1}{2}}} =$$

$$= \frac{am - (a+b)(a+c)}{\left[(a+b)\{m-(a+b)\}(a+c)\{m-(a+c)\}\right]^{\frac{1}{2}}} =$$

$$= \frac{ad-bc}{\left[(a+b)(c+d)(a+c)(b+d)\right]^{\frac{1}{2}}} \quad ; \quad (m = a+b+c+d) \tag{2.7}$$

Notemos, para finalizar, que, puesto que r es invariante bajo transformaciones lineales, los valores 0 y 1 son arbitrarios, ya que pueden ser transformados de forma lineal a otro par de valores.

Medidas basadas en coincidencias

Una forma intuitiva de medir la similaridad en variables dicotómicas es contar el número de veces que ambas variables toman el mismo valor de forma simultánea. Con ello dos variables serían más parecidas en tanto en cuanto mayor fuera el número de coincidencias a lo largo de los individuos.

No obstante, algunos factores influyen en las medidas que se pueden definir. Por ejemplo, una primera cuestión es qué hacer con las parejas del tipo $0-0$, ya que si las dicotomías son del tipo *presencia-ausencia*, los datos de la casilla d no poseen ningún atributo y no deberían tomar parte en la medida de asociación. Otra cuestión que surge es cómo ponderar las coincidencias y cómo las no coincidencias, o lo que es lo mismo, una diagonal u otra de la tabla (2.4).

A continuación exponemos algunas de las medidas que han ido surgiendo, atendiendo a varios criterios como los anteriormente expuestos.

1. **Medida de Russell y Rao**

$$\frac{a}{a+b+c+d} = \frac{a}{m} \tag{2.8}$$

Este coeficiente mide la probabilidad de que un individuo elegido al azar tenga el valor 1 en ambas variables. Notemos que este coeficiente excluye la pareja $0-0$, al contar el número de coincidencias pero no lo hace así al contar el número de posibles parejas. Asimismo, esta medida proporciona igual peso a las coincidencias y a las no coincidencias.

2. **Medida de parejas simples**

$$\frac{a+d}{a+b+c+d} = \frac{a+d}{m} \tag{2.9}$$

Este coeficiente mide la probabilidad de que un individuo elegido al azar presente una coincidencia de cualquier tipo, pesando de igual forma las coincidencias y las no coincidencias.

3. **Medida de Jaccard**

$$\frac{a}{a+b+c} \tag{2.10}$$

Esta medida mide la probabilidad condicionada de que un individuo elegido al azar presente un 1 en ambas variables, dado que las coincidencias del tipo $0-0$ han sido descartadas primero y por lo tanto han sido tratadas de forma irrelevante.

4. **Medida de Dice**

$$\frac{2a}{2a+b+c} \tag{2.11}$$

Esta medida excluye el par $0-0$ de forma completa, pesando de forma doble las coincidencias del tipo $1-1$. Se puede ver este coeficiente como una extensión de la medida de Jaccard, aunque su sentido probabilístico se pierde.

5. **Medida de Rogers-Tanimoto**

$$\frac{a+d}{a+d+2(b+c)} \tag{2.12}$$

Este coeficiente puede interpretarse como una extensión de la medida de parejas simples, pesando con el doble valor las no coincidencias.

6. **Medida de Kulcynski**

$$\frac{a}{b+c} \tag{2.13}$$

Esta medida muestra el cociente entre coincidencias y no coincidencias, excluyendo los pares $0-0$.

No son éstas las únicas medidas de este tipo que existen. Podíamos seguir citando muchas más y, entre ellas, a modo de ejemplo:

$$\begin{array}{cccc} \dfrac{a+d}{b+c} & \dfrac{a+d}{a+b+c} & \dfrac{2a}{2(a+d)+b+c} & \dfrac{2(a+d)}{2(a+d)+b+c} \\ \dfrac{2(a+d)}{2a+b+c} & \dfrac{a}{a+d+2(b+c)} & \dfrac{a}{a+2(b+c)} & \dfrac{a+d}{a+2(b+c)} \end{array} \tag{2.14}$$

2.3.4. Medidas basadas en probabilidades condicionadas.

Notemos que, de entre las medidas citadas con anterioridad, (2.8), (2.10) y (2.11) poseen interpretaciones probabilísticas razonables. Hay otras medidas que también poseen fundamentos probabilísticos.

Así, como ya se ha comentado con anterioridad, $\dfrac{a}{a+b}$ es la probabilidad condicionada de que un individuo elegido al azar presente el valor 1 en la variable X_j dado que ha presentado un 1 en la variable X_i. Asimismo, $\dfrac{a}{a+c}$ es la probabilidad condicionada de que un individuo elegido al azar presente un 1 en la variable X_i dado que lo ha presentado en la variable X_j.

Así podríamos pensar en una medida que marcara la probabilidad de que un individuo presente un 1 en una variable, dado que ha presentado un 1 en la otra, surgiendo la medida

$$\frac{1}{2}\left[\frac{a}{a+b}+\frac{a}{a+c}\right] \tag{2.15}$$

Como sabemos, no es claro que la codificación hecha sea la mejor. Por ello se puede optar por tener en cuenta también las otras coincidencias, dando lugar a la medida

$$\frac{1}{4}\left[\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right] \tag{2.16}$$

Estas expresiones son similares a las obtenidas a partir del coseno del ángulo entre variables en el caso de datos binarios, salvo que en lugar de tomar medias geométricas se toman medias aritméticas.

Por último se puede citar la medida de Hamann

$$\frac{2(a+d)}{a+b+c+d}-1=\frac{(a+d)-(b+c)}{a+b+c+d} \tag{2.17}$$

que indica la probabilidad de que un caso elegido al azar presente una coincidencia menos la probabilidad de que presente una diferencia en alguna de las variables.

2.4. Medidas de asociación entre individuos.

2.4.1. Distancias euclídea, de Minkowski y de Mahalanobis.

Consideremos ahora dos individuos tomados de la población, lo cual corresponde a tomar dos filas en la matriz de datos X:

$$x_i = (x_{i1}, \ldots, x_{in})'$$

$$x_j = (x_{j1}, \ldots, x_{jn})'$$

La métrica más conocida, que corresponde a la generalización a más de dos dimensiones de la distancia entre dos puntos en el plano, es la derivada de la norma $\mathbf{L}_2$ de un vector: [1]

$$||x_i||_2 = \sqrt{x_i' x_i} = \sqrt{\sum_{l=1}^{n} x_{il}^2}$$

obteniéndose, a partir de ella, la distancia euclídea

$$d_2(x_i, x_j) = ||x_i - x_j||_2 = \sqrt{(x_i - x_j)'(x_i - x_j)} = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2} \tag{2.18}$$

Esta métrica tiene la propiedad, al igual que la norma $\mathbf{L}_2$, de que todos sus valores son invariantes respecto de las transformaciones ortogonales $\widetilde{x}_i = \theta x_i$, donde θ es una matriz $n \times n$ que verifica $\theta'\theta = \theta\theta' = I$. En efecto:

$$||\theta x_i||_2 = \sqrt{x_i' \theta' \theta x_i} = \sqrt{x_i' x_i} = ||x_i||_2$$

y así se tiene

$$d_2(\theta x_i, \theta x_j) = d_2(x_i, x_j)$$

Además se verifica que estas transformaciones, además de las traslaciones, son las únicas para las cuales d_2 es invariante [2] .

En cuanto a las distancias de Minkowski, éstas proceden de las normas $\mathbf{L}_p$

$$||x_i||_p = \left(\sum_{l=1}^{n} |x_{il}|^p \right)^{\frac{1}{p}} \quad p \geq 1$$

dando origen a

$$d_p(x_i, x_j) = ||x_i - x_j||_p = \left(\sum_{l=1}^{n} |x_{il} - x_{jl}|^p \right)^{\frac{1}{p}} \tag{2.19}$$

Es fácil comprobar que esta distancia es invariante ante traslaciones, siendo éstas las únicas funciones para las cuales d_p posee esta propiedad.

Además se verifica la conocida relación

$$d_p(x_i, x_j) \leq d_q(x_i, x_j) \Leftrightarrow p \geq q$$

[1] Recordemos que dado un espacio vectorial X sobre un cuerpo K, una norma es una aplicación $||\cdot|| : X \longrightarrow K_0^+$ que verifica

1. $||x|| = 0 \Leftrightarrow x = 0$
2. $||\alpha x|| = |\alpha| ||x|| \quad \forall \alpha \in K \quad \forall x \in X$
3. $||x + y|| \leq ||x|| + ||y|| \quad \forall x, y \in X$

[2] En efecto, si consideramos $\widehat{x}_i = a + x_i$ y $\widehat{x}_j = a + x_j$, entonces se tiene:

$$d_2(\widehat{x}_i, \widehat{x}_j) = ||\widehat{x}_i - \widehat{x}_j||_2 = ||(a + x_i) - (a + x_j)||_2 = ||x_i - x_j||_2 = d_2(x_i, x_j)$$

Algunos casos particulares para valores de p concretos son [3]

1. Distancia d_1 o distancia ciudad (City Block) $(p = 1)$

$$d_1(x_i, x_j) = \sum_{l=1}^{n} |x_{il} - x_{jl}| \tag{2.20}$$

2. Distancia de Chebychev o distancia del máximo $(p = \infty)$

$$d_\infty(x_i, x_j) = \underset{l=1,\ldots,n}{\text{Max}} |x_{il} - x_{jl}| \tag{2.21}$$

Por otra parte, se puede generalizar la distancia euclídea, a partir de la norma

$$||x_i||_B = \sqrt{x_i' B x_i}$$

donde B es una matriz definida positiva. La métrica correspondiente a dicha norma es:

$$D_B(x_i, x_j) = \sqrt{(x_i - x_j)' B (x_i - x_j)} = \sqrt{\sum_{l=1}^{n}\sum_{h=1}^{n} b_{lh} x_{il} x_{jh}} \tag{2.22}$$

En el caso particular en que B sea una matriz diagonal, sus elementos son pesos positivos para las componentes del vector que corresponde a las variables en la matriz de datos.

Esta distancia se mantiene invariante frente a transformaciones (semejanzas) efectuadas por una matriz P que verifique $P'BP = B$. En efecto:

$$D_B(Px_i, Px_j) = \sqrt{(Px_i - Px_j)' B (Px_i - Px_j)} =$$

$$= \sqrt{(x_i - x_j)' P'BP (x_i - x_j)} = \sqrt{(x_i - x_j)' B (x_i - x_j)} = D_B(x_i, x_j)$$

La llamada métrica de Mahalanobis se obtiene tomando en 2.22 una matriz B determinada. Dicha matriz es la llamada matriz de varianzas-covarianzas de las variables (columnas de la matriz X de datos).

Los elementos de la matriz S, matriz de varianzas-covarianzas, se definen de la siguiente forma:

$$s_{uv} = \frac{1}{m}\sum_{l=1}^{m}(x_{lu} - \overline{x}_u)(x_{lv} - \overline{x}_v) \quad ; \quad u, v = 1, \ldots, n \tag{2.23}$$

Matricialmente tenemos dicha matriz expresada en la forma:

$$S = \frac{1}{m}\widetilde{X}'\widetilde{X} \quad \text{con} \quad \widetilde{X} = (\widetilde{x}_{ij}) \;\; ; \;\; \widetilde{x}_{ij} = x_{ij} - \overline{x}_j \;\; i = 1, \ldots, m \;\; ; \;\; j = 1, \ldots, n \tag{2.24}$$

A partir de la matriz S se puede definir la matriz de correlaciones, R, cuyos elementos son

$$\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} \quad ; \quad i, j = 1, \ldots, n$$

Notemos que si $m \geq n$, entonces la matriz de varianzas-covarianzas S es definida positiva y tiene sentido definir la distancia de Mahalanobis, para individuos, como:

$$D_S(x_i, x_j) = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)} \tag{2.25}$$

Esta distancia es invariante frente a transformaciones regidas por una matriz $C_{n\times n}$ no singular. En efecto, [4]

[3] Notemos que esta distancia generaliza a la distancia euclídea, en tanto en cuanto, esta última es un caso particular para $p = 2$.

[4] Notemos que la matriz de varianzas-covarianzas de las variables transformadas queda de la forma:

$$S = \frac{1}{m} C\widetilde{X}'\widetilde{X}C'$$

$$D_S(Cx_i, Cx_j) = \sqrt{(Cx_i - Cx_j)' \left[\frac{1}{m} C\widetilde{X}'\widetilde{X}C'\right]^{-1} (Cx_i - Cx_j)} =$$

$$= \sqrt{(x_i - x_j)' C'(C')^{-1} \left[\frac{1}{m}\widetilde{X}'\widetilde{X}\right]^{-1} C^{-1}C (x_i - x_j)} =$$

$$= \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)} = D_S(x_i, x_j)$$

Si, en particular, C es una matriz diagonal con los elementos no nulos, la transformación de X por C significa que el valor de cada variable en X es multiplicado por una constante, o sea, se ha hecho un cambio de escala. Por ello la métrica de Mahalanobis es invariante frente a cambios de escala, propiedad que no posee, por ejemplo, la métrica euclídea.

