

Contribuciones a la representación de datos multidimensionales mediante árboles aditivos

Antonio Arcas Pons

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tesisenxarxa.net) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tesisenred.net) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tesisenxarxa.net) service has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized neither its spreading and availability from a site foreign to the TDX service. Introducing its content in a window or frame foreign to the TDX service is not authorized (framing). This rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

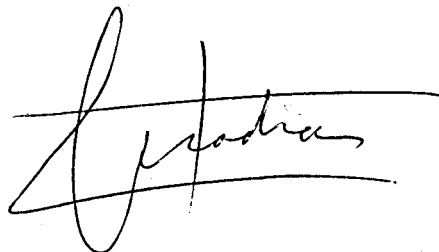
CONTRIBUCIONES A LA REPRESENTACION DE DATOS MULTIDIMENSIONALES

MEDIANTE ARBOLES ADITIVOS

Memoria presentada para
optar al título de
Doctor en Matemáticas, por
Antoni Arcas Pons

VºBº

EL DIRECTOR



Prof. D. Carlos M. Cuadras Avellana,

Catedrático de ~~Bio~~estadística.

Universidad de Barcelona.

Barcelona, 15 de Febrero de 1986

AGRADECIMIENTOS

Quisiera agradecer al Dr. C.Cuadras , director de esta memoria, sus constantes consejos y orientaciones durante la realización de la misma. También a los profesores del Dpto.de Bioestadística su apoyo constante , destacando en especial la inestimable e incansable cooperación del Prof. M.Salicrú así como la colaboración y supervisión del Dr.J.M.Oller en los aspectos de la memoria que tienen relación con su línea de trabajo.

Me siento también complacido de poder expresar mi gratitud al Dr.F.Sales por el estímulo que siempre me ha proporcionado. Asimismo quisiera agradecer al Dr.J.Cascante su interés y colaboración en la línea que hemos desarrollado conjuntamente con D. M.Salicrú.

Agradezco finalmente la supervisión del ejemplo al Dr.L. Serra del Dpt. de Genética.

A Roser y Xènia dedicarles el trabajo al darle verdadero sentido al mismo y ser la ayuda fundamental en los momentos difíciles.

INDICE

| | |
|--|----|
| PROLOGO | 1 |
| 1. REPRESENTACION DE UN CONJUNTO A TRAVES DEL ANALISIS MULTIVARIANTE | |
| 1.1. Introducci3n | 3 |
| 1.2. Principales m3todos de representaci3n | 5 |
| 1.3. Relaciones entre los distintos m3todos de representaci3n | 9 |
| 1.4. Objetivos de la memoria | 14 |
| 2. SOBRE LOS DISTINTOS METODOS DE REPRESENTACION DE DATOS MULTIDIMENSIONALES | |
| 2.1. Introducci3n | 16 |
| 2.2. Disimilaridades y distancias | 17 |
| 2.3. Propiedades generales del MDS | 25 |
| 2.4. Representaci3n mediante 3rboles ultram3tricos | 34 |
| 2.5. Representaci3n mediante grafos piramidales | 40 |
| 2.6. Ventajas de los 3rboles aditivos frente los ultram3tricos ... | 43 |
| 3. REPRESENTACION MEDIANTE ARBOLES ADITIVOS | |
| 3.1. Introducci3n | 49 |
| 3.2. Formalizaci3n de P. Buneman | 51 |
| 3.3. Formalizaci3n de la representaci3n de un conjunto asociada a una distancia aditiva | 56 |
| 3.4. Relaciones con las distancias ultram3tricas | 82 |
| 4. ESTRUCTURA DE VARIEDAD EN EL CONJUNTO DE DISTANCIAS ADITIVAS | |
| 4.1. Introducci3n | 89 |
| 4.2. Definiciones previas | 90 |
| 4.3. Estructura del conjunto de distancias aditivas | 92 |
| 4.4. Determinaci3n de la frontera de la variedad | 95 |