En la aplicación de las técnicas cluster la métrica de Mahalanobis presenta la desventaja de que el cálculo de la matriz S está basado en todos los individuos de forma conjunta y no trata, como sería de desear, de manera separada los objetos de cada cluster; además, su cálculo es mucho más laborioso que el de otras métricas. Por estas razones no suele emplearse en las técnicas cluster, si bien puede utilizarse dentro de cada cluster formado en una etapa determinada.

2.4.2. Correlación entre individuos.

Formalmente hablando, el coeficiente de correlación entre vectores de individuos puede ser usado como una medida de asociación entre individuos.

$$\text{Individuo i} \qquad x_i = (x_{i1}, x_{i2}, \ldots, x_{in})'$$

$$\text{Individuo j} \qquad x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})'$$

$$r_{ij} = \frac{\sum_{l=1}^{n} (x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{s_i s_j} \tag{2.26}$$

donde se ha definido

$$\overline{x}_h = \frac{1}{n}\sum_{l=1}^{n} x_{hl} \qquad h = i, j \quad \text{Media de cada individuo}$$

$$s_h^2 = \sum_{l=1}^{n} (x_{hl} - \overline{x}_h)^2 \qquad h = i, j \quad \text{Desviación cuadrática de cada individuo}$$

El principal problema de este coeficiente radica en el hecho de que en un vector de datos correspondiente a un individuo hay muchas unidades de medida diferentes, lo cual hace muy difícil comparar las *medias* y las *varianzas.*

No obstante, Cronbach y Gleser, en 1953, demostraron que este coeficiente posee un carácter métrico.

En efecto, sea x_{ik} el valor de la k-ésima variable sobre el i-ésimo individuo y transformemos ese dato en

$$\widehat{x}_{ik} = \frac{x_{ik} - \overline{x}_i}{s_i}$$

Entonces, la distancia euclídea al cuadrado entre dos individuos sobre los que se ha efectuado ese tipo de transformación será:

$$d_2^2(\widehat{x}_i, \widehat{x}_j) = \sum_{l=1}^{n} \left[\frac{x_{il} - \overline{x}_i}{s_i} - \frac{x_{jl} - \overline{x}_j}{s_j}\right]^2 =$$

$$= \sum_{l=1}^{n} \left[\frac{(x_{il} - \overline{x}_i)^2}{s_i^2} + \frac{(x_{jl} - \overline{x}_j)^2}{s_j^2} - 2\frac{(x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{s_i s_j}\right] = 2(1 - r_{ij})$$

Observemos que las dos medidas de la variable k-ésima, x_{ik} y x_{jk} son sometidas a transformaciones distintas

$$\widehat{x}_{ik} = \frac{x_{ik} - \overline{x}_i}{s_i}$$

$$\widehat{x}_{jk} = \frac{x_{jk} - \overline{x}_j}{s_j}$$

por lo que los nuevos valores no son comparables. Además, se observa que $1 - r$, complemento a uno del coeficiente de correlación, es una métrica (si $r_{ij} \longrightarrow 1 \Longrightarrow d(\widehat{x}_i, \widehat{x}_j) \longrightarrow 0$), pero lo es en el espacio en el que los datos se han transformado al tipificarlos.

Otra observación a hacer es que si se cambia la unidad de medida de una variable, cambia una componente en cada uno de los vectores de individuos: así si cambiamos la unidad de medida en la variable k-ésima, cambian los datos x_{ik} y x_{jk}; en consecuencia, cambian $\overline{x}_i$, $\overline{x}_j$, s_i y s_j y así cambia el coeficiente de correlación. Así pues, r_{ij}, es dependiente de cambios en unidades de medida. Es decir, estos cambios sopesan de manera distinta a las variables.

Por último, los valores de cada individuo pueden ser transformados de la siguiente manera

$$\widehat{x}_{ik} = \frac{x_{ik}}{\left(\sum_{l=1}^{n} x_{il}^2\right)^{\frac{1}{2}}}$$

Al igual que antes se puede demostrar, lo cual se deja como ejercicio al lector, que

$$d_2^2\left(\widehat{x}_i, \widehat{x}_j\right) = 2(1 - \cos(\alpha_{ij}))$$

donde

$$\cos(\alpha_{ij}) = \frac{\sum_{l=1}^{n} x_{il} x_{jl}}{\left(\sum_{l=1}^{n} x_{il}^2\right)^{\frac{1}{2}} \left(\sum_{l=1}^{n} x_{jl}^2\right)^{\frac{1}{2}}}$$

y, por lo tanto, $1 - \cos(\alpha_{ij})$ es una métrica.

2.4.3. Distancias derivadas de la distancia χ^2.

Hay muchas medidas de asociación que se basan en el estadístico χ^2, de uso familiar en el análisis de tablas de contingencia. Notemos

o_{ij} = valor observado en la celda i, j

e_{ij} = valor esperado bajo la hipótesis de independencia

Con dicha notación se define el estadístico χ^2 como

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{2.27}$$

donde p y q son el número de modalidades de las variables estudiadas.

Var **A**\Var **B**	1	$\cdots$	j	$\cdots$	q	
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1\cdot}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i\cdot}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot q}$	$n_{\cdot\cdot}$

(2.28)

Bajo la hipótesis de independencia de ambas variables, el valor esperado en la celda i, j es

$$e_{ij} = f_{i\cdot} f_{\cdot j} n_{\cdot\cdot} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$$

pero, por otra parte:

$$o_{ij} = n_{ij} = f_{ij} n_{..}$$

con lo cual

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij} - \dfrac{n_{i\cdot} n_{\cdot j}}{n_{..}}\right)^2}{\dfrac{n_{i\cdot} n_{\cdot j}}{n_{..}}} =$$

$$= n_{..} \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} n_{..} - f_{i\cdot} f_{\cdot j} n_{..})^2}{f_{i\cdot} f_{\cdot j}} = n_{..} \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2 n_{..}^2}{f_{i\cdot} f_{\cdot j} n_{..}^2} =$$

$$= n_{..} \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = n_{..} \left[\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{f_{ij}^2}{f_{i\cdot} f_{\cdot j}} - 1\right] =$$

$$= n_{..} \left[\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1\right]$$

Ahora bien, esta cantidad, que es muy útil para contrastes en tablas de contingencia, no lo es tanto como medida de asociación, puesto que aumenta cuando $n_{..}$ crece. Por ello se considera la medida Φ^2, llamada contingencia cuadrática media, definida como

$$\Phi^2 = \frac{\chi^2}{n_{..}} \tag{2.29}$$

Sin embargo, este coeficiente depende del tamaño de la tabla. Por ejemplo, supongamos que $p = q$ y que las variables están asociadas de forma perfecta, o sea, $n_{i\cdot} = n_{\cdot i} = n_{ii}\ \forall i$ (notemos que en tal caso sólo hay p casillas con valores distintos de cero). En este caso

$$\chi^2 = n_{..}(p-1)$$

$$\Phi^2 = p - 1$$

En el caso de una tabla rectangular con las variables perfectamente relacionadas, el número de casillas no nulas es $\text{Min}(p, q)$, por lo que

$$\chi^2 = n_{..} \,\text{Min}(p-1, q-1)$$

$$\Phi^2 = \text{Min}(p-1, q-1)$$

Con estas ideas en mente, se han hecho algunos intentos para normalizar la medida Φ^2 al rango $[0, 1]$. Por ejemplo:

$$\begin{aligned}
&\text{Medida de Tschuprow:} && T = \left(\frac{\Phi^2}{[(p-1)(q-1)]^{\frac{1}{2}}}\right)^{\frac{1}{2}} \\
&\text{Medida de Cramer:} && C = \left(\frac{\Phi^2}{\text{Min}(p-1, q-1)}\right)^{\frac{1}{2}} \\
&\text{Coeficiente de contingencia de Pearson:} && P = \left(\frac{\Phi^2}{1+\Phi^2}\right)^{\frac{1}{2}} = \left(\frac{\chi^2}{n_{..} + \chi^2}\right)^{\frac{1}{2}}
\end{aligned} \tag{2.30}$$

Obviamente, este tipo de medidas son empleadas en los casos en los que los datos que se poseen son conteos de frecuencias. Así, supongamos que tenemos m individuos sobre los que se han observado n variables. Sea x_{ij} la frecuencia observada de la j-ésima variable sobre el i-ésimo individuo.

	Var 1	$\cdots$	Var j	$\cdots$	Var n	
Ind. 1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1n}	$x_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Ind. i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{in}	$x_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Ind. m	x_{m1}	$\cdots$	x_{mj}	$\cdots$	x_{mn}	$x_{m\cdot}$
	$x_{\cdot 1}$	$\cdots$	$x_{\cdot j}$	$\cdots$	$x_{\cdot n}$	$x_{\cdot\cdot}$

Consideremos dos individuos x_i y x_j y sea la tabla $2 \times n$ formada a partir de ellos

	Var 1	$\cdots$	Var n	
Ind. i	x_{i1}	$\cdots$	x_{in}	$\sum_{l=1}^{n} x_{il}$
Ind. j	x_{j1}	$\cdots$	x_{jn}	$\sum_{l=1}^{n} x_{jl}$
	$x_{i1} + x_{j1}$	$\cdots$	$x_{in} + x_{jn}$	$\sum_{l=1}^{n} (x_{il} + x_{jl})$

Obviamente, cada individuo presenta un total de frecuencia marginal distinto ($x_{\cdot i}$ y $x_{\cdot j}$), por lo que no son comparables uno a uno. En este caso hay que buscar la semejanza teniendo en cuenta la proporcionalidad entre ambos. Por ello el empleo de distancias basadas en la distancia χ^2 es útil.

En nuestro caso, la forma que adopta el estadístico es:

$$\chi^2 = \sum_{l=1}^{n} \left[\frac{(x_{il} - e_{il})^2}{e_{il}} + \frac{(x_{jl} - e_{jl})^2}{e_{jl}} \right] \tag{2.31}$$

donde

$$e_{kh} = \frac{\sum_{l=1}^{n} x_{kl} (x_{ih} + x_{jh})}{\sum_{l=1}^{n} (x_{il} + x_{jl})} \quad ; \quad k = i, j \quad ; \quad h = 1, \cdots, n$$

y así, si $\chi^2 \longrightarrow 0$ se tiene la proporcionalidad buscada entre las dos filas y, por lo tanto, los dos individuos presentan el mismo perfil a lo largo de las variables, con lo cual dichos individuos serán parecidos.

2.4.4. Medidas no métricas: Coeficiente de Bray-Curtis.

Dados dos individuos

$$x_i = (x_{i1}, \ldots, x_{in})'$$

$$x_j = (x_{j1}, \ldots, x_{jn})'$$

el coeficiente de Bray-Curtis viene definido por la expresión

$$D_{i,j} = \frac{\sum_{l=1}^{n} |x_{il} - x_{jl}|}{\sum_{l=1}^{n} (x_{il} + x_{jl})} \tag{2.32}$$

El numerador no es otra cosa que la métrica $\mathbf{L}_1$, mientras que el denominador puede ser interpretado como una medida de la magnitud total de los dos individuos.

Hay que hacer notar que es aconsejable usar esta medida con datos no negativos, ya que pudiera haber cancelaciones en el denominador, pudiéndose obtener resultados poco aconsejables; por ejemplo, usando esta medida, no es aconsejable centrar los datos previamente. Además, puesto que para cada par de individuos se emplea un denominador distinto, esta medida no satisface siempre la desigualdad triangular.

2.4.5. Medidas para datos binarios.

Con alguna excepción, las medidas de asociación que se mencionaron para variables de tipo binario pueden ser aplicadas para medir la asociación entre individuos. En este caso la tabla de contingencia que se tiene es

Ind. I\Ind. J	1	0	Totales
1	a	b	$a+b$
0	c	d	$c+d$
Totales	$a+c$	$b+d$	$n=a+b+c+d$

(2.33)

Evidentemente, ahora a representa el número de veces que los individuos i y j presentan, de forma simultánea, un 1 sobre una misma variable.

Capítulo 3

Métodos Jerárquicos de Análisis Cluster.

3.1. Introducción.

Los llamados métodos jerárquicos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes.

1. Los métodos aglomerativos, también conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado.

2. Los métodos disociativos, también llamados descendentes, constituyen el proceso inverso al anterior. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

Para fijar ideas, centrémonos un segundo en los métodos aglomerativos. Sea n el conjunto de individuos de la muestra, de donde resulta el nivel $K = 0$, con n grupos. En el siguiente nivel se agruparán aquellos dos individuos que tengan la mayor similitud (o menor distancia), resultando así $n - 1$ grupos; a continuación, y siguiendo con la misma estrategia, se agruparán en el nivel posterior, aquellos dos individuos (o clusters ya formados) con menor distancia o mayor similitud; de esta forma, en el nivel L tendremos $n - L$ grupos formados. Si se continúa agrupando de esta forma, se llega al nivel $L = n - 1$ en el que sólo hay un grupo, formado por todos los individuos de la muestra.

Esta manera de formar nuevos grupos tiene la particularidad de que si en un determinado nivel se agrupan dos clusters, éstos quedan ya jerárquicamente agrupados para el resto de los niveles.

Los métodos jerárquicos permiten la construcción de un árbol de clasificación, que recibe el nombre de dendrograma (figura 3.1), en el cual se puede seguir de forma gráfica el procedimiento de unión seguido, mostrando que grupos se van uniendo, en que nivel concreto lo hacen, así como el valor de la medida de asociación entre los grupos cuando éstos se agrupan (valor que llamaremos nivel de fusión).

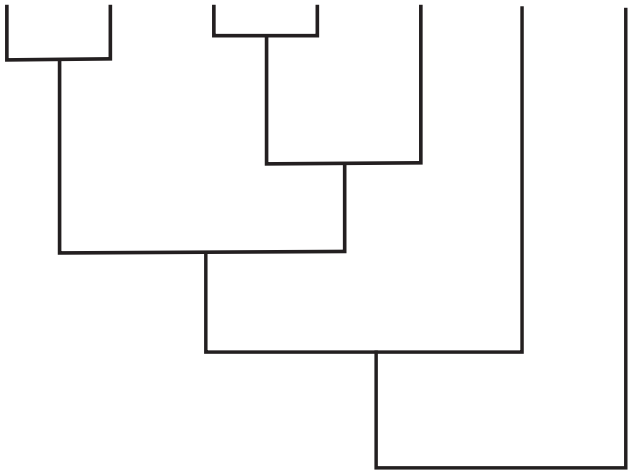
En resumen, la forma general de operar de estos métodos es bastante simple. Por ejemplo, en los métodos aglomerativos se parte de tantos grupos como individuos haya. A continuación se selecciona una medida de similitud, agrupándose los dos grupos o clusters con mayor similitud. Así se continúa hasta que:

1. Se forma un solo grupo.

2. Se alcanza el número de grupos prefijado.

3. Se detecta, a través de un contraste de significación, que hay razones estadísticas para no continuar agrupando clusters, ya que los más similares no son lo suficientemente homogéneos como para determinar una misma agrupación.

Figura 3.1: Dendrograma



3.2. Métodos Jerárquicos Aglomerativos.

A continuación vamos a presentar algunas de las estrategias que pueden ser empleadas a la hora de unir los clusters en las diversas etapas o niveles de un procedimiento jerárquico. Ninguno de estos procedimientos proporciona una solución óptima para todos los problemas que se pueden plantear, ya que es posible llegar a distintos resultados según el método elegido. El buen criterio del investigador, el conocimiento del problema planteado y la experiencia, sugerirán el método más adecuado. De todas formas, es conveniente, siempre, usar varios procedimientos con la idea de contrastar los resultados obtenidos y sacar conclusiones, tanto como si hubiera coincidencias en los resultados obtenidos con métodos distintos como si no las hubiera.

3.2.1. Estrategia de la distancia mínima o similitud máxima.

Esta estrategia recibe en la literatura anglosajona el nombre de *amalgamamiento simple (single linkage).* En este método se considera que la distancia o similitud entre dos clusters viene dada, respectivamente, por la mínima distancia (o máxima similitud) entre sus componentes.

Así, si tras efectuar la etapa K-ésima, tenemos ya formados $n-K$ clusters, la distancia entre los clusters C_i (con n_i elementos) y C_j (con n_j elementos) sería:

$$d(C_i,C_j) = \min_{\substack{x_l\in C_i\\ x_m\in C_j}} \{d(x_l,x_m)\} \quad l=1,\ldots,n_i \ ; \ m=1,\ldots,n_j \tag{3.1}$$

mientras que la similitud, si estuviéramos empleando una medida de tal tipo, entre los dos clusters sería:

$$s(C_i,C_j) = \max_{\substack{x_l\in C_i\\ x_m\in C_j}} \{s(x_l,x_m)\} \quad l=1,\ldots,n_i \ ; \ m=1,\ldots,n_j \tag{3.2}$$

Con ello, la estrategia seguida en el nivel $K+1$ será:

1. En el caso de emplear distancias, se unirán los clusters C_i y C_j si

$$d(C_i,C_j) = \min_{\substack{i_1,j_1=1,\ldots,n-K\\ i_1\neq j_1}} \{d(C_{i_1},C_{j_1})\} =$$

$$= \min_{\substack{i_1,j_1=1,\ldots,n-K\\ i_1\neq j_1}} \left\{ \min_{\substack{x_l\in C_{i_1}\\ x_m\in C_{j_1}}} \{d(x_l,x_m)\} \right\} \quad l=1,\ldots,n_{i_1} \ ; \ m=1,\ldots,n_{j_1}$$

2. En el caso de emplear similitudes, se unirán los clusters C_i y C_j si

$$s(C_i,C_j) = \max_{\substack{i_1,j_1=1,\ldots,n-K\\ i_1\neq j_1}} \{s(C_{i_1},C_{j_1})\} =$$

Aquí se calcula la distancia entre los objetos según la estrategia elegida (etapa)

Aquí se unen aquellos cluster con menor distancia al pasar a la siguiente etapa

$$= \underset{\substack{i_1,j_1=1,\ldots,n-K \\ i_1 \neq j_1}}{\text{Max}} \left\{ \underset{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}}{\text{Max}} \{s(x_l, x_m)\} \right\} \quad l = 1, \ldots, n_{i_1} \ ; \ m = 1, \ldots, n_{j_1}$$

donde, como es natural, se ha seguido la norma general de maximizar las similitudes o bien minimizar las distancias.

Ejemplo 3.1 *Partiendo de la matriz de distancias inicial entre 7 individuos*

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	2,15	0					
C	0,7	1,53	0				
D	1,07	1,14	0,43	0			
E	0,85	1,38	0,21	0,29	0		
F	1,16	1,01	0,55	0,22	0,41	0	
G	1,56	2,83	1,86	2,04	2,02	2,05	0

los pasos seguidos en un procedimiento cluster jerárquico ascendente, empleando la estrategia del amalgamamiento simple, serían los siguientes:

1. **Nivel K=1**

   $\text{Min}\{d(C_i, C_j)\} = d(C, E) = 0{,}21$, *por lo que el primer cluster que se forma es el cluster* (C, E).

2. **Nivel K=2**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**D**	**F**	**G**
A	0					
B	2,15	0				
(C,E)	0,7	1,38	0			
D	1,07	1,14	0,29	0		
F	1,16	1,01	0,41	0,22	0	
G	1,56	2,83	1,86	2,04	2,05	0

   Ahora bien, $\text{Min}\{d(C_i, C_j)\} = d(D, F) = 0{,}22$, *por lo que se forma el cluster* (D, F).

3. **Nivel K=3**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**(D,F)**	**G**
A	0				
B	2,15	0			
(C,E)	0,7	1,38	0		
(D,F)	1,07	1,01	0,29	0	
G	1,56	2,83	1,86	2,04	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{(C, E), (D, F)\} = 0{,}29$, *formándose el cluster* $((C, E), (D, F))$.

4. **Nivel K=4**

   La matriz de distancias en este paso es:

	A	**B**	**((C,E),(D,F))**	**G**
A	0			
B	2,15	0		
((C,E),(D,F))	0,7	1,01	0	
G	1,56	2,83	1,86	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{A, ((C, E), (D, F))\} = 0{,}7$, *formándose el cluster* $(A, ((C, E), (D, F)))$.

5. **Nivel K=5**

La matriz de distancias en este paso es:

	(A,((C,E),(D,F)))	**B**	**G**
(A,((C,E),(D,F)))	0		
B	1,01	0	
G	1,56	2,83	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{B, (A, ((C, E), (D, F)))\} = 1{,}01$, *formándose el cluster* $(B, (A, ((C, E), (D, F))))$.

6. **Nivel K=6**

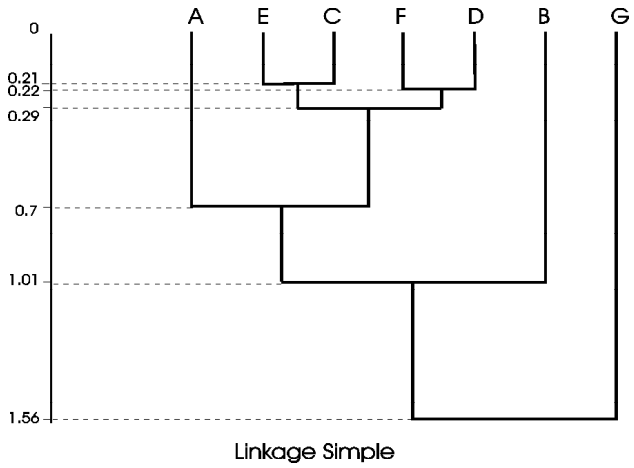
La matriz de distancias en este paso es:

	(B,(A,((C,E),(D,F))))	**G**
(B,(A,((C,E),(D,F))))	0	
G	1,56	0

Este será el último paso, en el cual, evidentemente, se tendrá un único cluster formado por los 7 individuos.

El dendrograma asociado es el de la figura 3.2

Figura 3.2: Método del amalgamamiento simple



3.2.2. Estrategia de la distancia máxima o similitud mínima.

En este método, también conocido como el procedimiento de *amalgamamiento completo (complete linkage)*, se considera que la distancia o similitud entre dos clusters hay que medirla atendiendo a sus elementos más dispares, o sea, la distancia o similitud entre clusters viene dada, respectivamente, por la máxima distancia (o mínima similitud) entre sus componentes.

Así pues, al igual que en la estrategia anterior, si estamos ya en la etapa K-ésima, y por lo tanto hay ya formados $n - K$ clusters, la distancia y similitud entre los clusters C_i y C_j (con n_i y n_j elementos respectivamente), serán:

$$d(C_i, C_j) = \max_{\substack{x_l \in C_i \\ x_m \in C_j}} \{d(x_l, x_m)\} \quad l = 1, \ldots, n_i \ ; \ m = 1, \ldots, n_j \tag{3.3}$$

$$s(C_i, C_j) = \min_{\substack{x_l \in C_i \\ x_m \in C_j}} \{s(x_l, x_m)\} \quad l = 1, \ldots, n_i \ ; \ m = 1, \ldots, n_j \tag{3.4}$$

y con ello, la estrategia seguida en el siguiente nivel, $K + 1$, será:

1. En el caso de emplear distancias, se unirán los clusters C_i y C_j si

$$d(C_i, C_j) = \underset{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}}{\text{Min}} \{d(C_{i_1}, C_{j_1})\} =$$

$$= \underset{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}}{\text{Min}} \left\{ \underset{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}}{\text{Max}} \{d(x_l, x_m)\} \right\} \quad l = 1, \ldots, n_{i_1} \ ; \ m = 1, \ldots, n_{j_1}$$

2. En el caso de emplear similitudes, se unirán los clusters C_i y C_j si

$$s(C_i, C_j) = \underset{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}}{\text{Max}} \{s(C_{i_1}, C_{j_1})\} =$$

$$= \underset{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}}{\text{Max}} \left\{ \underset{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}}{\text{Min}} \{s(x_l, x_m)\} \right\} \quad l = 1, \ldots, n_{i_1} \ ; \ m = 1, \ldots, n_{j_1}$$

Ejemplo 3.2 *En el mismo ejemplo anterior se tendrá:*

1. **Nivel K=1**

   $\text{Min}\{d(C_i, C_j)\} = d(C, E) = 0{,}21$, *por lo que el primer cluster que se forma es el cluster* (C, E).

2. **Nivel K=2**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**D**	**F**	**G**
A	0					
B	2,15	0				
(C,E)	0,85	1,53	0			
D	1,07	1,14	0,43	0		
F	1,16	1,01	0,55	0,22	0	
G	1,56	2,83	2,02	2,04	2,05	0

   Ahora bien, $\text{Min}\{d(C_i, C_j)\} = d(D, F) = 0{,}22$, *por lo que se forma el cluster* (D, F).

3. **Nivel K=3**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**(D,F)**	**G**
A	0				
B	2,15	0			
(C,E)	0,85	1,53	0		
(D,F)	1,16	1,14	0,55	0	
G	1,56	2,83	2,02	2,05	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{(C, E), (D, F)\} = 0{,}55$, *formándose el cluster* $((C, E), (D, F))$.