| | |
|---|-----|
| 4.5. Estructura del conjunto de distancias ultramétricas dentro de la variedad | 105 |
| 4.6. Ajuste por mínimos cuadrados en una carta de la variedad | 109 |
| 4.7. Representación espacial de distancias aditivas | 112 |
| 5. INFERENCIA EN ARBOLES ADITIVOS | |
| 5.1. Introducción | 119 |
| 5.2. Modelo probabilístico para el vector de distancias | 121 |
| 5.3. Contrastes en árboles aditivos | 124 |
| 6. ALGORITMOS DE CONSTRUCCION DE ARBOLES ADITIVOS | |
| 6.1. Introducción | 137 |
| 6.2. Principales algoritmos | 138 |
| 6.3. Ejemplo | 151 |
| CONCLUSIONES | 155 |
| BIBLIOGRAFIA | 158 |

PROLOGO

A raíz de la visita del Dr. F. J. Rohlf el año 1980 en la cual trató diversos aspectos sobre análisis multivariante con el Dr. C. M. Cuadras, comentando entre otros el tema de las representaciones mediante árboles aditivos, el Dr. C.M. Cuadras pensó entonces en la posibilidad de una línea de trabajo basada en el desarrollo de aspectos de modelos de representación en redes centrándonos en particular en los árboles ultramétricos y árboles aditivos desde una perspectiva del análisis multivariante. En este sentido, como primer paso realicé una tesina en la que se formalizaban diversos aspectos de la taxonomía numérica lo cual supuso una verdadera motivación para seguir en una línea parecida que ha desembocado en esta memoria, en la que se ha desarrollado un estudio formal sobre la representación utilizando árboles aditivos. La realización paralela del trabajo "Eigenanalysis and metric multidimensional scaling on hierarchical structures" por el Dr. C.M. Cuadras y las conexiones de este tema con la línea de trabajo llevada por el Dr. J.M. Oller sobre Geometría Riemmaniana aplicada al análisis de datos, han sido también un punto de apoyo muy importante en la realización de este estudio.

Podríamos dividir esta memoria en tres partes: En la primera parte (cap. 1 y 2) se relaciona la representación por árboles aditivos con otras técnicas de representación del análisis de datos. En la segunda parte (cap. 3) se obtiene una formalización de las relaciones entre árboles aditivos y distancias aditivas. En la tercera parte (cap. 4, 5 y 6) se analiza una estructura en el conjunto de árboles aditivos tratando posteriormente sobre inferencia y confección de algoritmos en árboles, conectando esta parte con los trabajos antes citados.

1.- REPRESENTACION DE UN CONJUNTO A TRAVES
DEL ANALISIS MULTIVARIANTE.

Resumen: En este capítulo se expone el concepto de representación de un conjunto a través del análisis multivariante presentando los principales métodos de representación junto con sus inter-relaciones.

Sumario:

- 1.1.- Introducción.
- 1.2.- Principales métodos de representación.
- 1.3.- Relaciones entre los distintos métodos de representación.
- 1.4.- Objetivos de la memoria.

1.1.- INTRODUCCION

En análisis de datos es interesante por lo general, considerar un modelo que refleje las analogías y diferencias existentes entre los elementos de un conjunto. Es también habitual que las proximidades y divergencias entre los mismos provengan de la observación de varios caracteres. Así, en análisis multivariante se estudian métodos que tienen por objetivo la búsqueda, representación e interpretación de los datos obtenidos a través de varias variables.

En la mayoría de los estudios aplicados a Biología, Medicina, Psicología, Educación, Agricultura, Economía, ... aparecen una serie de medidas asociadas a distintas variables relacionadas entre sí. En este sentido, Romeder (1973) expone un método de diagnóstico automático de dos clases de ictericia a partir del conocimiento de 84 variables; Bartlett(1951) relaciona varios factores económicos: consumo, precio, coste de producción y renta de los consumidores; Petit Pierre y Cuadras (1977) realizan una clasificación sistemática de coleópteros del género *Timarcha* a partir de 5 medidas biométricas; en Antropología, es común la determinación del sexo a través de algunas medidas craneales.

Es evidente que una serie de análisis univariantes tratados por separado resultarían del todo inadecuados, pues ignoraríamos las correlaciones entre las variables. Utilizaremos pues, en general, técnicas de análisis multivariante que ayudan a interpretar las relaciones entre las variables considerando la mutua información entre las mismas.

Cuadras (1981) establece una clasificación de los métodos más importantes del análisis multivariante atendiendo a las diferencias entre las distintas técnicas según se trate de un estudio en una o varias

poblaciones y uno o dos grupos de variables. Así, Análisis de Componentes Principales y Análisis Factorial utilizan una misma población y un solo grupo de variables; el primero tiene por objetivo la descripción de la información contenida en las variables mediante un número reducido de componentes, mientras el segundo pretende relacionar un conjunto de variables mediante un modelo lineal. Análisis canónico de poblaciones, análisis discriminante y análisis multivariante de la varianza utilizan varias poblaciones y un solo grupo de variables. Regresión múltiple y análisis de correlación canónica utilizan una población y dos grupos de variables.

Por otro lado, considera las técnicas que a partir de una colección, en general, heterogénea de objetos y una disimilaridad definida sobre ellos tienen por objetivo la representación de los mismos en un espacio geométrico modelo en el sentido siguiente: Si $S = \{s_1, \dots, s_n\}$ es el conjunto de objetos, δ una disimilaridad definida sobre S y (X, d) el espacio geométrico modelo, se trata de hallar una función

$$\phi : (S, \delta) \longrightarrow (X, d)$$

de modo que $d(\phi(s_i), \phi(s_j))$ se aproxime a $\delta(s_i, s_j)$ utilizando algún criterio de ajuste.

Al proceso de búsqueda de tal función es denominado en general como el problema de hallar una representación de los objetos de un conjunto.

1.2.- PRINCIPALES TECNICAS DE REPRESENTACION

Podemos dividir las representaciones en dos grupos según el tipo de espacio geométrico modelo: modelos espaciales o continuos y modelos de redes o discretos. El objetivo en los primeros es la representación de cada objeto como las coordenadas de un punto en un espacio euclídeo de modo que las distancias entre los mismos reflejen las relaciones de proximidad observadas entre ellos, mientras que el objetivo en los modelos de redes es la representación de cada objeto mediante un nudo en un grafo conexo, de modo que las relaciones entre los nudos en el grafo reflejen la disimilaridad observada.

Las principales técnicas relativas a modelos continuos están incluidas dentro del Multidimensional Scaling (MDS), definido por De Leeuw(1982) como aquellos métodos de representación en que $X = R^n$. Cuadras (1981) matizando más en la naturaleza de los datos considera dentro del MDS aquellas técnicas de representación de datos que tienen por objetivo la construcción de una configuración de puntos a través de una determinada información sobre las disimilaridades entre los objetos.

De los modelos de redes que se han formulado cabe citar los siguientes:

- a) Arboles ultramétricos: Se caracterizan por ser X un grafo conexo sin ciclos con un nudo equidistante a los nudos extremos y d una distancia ultramétrica.
- b) Arboles piramidales: Es un tipo especial de grafo conexo en donde d es una distancia compatible con un cierto orden definido sobre los objetos en el sentido que dados tres puntos la distancia entre los puntos extremos excede a aquellas entre puntos intermedios.

c) Arboles aditivos: En este caso X es un grafo conexo sin ciclos y d es una distancia verificando el axioma del cuarto punto (Buneman, 1971). Este modelo de representación será estudiado en esta memoria bajo los aspectos que citaremos posteriormente.

En el siguiente capítulo se realiza una exposición más detallada de estos métodos de representación continuos y discretos.