4. **Nivel K=4**

   La matriz de distancias en este paso es:

	A	**B**	**((C,E),(D,F))**	**G**
A	0			
B	2,15	0		
((C,E),(D,F))	1,16	1,53	0	
G	1,56	2,83	2,05	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{A, ((C, E), (D, F))\} = 1{,}16$,
   formándose el cluster $(A, ((C, E), (D, F)))$.

5. **Nivel K=5**

La matriz de distancias en este paso es:

	(A,((C,E),(D,F)))	**B**	**G**
(A,((C,E),(D,F)))	0		
B	2,15	0	
G	2,05	2,83	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{G, (A, ((C, E), (D, F)))\} = 2{,}05$*, formándose el cluster* $(G, (A, ((C, E), (D, F))))$.

6. **Nivel K=6**

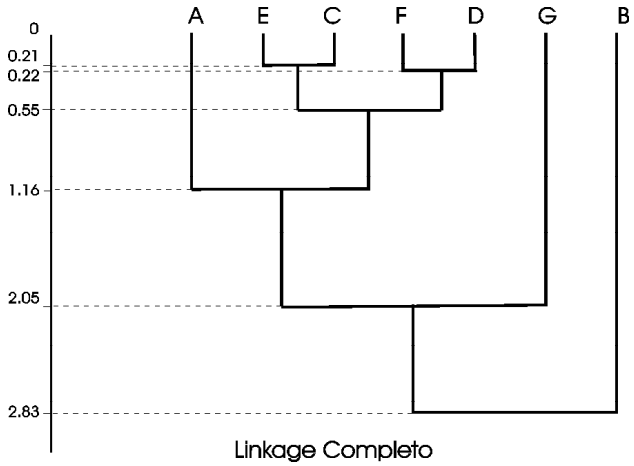
La matriz de distancias en este paso es:

	(G,(A,((C,E),(D,F))))	**B**
(G,(A,((C,E),(D,F))))	0	
B	2,83	0

Este será el último paso, en el cual, evidentemente, se tendrá un único cluster formado por los 7 individuos.

El dendrograma asociado es el de la figura 3.3

Figura 3.3: Método del amalgamamiento completo



3.2.3. Estrategia de la distancia, o similitud, promedio no ponderada. (Weighted aritmethic average)

En esta estrategia la distancia, o similitud, del cluster C_i con el C_j se obtiene como la media aritmética entre la distancia, o similitud, de las componentes de dichos clusters.

Así, si el cluster C_i (con n_i elementos) está compuesto, a su vez, por dos clusters C_{i_1} y C_{i_2} (con n_{i_1} y n_{i_2} elementos respectivamente), y el cluster C_j posee n_j elementos, la distancia, o similitud, entre ellos se calcula como

$$d(C_i, C_j) = \frac{d(C_{i_1}, C_j) + d(C_{i_2}, C_j)}{2} \tag{3.5}$$

Notemos que en este método no se tiene en cuenta el tamaño de ninguno de los clusters involucrados en el cálculo, lo cual significa que concede igual importancia a la distancia $d(C_{i_1}, C_j)$ que a la distancia $d(C_{i_2}, C_j)$.

Ejemplo 3.3 *Continuando con el ejemplo anterior, ahora tendremos:*

1. **Nivel K=1**

   $\text{Min}\{d(C_i, C_j)\} = d(C, E) = 0{,}21$, *por lo que el primer cluster que se forma es el cluster* (C, E).

2. **Nivel K=2**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**D**	**F**	**G**
A	0					
B	2,15	0				
(C,E)	0,775	1,455	0			
D	1,07	1,14	0,36	0		
F	1,16	1,01	0,48	0,22	0	
G	1,56	2,83	1,94	2,04	2,05	0

   Ahora bien, $\text{Min}\{d(C_i, C_j)\} = d(D, F) = 0{,}22$, *por lo que se forma el cluster* (D, F).

3. **Nivel K=3**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**(D,F)**	**G**
A	0				
B	2,15	0			
(C,E)	0,775	1,455	0		
(D,F)	1,115	1,075	0,42	0	
G	1,56	2,83	1,94	2,045	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{(C, E), (D, F)\} = 0{,}42$, *formándose el cluster* $((C, E), (D, F))$.

4. **Nivel K=4**

   La matriz de distancias en este paso es:

	A	**B**	**((C,E),(D,F))**	**G**
A	0			
B	2,15	0		
((C,E),(D,F))	0,945	1,265	0	
G	1,56	2,83	1,9925	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{A, ((C, E), (D, F))\} = 0{,}945$, *formándose el cluster* $(A, ((C, E), (D, F)))$.

5. **Nivel K=5**

   La matriz de distancias en este paso es:

	(A,((C,E),(D,F)))	**B**	**G**
(A,((C,E),(D,F)))	0		
B	1,7075	0	
G	1,77625	2,83	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{B, (A, ((C, E), (D, F)))\} = 1{,}7075$, *formándose el cluster* $(B, (A, ((C, E), (D, F))))$.

6. **Nivel K=6**

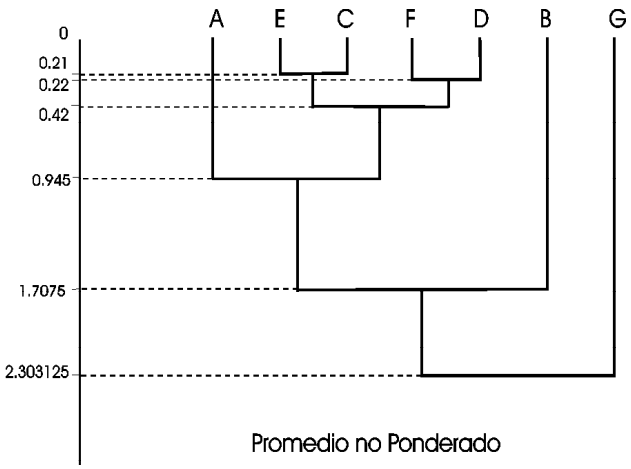
   La matriz de distancias en este paso es:

	(B,(A,((C,E),(D,F))))	**G**
(B,(A,((C,E),(D,F))))	0	
G	2,303125	0

   Este será el último paso, en el cual, evidentemente, se tendrá un único cluster formado por los 7 individuos.

El dendrograma asociado a este ejemplo es el de la figura 3.4

Figura 3.4: Método del promedio no ponderado



3.2.4. Estrategia de la distancia, o similitud, promedio ponderada. (unweighted aritmethic average)

Se considera que la distancia, o similitud, entre dos clusters, viene definida por el promedio ponderado de las distancias, o similitudes, de los componentes de un cluster respecto a los del otro.

Sea dos clusters, C_i y C_j; supongamos que el cluster C_i está formado, a su vez, por otros dos clusters, C_{i_1} y C_{i_2}, con n_{i_1} y n_{i_2} elementos respectivamente. Sea $n_i = n_{i_1} + n_{i_2}$ el número de elementos de C_i y n_j el número de elementos que componen C_j. Entonces, en términos de distancias (igual puede hacerse para similitudes), la distancia promedio ponderada sería, notando $x_i \in C_i$, $x_{i_1} \in C_{i_1}$, $x_{i_2} \in C_{i_2}$, $x_j \in C_j$

$$d(C_i, C_j) = \frac{1}{(n_{i_1} + n_{i_2})n_j} \sum_{i=1}^{n_{i_1}+n_{i_2}} \sum_{j=1}^{n_j} d(x_i, x_j) =$$

$$= \frac{1}{(n_{i_1} + n_{i_2})n_j} \sum_{i_1=1}^{n_{i_1}} \sum_{j=1}^{n_j} d(x_{i_1}, x_j) + \frac{1}{(n_{i_1} + n_{i_2})n_j} \sum_{i_2=1}^{n_{i_2}} \sum_{j=1}^{n_j} d(x_{i_2}, x_j) =$$

$$= \frac{n_{i_1}}{(n_{i_1} + n_{i_2})n_{i_1} n_j} \sum_{i_1=1}^{n_{i_1}} \sum_{j=1}^{n_j} d(x_{i_1}, x_j) + \frac{n_{i_2}}{(n_{i_1} + n_{i_2})n_{i_2} n_j} \sum_{i_2=1}^{n_{i_2}} \sum_{j=1}^{n_j} d(x_{i_2}, x_j) =$$

$$= \frac{n_{i_1}}{n_{i_1} + n_{i_2}} d(C_{i_1}, C_j) + \frac{n_{i_2}}{n_{i_1} + n_{i_2}} d(C_{i_2}, C_j) =$$

$$= \frac{n_{i_1} d(C_{i_1}, C_j) + n_{i_2} d(C_{i_2}, C_j)}{n_{i_1} + n_{i_2}} \tag{3.6}$$

con lo cual la distancia $d(C_i, C_j)$ es el promedio ponderado de las distancias de cada uno de los dos clusters previos, C_{i_1} y C_{i_2}, con respecto al cluster C_j.

Ejercicio 3.1 *Comprobar que, con la estrategia de la distancia promedio ponderada, se tiene*

$$d([(a,b),c],[(d,e),f]) = \frac{2d\left((d,e),[(a,b),c]\right) + d\left(f,[(a,b),c]\right)}{3} =$$

$$= \frac{d(a,d) + d(a,e) + d(a,f) + d(b,d) + d(b,e) + d(b,f) + d(c,d) + d(c,e) + d(c,f)}{9}$$

Ejemplo 3.4 *Siguiendo con el ejemplo tratado anteriormente, ahora tendremos:*

1. **Nivel K=1**

   $\text{Min}\{d(C_i, C_j)\} = d(C, E) = 0{,}21$, *por lo que el primer cluster que se forma es el cluster* (C, E).

2. **Nivel K=2**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**D**	**F**	**G**
A	0					
B	2,15	0				
(C,E)	0,775	1,455	0			
D	1,07	1,14	0,36	0		
F	1,16	1,01	0,48	0,22	0	
G	1,56	2,83	1,94	2,04	2,05	0

   Ahora bien, $\text{Min}\{d(C_i, C_j)\} = d(D, F) = 0{,}22$*, por lo que se forma el cluster* (D, F).

3. **Nivel K=3**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**(D,F)**	**G**
A	0				
B	2,15	0			
(C,E)	0,775	1,455	0		
(D,F)	1,115	1,075	0,42	0	
G	1,56	2,83	1,94	2,045	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{(C, E), (D, F)\} = 0{,}42$*, formándose el cluster* $((C, E), (D, F))$.

4. **Nivel K=4**

   La matriz de distancias en este paso es:

	A	**B**	**((C,E),(D,F))**	**G**
A	0			
B	2,15	0		
((C,E),(D,F))	0,945	1,265	0	
G	1,56	2,83	1,9925	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{A, ((C, E), (D, F))\} = 0{,}945$*, formándose el cluster* $(A, ((C, E), (D, F)))$.

5. **Nivel K=5**

   La matriz de distancias en este paso es:

	(A,((C,E),(D,F)))	**B**	**G**
(A,((C,E),(D,F)))	0		
B	1,442	0	
G	1,906	2,83	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{B, (A, ((C, E), (D, F)))\} = 1{,}442$*, formándose el cluster* $(B, (A, ((C, E), (D, F))))$.

6. **Nivel K=6**

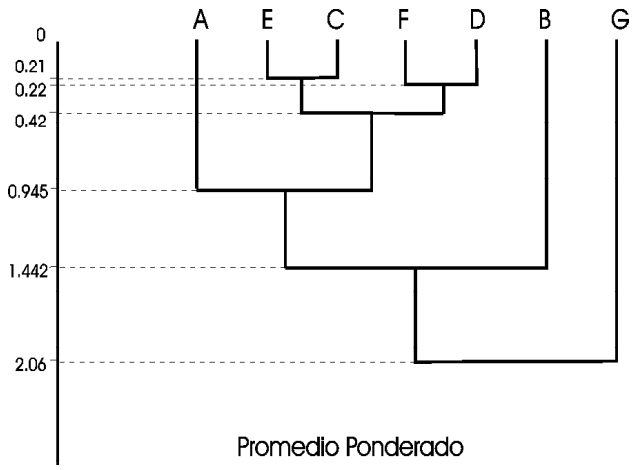
   La matriz de distancias en este paso es:

	(B,(A,((C,E),(D,F))))	**G**
(B,(A,((C,E),(D,F))))	0	
G	2,06	0

   Este será el último paso, en el cual, evidentemente, se tendrá un único cluster formado por los 7 individuos.

El dendrograma asociado a este ejemplo es el de la figura 3.5

Figura 3.5: Método del promedio ponderado



3.2.5. Métodos basados en el centroide.

En estos métodos, la semejanza entre dos clusters viene dada por la semejanza entre sus centroides, esto es, los vectores de medias de las variables medidas sobre los individuos del cluster.

Entre ellos distinguiremos dos:

1. Método del centroide ponderado, en el que los tamaños de los clusters son considerados a la hora de efectuar los cálculos.
2. Método del centroide no ponderado, o método de la mediana, en el cual los tamaños de los clusters no son considerados.

Veamos cada uno de ellos por separado:

1. En cuanto al primero de ellos y centrándonos en la distancia euclídea al cuadrado, supongamos que pretendemos medir la distancia entre los clusters C_j (compuesto por n_j elementos) y C_i (formado a su vez por dos clusters, C_{i_1} y C_{i_2}, con n_{i_1} y n_{i_2} elementos, respectivamente). Sean m^j, m^{i_1} y m^{i_2} los centroides de los clusters anteriormente citados (obviamente, esos centroides son vectores n dimensionales).

   Así, el centroide del cluster C_i vendrá dado en notación vectorial por:

$$m^i = \frac{n_{i_1} m^{i_1} + n_{i_2} m^{i_2}}{n_{i_1} + n_{i_2}}$$

   cuyas componentes serán:

$$m_l^i = \frac{n_{i_1} m_l^{i_1} + n_{i_2} m_l^{i_2}}{n_{i_1} + n_{i_2}} \quad l = 1, \ldots, n$$

   Con ello, la distancia euclídea al cuadrado entre los clusters C_i y C_j vendrá dada por:

$$d_2^2(C_j, C_i) = \sum_{l=1}^{n} \left(m_l^j - m_l^i \right)^2 = \sum_{l=1}^{n} \left[m_l^j - \frac{n_{i_1} m_l^{i_1} + n_{i_2} m_l^{i_2}}{n_{i_1} + n_{i_2}} \right]^2 =$$

$$= \sum_{l=1}^{n} \left[(m_l^j)^2 - 2 m_l^j \frac{n_{i_1} m_l^{i_1} + n_{i_2} m_l^{i_2}}{n_{i_1} + n_{i_2}} + \right.$$

$$\left. + \frac{(n_{i_1})^2 (m_l^{i_1})^2 + (n_{i_2})^2 (m_l^{i_2})^2 + n_{i_1} n_{i_2} (m_l^{i_1})^2 + n_{i_1} n_{i_2} (m_l^{i_2})^2}{(n_{i_1} + n_{i_2})^2} + \right.$$

$$+\frac{-n_{i_1}n_{i_2}(m_l^{i_1})^2 - n_{i_1}n_{i_2}(m_l^{i_2})^2 + 2n_{i_1}n_{i_2}m_l^{i_1}m_l^{i_2}}{(n_{i_1}+n_{i_2})^2}\Bigg] =$$

$$= \sum_{l=1}^{n}\Bigg[(m_l^j)^2 - 2m_l^j\frac{n_{i_1}m_l^{i_1}+n_{i_2}m_l^{i_2}}{n_{i_1}+n_{i_2}} +$$

$$+\frac{(n_{i_1}+n_{i_2})n_{i_1}(m_l^{i_1})^2 + (n_{i_1}+n_{i_2})n_{i_2}(m_l^{i_2})^2}{(n_{i_1}+n_{i_2})^2} -$$

$$-\frac{n_{i_1}n_{i_2}}{(n_{i_1}+n_{i_2})^2}\left[(m_l^{i_1})^2 + (m_l^{i_2})^2 - 2m_l^{i_1}m_l^{i_2}\right]\Bigg] =$$

$$= \sum_{l=1}^{n}\Bigg[\frac{n_{i_1}(m_l^j)^2 + n_{i_2}(m_l^j)^2}{n_{i_1}+n_{i_2}} + \frac{n_{i_1}(m_l^{i_1})^2}{n_{i_1}+n_{i_2}} + \frac{n_{i_2}(m_l^{i_2})^2}{n_{i_1}+n_{i_2}} -$$

$$-2m_l^j\frac{n_{i_1}m_l^{i_1}}{n_{i_1}+n_{i_2}} - 2m_l^j\frac{n_{i_2}m_l^{i_2}}{n_{i_1}+n_{i_2}} - \frac{n_{i_1}n_{i_2}}{(n_{i_1}+n_{i_2})^2}\left[m_l^{i_1}-m_l^{i_2}\right]^2\Bigg] =$$

$$= \sum_{l=1}^{n}\Bigg[\frac{n_{i_1}}{n_{i_1}+n_{i_2}}\left[m_l^j - m_l^{i_1}\right]^2 + \frac{n_{i_2}}{n_{i_1}+n_{i_2}}\left[m_l^j - m_l^{i_2}\right]^2 -$$

$$-\frac{n_{i_1}n_{i_2}}{(n_{i_1}+n_{i_2})^2}\left[m_l^{i_1}-m_l^{i_2}\right]^2\Bigg] =$$

$$= \frac{n_{i_1}}{n_{i_1}+n_{i_2}}d_2^2(C_{i_1},C_j) + \frac{n_{i_2}}{n_{i_1}+n_{i_2}}d_2^2(C_{i_2},C_j) - \frac{n_{i_1}n_{i_2}}{(n_{i_1}+n_{i_2})^2}d_2^2(C_{i_1},C_{i_2}) \tag{3.7}$$

Comentario 3.2.1 *Nótese que la relación anterior se ha establecido para el caso particular de la distancia euclídea. No obstante, dicha relación se sigue verificando si la distancia empleada viene definida a partir de una norma que proceda de un producto escalar.* [1]

Esta hipótesis puede relajarse aún más hasta considerar distancias que procedan de una norma que verifique la ley del paralelogramo

$$||x+y||^2 + ||x-y||^2 = 2\left[||x||^2 + ||y||^2\right]$$

ya que en tales circunstancias se puede definir un producto escalar a partir de ella como

$$< x, y >= \frac{1}{4}\left[||x+y||^2 - ||x-y||^2\right]$$

2. Una desventaja del procedimiento anterior estriba en que si los tamaños n_{i_1} y n_{i_2} de los componentes de C_i son muy diferentes entre sí, se corre el peligro de que el centroide de dicho cluster, m^i, esté influenciado excesivamente por el componente con tamaño superior y las cualidades del grupo pequeño no se tengan prácticamente en cuenta.

   Así la estrategia de la distancia mediana, al considerar de forma arbitraria que $n_{i_1} = n_{i_2}$, provoca que el centroide del cluster C_i esté situado entre los clusters C_{i_1} y C_{i_2} y con ello el centroide del cluster (C_i, C_j) esté localizado en el punto central o mediana del triángulo formado por los clusters C_{i_1}, C_{i_2} y C_j.

   Salvo esta diferencia, la estrategia de la distancia mediana es análoga a la anterior y, por lo tanto, goza de sus mismas características. Así, si estamos hablando de distancias, la distancia entre el cluster C_i y el C_j viene dada por

[1] Dado un producto escalar en un espacio vectorial, se puede definir la norma de un vector como la raíz cuadrada positiva del producto escalar del vector por sí mismo.

$$d(C_i, C_j) = \frac{1}{2}\left[d(C_{i_1}, C_j) + d(C_{i_2}, C_j)\right] - \frac{1}{4}d(C_{i_1}, C_{i_2})$$

y si hablamos de similitudes

$$s(C_i, C_j) = \frac{1}{2}\left[s(C_{i_1}, C_j) + s(C_{i_2}, C_j)\right] + \frac{1}{4}\left[1 - s(C_{i_1}, C_{i_2})\right]$$

Notemos que una característica de los métodos basados en el centroide y sus variantes es que el valor de similaridad o la distancia asociada con los clusters enlazados puede aumentar o disminuir de una etapa a otra. Por ejemplo, cuando la medida es una distancia, la distancia entre los centroides puede ser menor que la de otro par de centroides unidos en una etapa anterior. Esto puede ocurrir ya que los centroides, en cada etapa, pueden cambiar de lugar. Este problema puede llevar a que el dendrograma resultante sea complicado de interpretar.

Ejemplo 3.5 *Consideremos los siguientes individuos sobre los cuales se han medido dos variables y apliquemos los métodos del centroide ponderado y el de la mediana, empleando para ello la distancia euclídea al cuadrado.*

Ver en transparencia

Individuo	X_1	X_2
A	10	5
B	20	20
C	30	10
D	30	15
E	5	10

Método del Centroide Ponderado.

1. **Nivel 1:**

   La matriz inicial de distancias es

	A	**B**	**C**	**D**	**E**
A	0				
B	325	0			
C	425	200	0		
D	500	125	25	0	
E	50	325	625	650	0

   A la vista de esta matriz se unen los individuos C y D. El centroide del cluster (C, D) es $(30, 12{,}5)$.

2. **Nivel 2:**

   La matriz de distancias en este paso es

	A	**B**	**(C,D)**	**E**
A	0			
B	325	0		
(C,D)	456,25	156,25	0	
E	50	325	631,25	0

   uniéndose en este nivel los individuos A y E. El centroide del cluster (A, E) es $(7{,}5, 7{,}5)$.

3. **Nivel 3:**

   La matriz de distancias en este nivel es

	(A,E)	**B**	**(C,D)**
(A,E)	0		
B	312,5	0	
(C,D)	531,25	156,25	0

   En este nivel se unen los clusters (C, D) y B. El centroide del cluster (B, C, D) es $(26{,}66, 15)$.

4. **Nivel 4:**

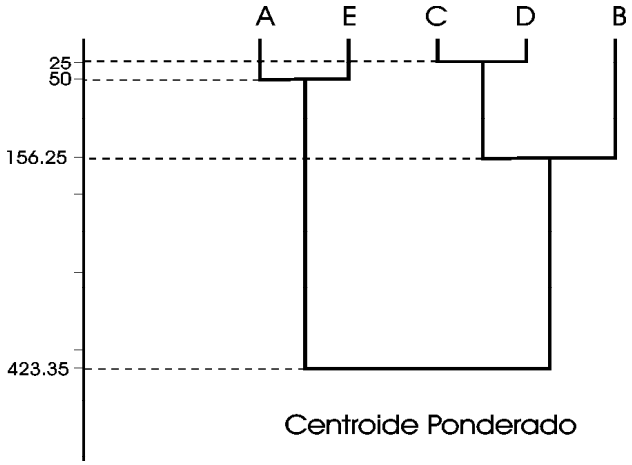
   La matriz de distancias en este nivel es

	(A,E)	**(B,C,D)**
(A,E)	0	
(B,C,D)	423,35	0

completándose así la jerarquía. El centroide es el punto $(19, 12)$.

El dendrograma asociado es el de la figura 3.6

Figura 3.6: Método del centroide ponderado



Método de la mediana.

1. **Nivel 1:**

   La matriz inicial de distancias es

	A	**B**	**C**	**D**	**E**
A	0				
B	325	0			
C	425	200	0		
D	500	125	25	0	
E	50	325	625	650	0

Ver en transparencia

A la vista de esta matriz se unen los individuos C *y* D. *El centroide del cluster* (C, D) *es* $(30, 12{,}5)$.

2. **Nivel 2:**

   La matriz de distancias en este paso es

	A	**B**	**(C,D)**	**E**
A	0			
B	325	0		
(C,D)	456,25	156,25	0	
E	50	325	631,25	0

uniéndose en este nivel los individuos A *y* E. *El centroide del cluster* (A, E) *es* $(7{,}5, 7{,}5)$.

3. **Nivel 3:**

   La matriz de distancias en este nivel es

	(A,E)	B	(C,D)
(A,E)	0		
B	312,5	0	
(C,D)	531,25	156,25	0

En este nivel se unen los clusters (C, D) *y* B. *El centroide del cluster* (B, C, D) *es* $(25, 16{,}25)$.

4. **Nivel 4:**

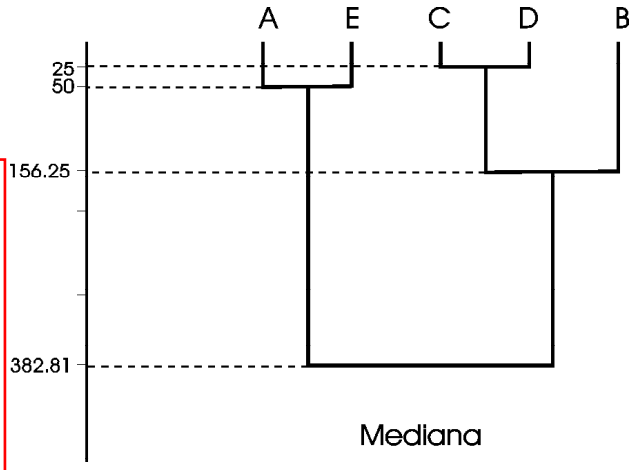
La matriz de distancias en este nivel es

	(A,E)	(B,C,D)
(A,E)	0	
(B,C,D)	382,81	0

completándose así la jerarquía. El centroide es el punto $(16{,}25, 11{,}875)$

El dendrograma asociado es el de la figura 3.7

Figura 3.7: Método de la mediana



Esta tabla mide los valores de 2 variables sobre 5 individuos

	X1	X2
I1	1	3
I2	2	4
I3	5	1
I4	7	3
I5	7	2

Utilizar (distancia euclídea)^2

3.2.6. Método de Ward.

El método de Ward es un procedimiento jerárquico en el cual, en cada etapa, se unen los dos clusters para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del cluster.