Ejemplos a partir de datos reales de representaciones mediante modelos citados anteriormente son:

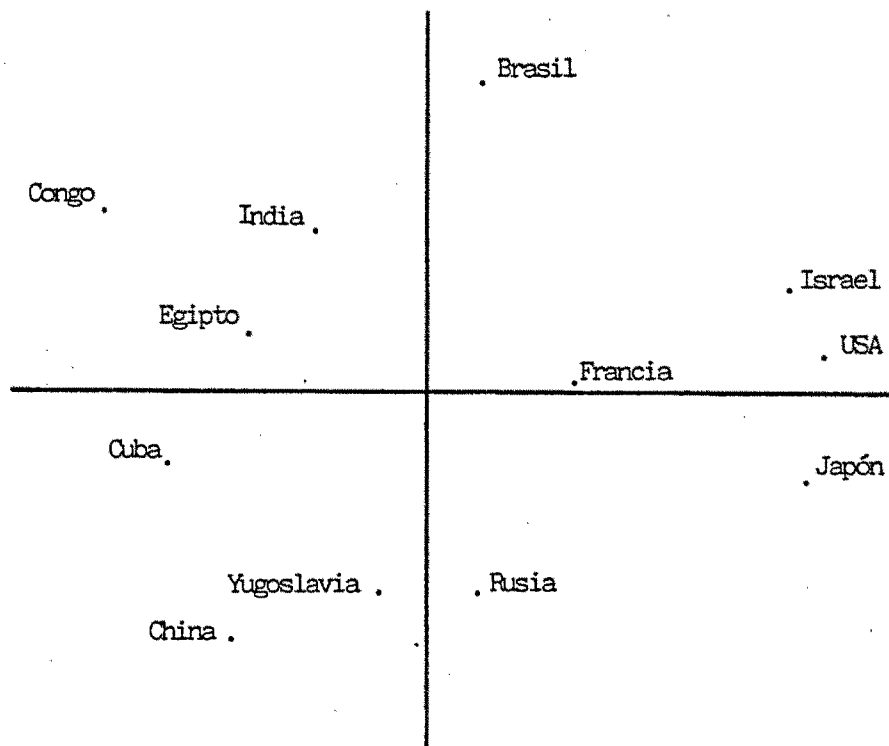


Figura 1: Representación por MDS de un conjunto de 12 naciones según un estudio de Wish (1970) y Wish, Deutch, Biener (1972).

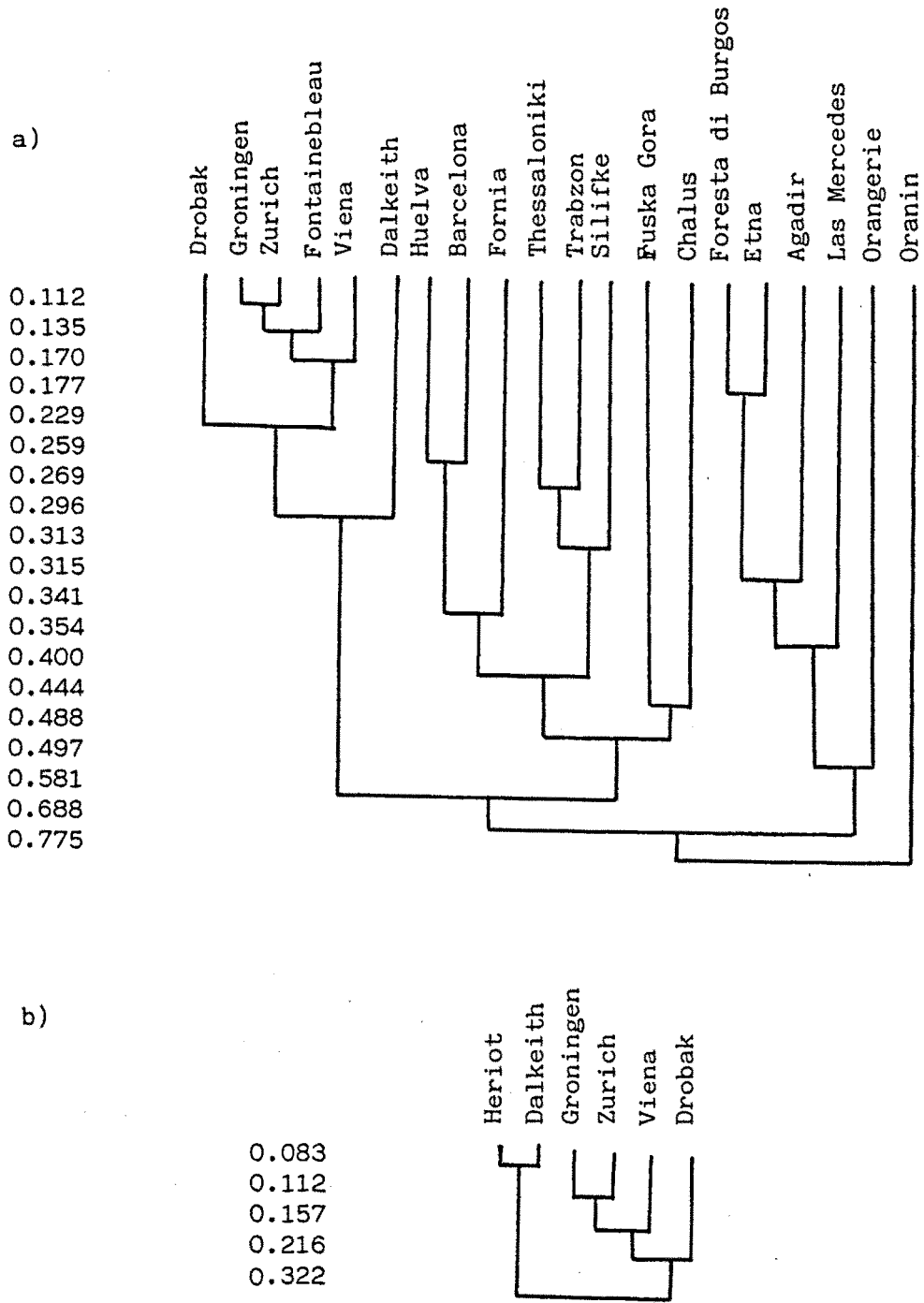


Figura 2:

a) Clasificación taxonómica de 20 poblaciones de Drosophila Subobscura realizada por Salicrú (1983) partiendo de una matriz de distancias obtenida por Prevosti (1974) basadas en las frecuencias de las ordenaciones cromosómicas producidas por inversiones y aplicando el método UPGMA.

b) Se observa como el estrecho de Drover es una barrera genética según resultados del mismo estudio anterior.