Notemos por

- x_{ij}^k al valor de la $j-$ésima variable sobre el $i-$ésimo individuo del $k-$ésimo cluster, suponiendo que dicho cluster posee n_k individuos.
- m^k al centroide del cluster k, con componentes m_j^k.
- E_k a la suma de cuadrados de los errores del cluster k, o sea, la distancia euclídea al cuadrado entre cada individuo del cluster k a su centroide

$$E_k = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k - m_j^k)^2 = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k)^2 - n_k\sum_{j=1}^{n}(m_j^k)^2$$

- E a la suma de cuadrados de los errores para todos los clusters, o sea, si suponemos que hay h clusters

$$E = \sum_{k=1}^{h} E_k$$

El proceso comienza con m clusters, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del cluster y por lo tanto en este primer paso se tendrá $E_k = 0$ para cada cluster y con ello, $E = 0$. El objetivo del método de Ward es encontrar en cada etapa aquellos dos clusters cuya unión proporcione el menor incremento en la suma total de errores, E.

Supongamos ahora que los clusters C_p y C_q se unen resultando un nuevo cluster C_t. Entonces el incremento de E será

$$\Delta E_{pq} = E_t - E_p - E_q =$$

$$= \left[\sum_{i=1}^{n_t}\sum_{j=1}^{n}(x_{ij}^t)^2 - n_t\sum_{j=1}^{n}(m_j^t)^2\right] - \left[\sum_{i=1}^{n_p}\sum_{j=1}^{n}(x_{ij}^p)^2 - n_p\sum_{j=1}^{n}(m_j^p)^2\right] - \left[\sum_{i=1}^{n_q}\sum_{j=1}^{n}(x_{ij}^q)^2 - n_q\sum_{j=1}^{n}(m_j^q)^2\right] =$$

$$= n_p\sum_{j=1}^{n}(m_j^p)^2 + n_q\sum_{j=1}^{n}(m_j^q)^2 - n_t\sum_{j=1}^{n}(m_j^t)^2$$

Ahora bien

$$n_t m_j^t = n_p m_j^p + n_q m_j^q$$

de donde

$$n_t^2(m_j^t)^2 = n_p^2(m_j^p)^2 + n_q^2(m_j^q)^2 + 2n_p n_q m_j^p m_j^q$$

y como

$$2m_j^p m_j^q = (m_j^p)^2 + (m_j^q)^2 - (m_j^p - m_j^q)^2$$

se tiene

$$n_t^2(m_j^t)^2 = n_p(n_p + n_q)(m_j^p)^2 + n_q(n_p + n_q)(m_j^q)^2 - n_p n_q(m_j^p - m_j^q)^2$$

Dado que $n_t = n_p + n_q$, dividiendo por n_t^2 se obtiene

$$(m_j^t)^2 = \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

con lo cual se obtiene la siguiente expresión de ΔE_{pq}:

$$n_p\sum_{j=1}^{n}(m_j^p)^2 + n_q\sum_{j=1}^{n}(m_j^q)^2 - n_t\sum_{j=1}^{n}\left[\frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2\right]$$

$$\Delta E_{pq} = n_p\sum_{j=1}^{n}(m_j^p)^2 + n_q\sum_{j=1}^{n}(m_j^q)^2 - m_p\sum_{j=1}^{n}(m_j^p)^2 - n_q\sum_{j=1}^{n}(m_j^q)^2 + \frac{n_p n_q}{n_t}\sum_{j=1}^{n}(m_j^p - m_j^q)^2$$

$$= \frac{n_p n_q}{n_t}\sum_{j=1}^{n}(m_j^p - m_j^q)^2$$

Así el menor incremento de los errores cuadráticos es proporcional a la distancia euclídea al cuadrado de los centroides de los clusters unidos. La suma E es no decreciente y el método, por lo tanto, no presenta los problemas de los métodos del centroide anteriores.

Veamos, para finalizar, cómo se pueden calcular los distintos incrementos a partir de otros calculados con anterioridad.

Sea C_t el cluster resultado de unir C_p y C_q y sea C_r otro cluster distinto a los otros dos. El incremento potencial en E que se produciría con la unión de C_r y C_t es

$$\Delta E_{rt} = \frac{n_r n_t}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^t)^2$$

Teniendo en cuenta que

$$\begin{aligned} m_j^t = & \quad \frac{n_p m_j^p + n_q m_j^q}{n_t} \\ n_t = & \quad n_p + n_q \end{aligned}$$

y la expresión

$$(m_j^t)^2 = \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

se deduce

$$(m_j^r - m_j^t)^2 = (m_j^r)^2 + (m_j^t)^2 - 2m_j^r m_j^t =$$

$$= (m_j^r)^2 + \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2 - 2m_j^r \frac{n_p m_j^p + n_q m_j^q}{n_t} =$$

$$= \frac{n_p(m_j^r)^2 + n_q(m_j^r)^2}{n_t} + \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 -$$

$$- \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2 - 2m_j^r \frac{n_p m_j^p + n_q m_j^q}{n_t} =$$

$$= \frac{n_p}{n_t}(m_j^r - m_j^p)^2 + \frac{n_q}{n_t}(m_j^r - m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

con lo cual

$$\Delta E_{rt} = \frac{n_r n_t}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^t)^2 =$$

$$= \frac{n_r n_t}{n_r + n_t}\sum_{j=1}^{n}\left[\frac{n_p}{n_t}(m_j^r - m_j^p)^2 + \frac{n_q}{n_t}(m_j^r - m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2\right] =$$

$$= \frac{n_r n_p}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^p)^2 + \frac{n_q n_r}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^q)^2 - \frac{n_r n_p n_q}{n_t(n_r + n_t)}\sum_{j=1}^{n}(m_j^p - m_j^q)^2 =$$

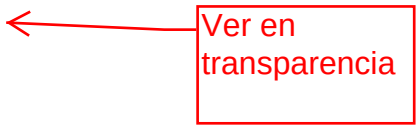
$$= \frac{1}{n_r + n_t}\sum_{j=1}^{n}\left[n_r n_p(m_j^r - m_j^p)^2 + n_r n_q(m_j^r - m_j^q)^2 - \frac{n_r n_p n_q}{n_p + n_q}(m_j^p - m_j^q)^2\right] =$$

$$= \frac{1}{n_r + n_t}\left[(n_r + n_p)\Delta E_{rp} + (n_r + n_q)\Delta E_{rq} - n_r \Delta E_{pq}\right]$$

Al igual que en los anteriores métodos del centroide se puede demostrar que la relación anterior se sigue verificando para una distancia que venga definida a partir de una norma que proceda de un producto escalar o que verifique la ley del paralelogramo.

Ejemplo 3.6 *Veamos cómo funciona este procedimiento en el caso de 5 individuos sobre los cuales se miden dos variables. Los datos son los siguientes*

Individuo	X_1	X_2
A	10	5
B	20	20
C	30	10
D	30	15
E	5	10



Nivel 1

En este primer paso hemos de calcular las $\binom{5}{2} = 10$ *posibles combinaciones.*

Partición	*Centroides*	E_k	E	ΔE
$(A,B),C,D,E$	$C_{AB}=(15,12{,}5)$	$E_{AB}=162{,}5$ $E_C=E_D=E_E=0$	162,5	162,5
$(A,C),B,D,E$	$C_{AC}=(20,7{,}5)$	$E_{AC}=212{,}5$ $E_B=E_D=E_E=0$	212,5	212,5
$(A,D),B,C,E$	$C_{AD}=(20,10)$	$E_{AD}=250$ $E_B=E_C=E_E=0$	250	250
$(A,E),B,C,D$	$C_{AE}=(7{,}5,7{,}5)$	$E_{AE}=25$ $E_B=E_C=E_D=0$	25	25
$(B,C),A,D,E$	$C_{BC}=(25,15)$	$E_{BC}=100$ $E_A=E_D=E_E=0$	100	100
$(B,D),A,C,E$	$C_{BD}=(25,17{,}5)$	$E_{BD}=62{,}5$ $E_A=E_C=E_E=0$	62,5	62,5
$(B,E),A,C,D$	$C_{BE}=(12{,}5,15)$	$E_{BE}=162{,}5$ $E_A=E_C=E_D=0$	162,5	162,5
$(C,D),A,B,E$	$C_{CD}=(30,12{,}5)$	$E_{CD}=12{,}5$ $E_A=E_B=E_E=0$	12,5	12,5
$(C,E),A,B,D$	$C_{CE}=(17{,}5;10)$	$E_{CE}=312{,}5$ $E_A=E_B=E_D=0$	312,5	312,5
$(D,E),A,B,C$	$C_{DE}=(17{,}5;12{,}5)$	$E_{DE}=325$ $E_A=E_B=E_C=0$	325	325

de donde se deduce que en esta etapa se unen los elementos C y D. La configuración actual es $(C,D),A,B,E$.

Nivel 2

A partir de la configuración actual tomamos las $\binom{4}{2}=6$ combinaciones posibles.

Partición	*Centroides*	E_k	E	ΔE
$(A,C,D),B,E$	$C_{ACD}=(23{,}33,10)$	$E_{ACD}=316{,}66$ $E_B=E_E=0$	316,66	304,16
$(B,C,D),A,E$	$C_{BCD}=(26{,}66,15)$	$E_{BCD}=116{,}66$ $E_A=E_E=0$	116,66	104,16
$(C,D,E),A,B$	$C_{CDE}=(21{,}66,11{,}66)$	$E_{CDE}=433{,}33$ $E_A=E_B=0$	433,33	420,83
$(A,B),(C,D),E$	$C_{AB}=(15,12{,}5)$ $C_{CD}=(30,12{,}5)$	$E_{AB}=162{,}5$ $E_{CD}=12{,}5$ $E_E=0$	175	162,5
$(A,E),(C,D),B$	$C_{AE}=(7{,}5,7{,}5)$ $C_{CD}=(30,12{,}5)$	$E_{AE}=25$ $E_{CD}=12{,}5$ $E_B=0$	37,5	25
$(B,E),(C,D),A$	$C_{BE}=(12{,}5,15)$ $C_{CD}=(30,12{,}5)$	$E_{BE}=162{,}5$ $E_{CD}=12{,}5$ $E_A=0$	175	162,5

de donde se deduce que en esta etapa se unen los elementos A y E. La configuración actual es $(A,E),(C,D),B$.

Paso 3

A partir de la configuración actual tomamos las $\binom{3}{2}=3$ combinaciones posibles.

Partición	*Centroides*	E_k	E	ΔE
$(A,C,D,E),B$	$C_{ACDE}=(18{,}75,10)$	$E_{ACDE}=568{,}75$ $E_B=0$	568,75	531,25
$(A,B,E),(C,D)$	$C_{ABE}=(11{,}66,11{,}66)$ $C_{CD}=(30,12{,}5)$	$E_{ABE}=233{,}33$ $E_{CD}=12{,}5$	245,8	208,3
$(A,E),(B,C,D)$	$C_{AE}=(7{,}5,7{,}5)$ $C_{BCD}=(26{,}66,15)$	$E_{AE}=25$ $E_{BCD}=116{,}66$	141,66	104,16

de donde se deduce que en esta etapa se unen los clusters B y (C,D). La configuración actual es $(A,E),(B,C,D)$.

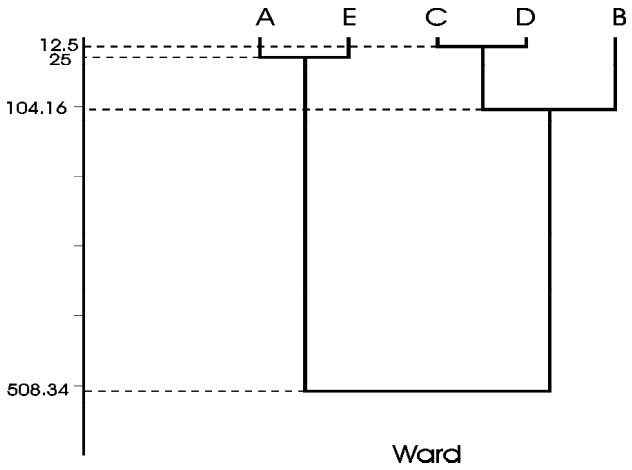
Paso 4

Evidentemente en este paso se unirán los dos clusters existentes. Los valores del centroide y de los incrementos de las distancias serán los siguientes

Partición	*Centroide*	E	ΔE
(A, B, C, D, E)	$C_{ABCDE} = (19, 12)$	650	508,34

El dendrograma asociado es el de la figura 3.8

Figura 3.8: Método de Ward



3.3. Fórmula de recurrencia de Lance y Williams.

A continuación vamos a exponer una expresión debida a Lance y Williams en 1967 que intenta aglutinar todos los métodos anteriores bajo una misma fórmula. Concretamente la expresión que dedujeron dichos autores proporciona la distancia entre un grupo K y otro grupo (I, J) formado en una etapa anterior por la fusión de dos grupos. Obviamente dicha expresión tiene importantes aplicaciones desde el punto de vista computacional ya que permite una reducción considerable en los cálculos.

La fórmula en cuestión es la siguiente

$$d(K,(I,J)) = \alpha_I d(K,I) + \alpha_J d(K,J) + \beta d(I,J) + \gamma |d(K,I) - d(K,J)|$$

De esta manera el cálculo de las distancias entre grupos usadas por las técnicas jerárquicas descritas anteriormente son casos particulares de la expresión anterior, para una elección conveniente de los parámetros α_I, α_J, β y γ. Algunos de estos coeficientes han sido ya deducidos en la descripción de los métodos anteriores (métodos del promedio ponderado y no ponderado, método del centroide, método de la mediana y método de Ward).

Veamos ahora cómo el método del *amalgamamiento simple* y el del *amalgamamiento completo* pueden ser también englobados bajo esta filosofía.

Amalgamamiento simple

Supongamos que en una etapa se dispone de un cluster C_j y de otro C_i que es fruto de la unión de otros dos clusters, C_{i_1} y C_{i_2} en una etapa anterior. El método del amalgamamiento simple determina que la distancia entre ambos clusters se establece como la menor distancia existente entre los elementos de ambos clusters; evidentemente, al estar constituido el cluster C_i por otros dos clusters C_{i_1} y C_{i_2}, dicho criterio equivale a calcular el mínimo de las distancias entre el cluster C_j y C_{i_1} y entre C_j y C_{i_2}. Teniendo en cuenta la siguiente igualdad (de fácil comprobación)

$$\text{Min}(a,b) = \frac{1}{2}(a+b) - \frac{1}{2}|a-b|$$

se tiene

$$d(C_j, C_i) = \text{Min}\{d(C_j, C_{i_1}), d(C_j, C_{i_2})\} =$$

$$= \frac{1}{2}d(C_j, C_{i_1}) + \frac{1}{2}d(C_j, C_{i_2}) - \frac{1}{2}|d(C_j, C_{i_1}) - d(C_j, C_{i_2})|$$

que corresponde a la expresión anterior con

$$\alpha_I = \alpha_J = \frac{1}{2} \;;\; \beta = 0 \;;\; \gamma = -\frac{1}{2}$$

Amalgamamiento completo

En las mismas hipótesis que en el caso anterior y usando la expresión

$$\text{Max}(a, b) = \frac{1}{2}(a + b) + \frac{1}{2}|a - b|$$

se tiene para el método del *amalgamamiento completo*

$$d(C_j, C_i) = \text{Max}\{d(C_j, C_{i_1}), d(C_j, C_{i_2})\} =$$

$$= \frac{1}{2}d(C_j, C_{i_1}) + \frac{1}{2}d(C_j, C_{i_2}) + \frac{1}{2}|d(C_j, C_{i_1}) - d(C_j, C_{i_2})|$$

que corresponde a la fórmula de Lance y Williams con

$$\alpha_I = \alpha_J = \frac{1}{2} \;;\; \beta = 0 \;;\; \gamma = \frac{1}{2}$$

Extrayendo los resultados obtenidos en apartados anteriores para otros procedimientos se puede comprobar la validez de la fórmula de recurrencia para dichos parámetros. Concretamente:

1. **Método del promedio no ponderado**

$$\alpha_I = \alpha_J = \frac{1}{2} \;;\; \beta = \gamma = 0$$

2. **Método del promedio ponderado**

$$\alpha_I = \frac{n_{i_1}}{n_{i_1} + n_{i_2}} \;;\; \alpha_J = \frac{n_{i_2}}{n_{i_1} + n_{i_2}} \;;\; \beta = \gamma = 0$$

3. **Método del centroide**

   Para la distancia euclídea al cuadrado se tiene

$$\alpha_I = \frac{n_{i_1}}{n_{i_1} + n_{i_2}} \;;\; \alpha_J = \frac{n_{i_2}}{n_{i_1} + n_{i_2}} \;;\; \beta = -\alpha_I\alpha_J \;;\; \gamma = 0$$

4. **Método de la mediana**

$$\alpha_I = \alpha_J = \frac{1}{2} \;;\; \beta = -\frac{1}{4} \;;\; \gamma = 0$$

5. **Método de Ward**

   Para la distancia euclídea al cuadrado se tiene

$$\alpha_I = \frac{n_{i_1} + n_j}{n_{i_1} + n_{i_2} + n_j} \;;\; \alpha_J = \frac{n_{i_2} + n_j}{n_{i_1} + n_{i_2} + n_j} \;;\; \beta = -\frac{n_j}{n_{i_1} + n_{i_2} + n_j} \;;\; \gamma = 0$$

3.4. Métodos Jerárquicos Disociativos.

Como se comentó en la introducción de este capítulo, los métodos disociativos, constituyen el proceso inverso a los aglomerativos. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez menores. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

En cuanto a la clasificación de estos métodos se puede decir que la filosofía de los métodos aglomerativos puede mantenerse para este otro tipo de procedimientos en lo que concierne a la forma de calcular la distancia entre los grupos, si bien, como es lógico, al partir de un grupo único que hay que subdividir, se seguirá la estrategia de maximizar las distancias, o minimizar las similaridades, puesto que buscamos ahora los individuos menos similares para separarlos del resto del conglomerado.

Esta clase de métodos son esencialmente de dos tipos:

1. Monotéticos, los cuales dividen los datos sobre la base de un solo atributo y suelen emplearse cuando los datos son de tipo binario.

2. Politéticos, cuyas divisiones se basan en los valores tomados por todas las variables.

Esta clase de procedimientos es bastante menos popular que los ascendentes por lo que la literatura sobre ellos no es muy extensa. Una cuestión importante que puede surgir en su desarrollo es el hecho de cuándo un cluster determinado debe dejar de dividirse para proceder con la división de otro conglomerado distinto. Dicha cuestión puede resolverse con la siguiente variante expuesta por MacNaughton-Smith en 1964 y que está concebida para aquellas medidas de asociación que sean positivas.

Dicho procedimiento comienza con la eliminación del grupo principal de aquel individuo cuya distancia sea mayor, o cuya similaridad sea menor, al cluster formado por los restantes individuos, tomando como base para calcular dichas distancias o similaridades cualquiera de los procedimientos anteriormente descritos en los métodos ascendentes. Así se tiene un cluster unitario y otro formado por los restantes individuos.

A continuación se añadirá al cluster unitario aquel elemento cuya distancia (similaridad) total al resto de los elementos que componen su actual cluster menos la distancia (similaridad) al cluster anteriormente formado sea máxima (mínima). Cuando esta diferencia sea negativa dicho elemento no se añade y se repite el proceso sobre los dos subgrupos.

Ejemplo 3.7 *Retomemos la matriz de distancias del ejemplo 3.1*

El método de cálculo de las distancias será la del método del amalgamamiento simple. *(Se propone como ejercicio el empleo de los otros tipos de estrategia).*

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	2,15	0					
C	0,7	1,53	0				
D	1,07	1,14	0,43	0			
E	0,85	1,38	0,21	0,29	0		
F	1,16	1,01	0,55	0,22	0,41	0	
G	1,56	2,83	1,86	2,04	2,02	2,05	0

Paso 1

Las distancias de cada individuo al cluster formado por el resto es

A	0,7
B	1,01
C	0,21
D	0,22
E	0,21
F	0,22
G	1,56

por lo que el individuo empleado para comenzar la división será el individuo etiquetado **G** *(notemos que ahora el criterio que se sigue es maximizar la distancia). Tenemos con ello dos clusters,* (G) *y* (A, B, C, D, E, F).

Paso 2

A continuación calculamos la distancia de cada individuo del cluster principal al resto, la distancia de cada individuo de dicho grupo al nuevo cluster formado así como la diferencia entre ambas.

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
A	0,7	1,56	−0,86
B	1,01	2,83	−1,82
C	0,21	1,86	−1,65
D	0,22	2,04	−1,82
E	0,21	2,02	−1,81
F	0,22	2,05	−1,83

A la vista de estos resultados es obvio que ningún elemento se añadirá al cluster anterior, por lo que procede comenzar con la división del grupo principal, empezando por el individuo **B**. *Tenemos así la división* (G), (B) (A, C, D, E, F).

Paso 3

Volvemos a calcular la distancia entre cada individuo del cluster (A, C, D, E, F) *así como la distancia de cada individuo de dicho grupo al nuevo cluster formado y la diferencia entre ambas.*

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
A	0,7	2,15	−1,45
C	0,21	1,53	−1,32
D	0,22	1,14	−0,92
E	0,21	1,38	−1,17
F	0,22	1,01	−0,79

de donde se deduce que ningún individuo se añadirá al nuevo cluster formado. Ahora se empezará a dividir el cluster (A, C, D, E, F) *por el individuo* **A**.

Paso 4

Calculamos la distancia entre cada individuo del cluster (C, D, E, F) *así como la distancia de cada individuo de dicho grupo al nuevo cluster formado y la diferencia entre ambas.*

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
C	0,21	0,7	−0,49
D	0,22	1,07	−0,85
E	0,21	0,85	−0,64
F	0,22	1,16	−0,94

Ningún elemento se añadirá al cluster formado por el individuo **A**. *Elegimos ahora el individuo* **D** *(también podíamos haber elegido el* **F***).*

Paso 5

Calculamos la distancia entre cada individuo del cluster (C, E, F) *así como la distancia de cada individuo de dicho grupo al nuevo cluster formado y la diferencia entre ambas.*

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
C	0,21	0,43	−0,22
E	0,21	0,29	−0,08
F	0,41	0,22	0,19

A la vista del resultado anterior se tiene que el individuo **F** *se suma al individuo* **D**. *Vemos si algún otro individuo se une*

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
C	0,21	0,43	−0,22
E	0,21	0,29	−0,08

con lo cual no se añade ningún individuo.

Paso 6

El proceso se seguiría ahora descomponiendo los dos clusters que quedan, a saber, (D, F) y (C, E), empezando con el primero de ellos pues es el que más distancia presenta entre sus elementos.

Las técnicas monotéticas son generalmente empleadas cuando los datos son de tipo binario. Ahora la división se inicia en aquellos individuos que poseen y aquellos que no poseen algún atributo específico. Teniendo en cuenta este criterio, para un conjunto de datos con m variables binarias hay m divisiones potenciales del conjunto inicial, $m-1$ para cada uno de los dos subgrupos formados y así sucesivamente; de ello se deduce que hay que determinar algún criterio para elegir la variable sobre la cual se va a proceder a la división.

El criterio que suele ser más usual es el basado en los estadísticos del tipo χ^2 obtenidos a partir de la tabla de doble entrada para cada par de variables

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

y tomar la variable k tal que $\sum_{j \neq k} \chi^2_{jk}$ sea máximo.

Otros criterios alternativos pueden ser

$$\text{Max} \sum \sqrt{\chi^2_{jk}}$$

$$\text{Max} \sum |ad-bc|$$

$$\text{Max} \sum (ad-bc)^2$$

Por ejemplo consideremos el siguiente ejemplo en el cual se tienen 5 individuos sobre los cuales se miden tres variables de tipo binario

X_1	X_2	X_3
0	1	1
1	1	0
1	1	1
1	1	0
0	0	1

Calculemos primero los estadísticos χ^2 para cada par de variables. Por ejemplo, para las variables X_1 y X_2 se tiene

$X_2 \backslash X_1$	1	0	Total
1	3	1	4
0	0	1	1
Total	3	2	5

de donde

$$\chi^2_{12} = \frac{45}{24} = 1{,}875$$

Asimismo $\chi^2_{13} = \frac{80}{36} = 2{,}22$ y $\chi^2_{23} = \frac{20}{24} = 0{,}83$. Ahora, aplicando el criterio $\text{Max} \sum_{j \neq k} \chi^2_{jk}$, se tiene

$$\begin{gathered} \chi^2_{12} + \chi^2_{13} = 4{,}09 \\ \chi^2_{12} + \chi^2_{23} = 2{,}7 \\ \chi^2_{13} + \chi^2_{23} = 3{,}05 \end{gathered}$$

de donde la división se basará en la determinación de quien posee la característica asociada a la variable X_1 y quien no, obteniéndose así los dos clusters (I_2, I_3, I_4) y (I_1, I_5). De forma sucesiva se seguiría aplicando este criterio a ambos subgrupos.

3.5. La matriz cofenética. Coeficiente de correlación cofenético.

Los métodos jerárquicos imponen una estructura sobre los datos y es necesario con frecuencia considerar si es aceptable o si se introducen distorsiones inaceptables en las relaciones originales. El método más usado para verificar este hecho, o sea, para ver la relación entre el dendrograma y la matriz de proximidades original, es el coeficiente de correlación cofenético, el cual es simplemente la correlación entre los $\frac{n(n-1)}{2}$ elementos de la parte superior de la matriz de proximidades observada y los correspondientes en la llamada matriz cofenética, C, cuyos elementos, c_{ij}, se definen como aquellos que determinan la proximidad entre los elementos i y j cuando éstos se unen en el mismo cluster.

Así, si tras el empleo de varios procedimientos cluster distintos, éstos conducen a soluciones parecidas, surge la pregunta de qué método elegiremos como definitivo. La respuesta la da el coeficiente cofenético, ya que aquel método que tenga un coeficiente cofenético más elevado será aquel que presente una menor distorsión en las relaciones originales existentes entre los elementos en estudio.

Ejemplo 3.8 *Calculemos las matrices cofenéticas y los coeficientes de correlación cofenéticos asociados a los ejemplos 3.1 a 3.4*

1. **Método del amalgamamiento simple**

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	1,01	0					
C	0,7	1,01	0				
D	0,7	1,01	0,29	0			
E	0,7	1,01	0,21	0,29	0		
F	0,7	1,01	0,29	0,22	0,29	0	
G	1,56	1,56	1,56	1,56	1,56	1,56	0

siendo el coeficiente de correlación cofenético 0.911438774

2. **Método del amalgamamiento completo**

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	2,83	0					
C	1,16	2,83	0				
D	1,16	2,83	0,55	0			
E	1,16	2,83	0,21	0,55	0		
F	1,16	2,83	0,55	0,22	0,55	0	
G	2,05	2,83	2,05	2,05	2,05	2,05	0

siendo el coeficiente de correlación cofenético 0.788405653

3. **Método de la distancia promedio no ponderada**

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	1,7075	0					
C	0,945	1,7075	0				
D	0,945	1,7075	0,41	0			
E	0,945	1,7075	0,21	0,41	0		
F	0,945	1,7075	0,41	0,22	0,41	0	
G	2,303125	2,303125	2,303125	2,303125	2,303125	2,303125	0

siendo el coeficiente de correlación cofenético 0.911167777

4. **Método de la distancia promedio ponderada**

	A	B	C	D	E	F	G
A	0						
B	1,442	0					
C	0,945	1,442	0				
D	0,945	1,442	0,42	0			
E	0,945	1,442	0,21	0,42	0		
F	0,945	1,442	0,42	0,22	0,42	0	
G	2,06	2,06	2,06	2,06	2,06	2,06	0

siendo el coeficiente de correlación cofenético 0.911359728

3.6. El problema del número de clusters a determinar.

Con frecuencia, cuando se emplean técnicas clusters jerárquicas, el investigador no está interesado en la jerarquía completa sino en un subconjunto de particiones obtenidas a partir de ella. Las particiones se obtienen cortando el dendrograma o seleccionando una de las soluciones en la sucesión encajada de clusters que comprende la jerarquía.

Desafortunadamente este paso fundamental está entre los problemas que todavía no están totalmente resueltos. Entre las razones más importantes que se pueden citar para que dicho problema siga siendo un campo abierto están las siguientes:

1. **La inexistencia de una hipótesis nula apropiada.**

   En efecto, la dificultad para crear una hipótesis nula operativa radica en la falta de una definición clara y comprensiva de lo que significa *no estructura* en un conjunto de datos. El concepto de *no estructura* (que podía ser una posible hipótesis nula) está bastante lejos de ser clara, lo cual conlleva a no saber qué tipos de contrastes hay que desarrollar para determinar si una determinada estructura está presente o no en el conjunto de datos. Dubes y Jain (1980) comentan sobre este hecho lo siguiente:

   > *... el rechazo de la hipótesis nula no es significativo porque no han sido desarrolladas hipótesis alternativas significativas; todavía no existe una definición útil y práctica de estructura cluster, matemáticamente hablando.*

2. **La naturaleza compleja de las distribuciones muestrales multivariantes.**

   Igualmente intratable es el problema de la mixtura de las distribuciones muestrales multivariantes en el análisis de datos reales. Aunque son muchos los aspectos conocidos y desarrollados acerca de la distribución normal multivariante, no es ni esperable ni razonable que los datos que se manejen en estos estudios obedezcan a dicha ley, sino que existirán mixturas de diversas distribuciones muestrales que pueden ser incluso desconocidas.

Las soluciones propuestas a estas cuestiones han sido múltiples. En algunos campos de aplicación, como puede ser algunos tipos de investigaciones en las ciencias biológicas, el problema de determinar el número de clusters no es un tema que parezca excesivamente importante ya que el objetivo puede ser simplemente explorar el patrón general de las relaciones existentes entre los individuos objeto de estudio, lo cual puede ser observado a partir del dendrograma. Sin embargo hay campos de aplicación en los cuales se pretende ir más lejos en el estudio y obtener una clasificación de los individuos lo más realista posible, lo cual conlleva tener que estudiar con más énfasis el problema del número de clusters a determinar. Esta cuestión ha motivado la aparición de múltiples *reglas*. Algunas de estas reglas son simples métodos heurísticos, otras están basadas en contrastes de hipótesis formales, los cuales han sido desarrollados al amparo de la hipótesis de la existencia de una determinada distribución muestral (casi siempre la normal multivariante), mientras que otros son procedimientos asimismo heurísticos pero que extraen la filosofía de los contrastes existentes en poblaciones normales. A continuación vamos a citar algunas de estas reglas, si bien hay que decir que son muchísimos los procedimientos que en los últimos años han sido desarrollados, con frecuencia orientados a técnicas particulares.

- La primera técnica que podemos citar se basa simplemente en cortar el dendrograma de forma subjetiva tras visualizarlo. Obviamente este procedimiento no es nada satisfactorio puesto que está generalmente sesgado por la opinión que el investigador posee sobre sus datos.
- Un método más formal, pero asimismo heurístico, se basa en representar en una gráfica el número de clusters que se observan en los distintos niveles del dendrograma frente a los niveles de fusión a los que los

clusters se unen en cada nivel. La presencia de una pendiente poco pronunciada sugiere que la siguiente unión de clusters no aporta apenas información adicional sobre la aportada en el nivel anterior. Este método, por lo tanto, se basa en la existencia de *pequeños* saltos o discontinuidades en los niveles de fusión.

• Mojena (1977) siguió con la idea de estudiar los saltos relativos en los valores de fusión y sugirió otro procedimiento heurístico bastante divulgado y que ha sido fuente de bastantes investigaciones posteriores. En su método se compara el valor de fusión de cada etapa con el promedio de los valores de fusión sumado con el producto de una cierta constante por la cuasidesviación típica de los valores de fusión. Cuando un valor de fusión supera dicha cantidad se concluye que el nivel precedente es el que origina la solución óptima. Mojena sugirió que el valor de la constante debía de estar comprendido en el rango de 2.75 a 3.50 si bien Milligan, en 1985, tras una detallada investigación de valores en función del número de clusters, establece que el valor óptimo para dicha constante debe ser 1.25.

• Beale en (1969) propuso el uso de un contraste basado en la distribución F de Snedecor para contrastar la hipótesis de la existencia de c_2 clusters frente a la existencia de c_1 clusters, siendo $c_2 > c_1$. Para ello se consideran la suma, para cada partición, de las desviaciones cuadráticas medias de los elementos de cada cluster a su centroide, llamémoslas DC_1 y DC_2:

$$DC_1 = \frac{1}{n - c_1} \sum_{i=1}^{c_1} \sum_{j=1}^{n_i} ||x_{ij} - \overline{x}_i||^2$$

$$DC_2 = \frac{1}{n - c_2} \sum_{i=1}^{c_2} \sum_{j=1}^{n_i} ||x_{ij} - \overline{x}_i||^2$$

donde se ha supuesto que el cluster $i-$ésimo posee n_i elementos y n es el total de la muestra. El estadístico considerado es

$$F(p(c_2 - c_1), p(n - c_2)) = \frac{\dfrac{DC_1 - DC_2}{DC_2}}{\left[\left(\dfrac{n - c_1}{n - c_2}\right)\left(\dfrac{c_2}{c_1}\right)^{\frac{2}{p}} - 1\right]}$$

Un resultado significativo indica que la división en c_2 clusters representa una mejoría frente a la división en c_1 clusters. Notemos que este contraste no impone ninguna distribución concreta de la muestra.

Los siguientes métodos que vamos a comentar ahora proceden en su mayoría de la abstracción de procedimientos inherentes en su mayoría al análisis multivariante paramétrico. Para su desarrollo, definimos las siguientes matrices:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})(x_{ij} - \overline{x})'$$

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)'$$

$$B = \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'$$

Estas matrices representan, respectivamente, la dispersión total de todos los individuos respecto de su centroide, la suma de las dispersiones en cada grupo (desviación intra clusters) y la dispersión entre los grupos (desviación entre clusters). Asimismo k representa el número total de clusters y n es el tamaño total de la muestra ($n = n_1 + \cdots + n_k$).

Se puede comprobar que se cumple la igualdad $T = W + B$. Dicha igualdad es la extensión al caso multivariante de la conocida descomposición de la variabilidad del análisis de la varianza de una vía. Para fijar ideas y particularizando al caso unidimensional, es obvio que en tales circunstancias un criterio lógico para determinar el número de clusters sería elegir aquella partición que tuviera el menor valor en la desviación intra-clusters o, equivalentemente, el mayor valor en la desviación entre-clusters.

Siguiendo con esta idea se puede extender dicha situación al caso multivariante, si bien el empleo de las matrices antes reseñadas no hace tan inmediata dicha extensión. Por ello surgen diversos procedimientos, entre los cuales podemos citar los siguientes:

1. **Minimización de la traza de W.**

   Esta es la extensión más inmediata al criterio anteriormente comentado para el caso unidimensional. Evidentemente esto es equivalente a minimizar la suma de los cuadrados de las distancias euclídeas entre cada individuo a la media del cluster al que ha sido asignado.

   Hay que hacer notar que este criterio está implícito en diversos métodos no jerárquicos que serán descritos en el capítulo siguiente, como el de Forgy, Jancey y el de las k-medias, así como, dentro de los métodos jerárquicos, el de Ward.

   Notemos asimismo que como $T = W + B$, entonces $\text{tr}[T] = \text{tr}[W] + \text{tr}[B]$, por lo que minimizar la traza de W equivale a maximizar la traza de B ya que, sea cual sea la configuración de clusters que se establezca, la matriz T no varía y, por tanto, tampoco su traza.

2. **Minimización de $k^2|W|$.**

   Marriot en 1971 sugiere el empleo de $k^2|W|$, tomándose el valor de k tal que haga esa cantidad mínimo.

3. **Minimización del determinante de W.**

   En el análisis de la varianza multivariante de una vía (MANOVA) son diversos los criterios empleados basados en la distribución de la razón de verosimilitudes. Entre ellos destaca el criterio de Wilks, el cual considera el cociente

   $$\Lambda = \frac{|W|}{|T|} = \frac{|W|}{|W + B|}$$

   rechazándose la hipótesis nula de igualdad de las medias poblacionales si ese cociente es menor que un valor predeterminado o, lo que es equivalente, si el cociente

   $$\frac{|T|}{|W|}$$

   es mayor que un determinado valor.

   Es evidente que en nuestro ambiente no podemos aplicar este contraste ya que carecemos de las hipótesis de normalidad multivariante, pero se puede *abstraer* la filosofía de dicho contraste y aplicarlo para nuestros propósitos, lo cual no deja de ser un método puramente heurístico. Así pues y puesto que para todas las particiones de los individuos en k grupos la matriz T permanece constante, Friedman y Rubin sugirieron en 1967 la maximización de dicho cociente, lo cual equivale a la minimización de $|W|$.

4. **Maximización de la traza de BW^{-1}.**

   Siguiendo con la misma idea anterior, otro de los criterios que se pueden aplicar en el análisis de la varianza multivariante de una vía es el debido a Lawley y Hotelling, quienes proponen el empleo del estadístico

   $$\text{tr}[BW^{-1}]$$

   siendo rechazada la hipótesis nula cuando dicha traza supere un cierto valor impuesto de antemano.

   En nuestro caso, y siempre abstrayendo la filosofía del criterio expuesto, debemos seleccionar aquella partición que produzca la maximización de esa traza.

5. Por otro lado, Calinski y Harabasz (1974) proponen el estadístico

   $$C = \frac{\dfrac{\text{tr}[B]}{k-1}}{\dfrac{\text{tr}[W]}{n-k}}$$

   tomando como número óptimo de clusters aquel que produzca el mayor valor de C.