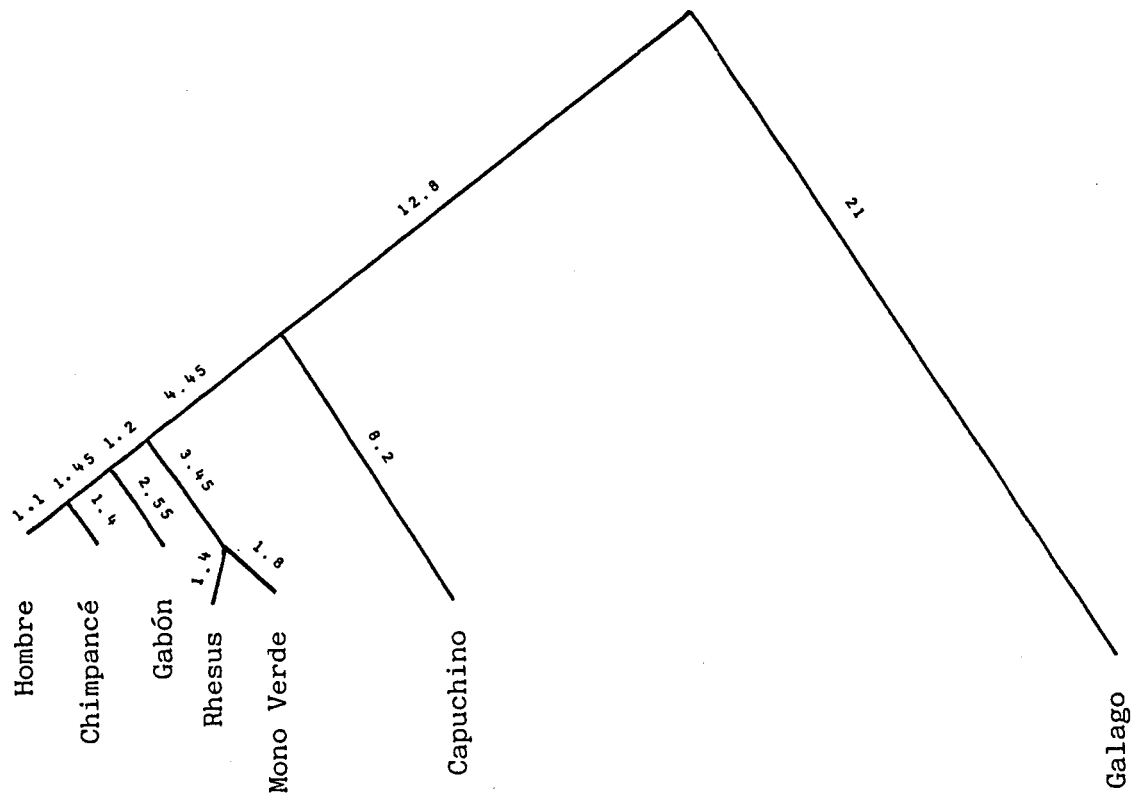


Figura 3: Representación mediante un árbol aditivo de especies de primates basada en técnicas de hibridación del DNA según un estudio de Kohne et al., 1972.

1.3.- RELACIONES ENTRE LOS DISTINTOS METODOS DE REPRESENTACION

Una de las principales dificultades con que nos encontramos al obtener una disimilaridad (u otra información relativa a la misma) sobre un conjunto de objetos es decidir el tipo de representación, modelos continuos o en redes, que es más apropiado a nuestros datos. Este problema ha sido discutido por diversos autores (Carroll, 1976), (Tversky, 1977), (Cunningham, 1978) coincidiendo en señalar que datos que se ajusten a una estructura factorial (en términos de un número reducido de factores) son preferiblemente representables mediante modelos continuos, mientras que si se puede pensar intuitivamente en alguna estructura jerárquica o filogenética podría tener más sentido una representación mediante modelos de redes.

En un interesante trabajo, Carroll (1976) concluye en la necesidad de que en un futuro próximo se desarrollen modelos que contengan aspectos continuos y discretos, pues la creciente complejidad de los datos en los estudios experimentales así lo van a exigir.

Pruzansky, Tversky y Carroll (1982) realizan un estudio en que tratan sobre las relaciones entre modelos continuos y en redes desde una perspectiva empírica. Intentan desarrollar una metodología que ayude a decidir el tipo de representación a utilizar cuando se tiene un conjunto de datos obtenidos a través de una disimilaridad dada sobre el conjunto de objetos. Como técnica MDS se utiliza el ajuste a un plano mientras que como modelo en red se utiliza árboles aditivos.

En una primera fase, mediante técnicas de simulación se compara el grado de ajuste de conjuntos de datos obtenidos artificialmente a representaciones en el plano y mediante estructuras de árbol. La generación de datos está basada en la obtención de una distancia a través de un plano o un árbol (reduciéndola a varianza unidad) sumándole a cada una de las distancias una componente de error a través de una distribución $N(0, \sigma^2)$ para distintos valores de σ^2 ($\sigma^2 = 0, \sigma^2 = 0.25, \sigma^2 = 0.50$) y utilizando muestras de tamaños diferentes ($n = 12, 24, 36$) obteniendo 11 réplicas para cada condición, con un total de $2 \times 9 \times 11 = 198$ condiciones experimentales. A fin de realizar el ajuste a un plano o un árbol utiliza los programas KYST (Kruskal et al 1973) y ADDTREE (Tversky, 1977) basados en sendos algoritmos de transformación de una distancia a un modelo espacial y a un árbol aditivo respectivamente (en capítulos sucesivos se comentarán aspectos relativos a estos programas). La comparación entre la distancia original y la distancia estimada se realiza mediante el coeficiente de correlación lineal r_1 . Las principales conclusiones a las que se llegan en este estudio son:

- a) Los datos surgidos del plano se adaptan mejor utilizando MDS mientras los datos surgidos del árbol se adaptan mejor a través de modelos de redes, resultando las correlaciones entre las distancias originales y estimadas prácticamente idénticas para ambos casos.
- b) KYST tiende a adaptar mejor datos surgidos de árbol que ADDTREE datos surgidos del plano.

En todo caso, a partir del modelo apropiado se obtiene siempre una correlación superior.

En la figura 4 se pueden observar los resultados para muestras de tamaño 24.

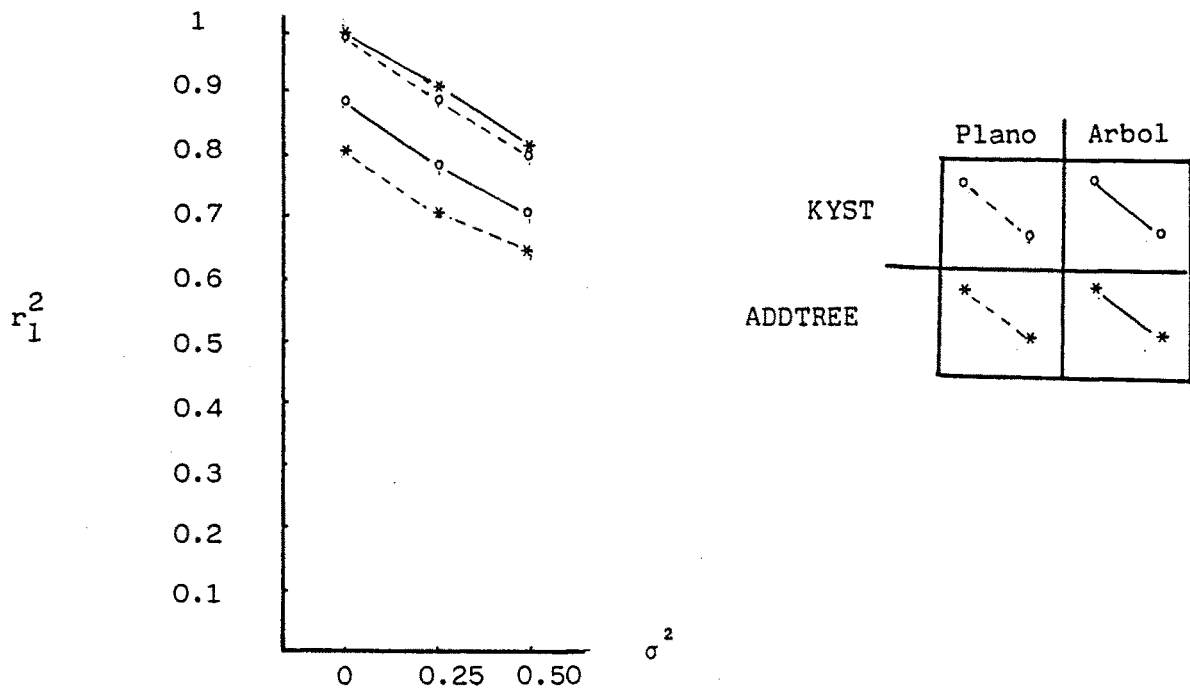


Figura 4.

En una simulación parecida perturba la tercera parte (superior o inferior) de las distancias de la distribución, tomando en este caso 5 réplicas por condición. La conclusión que obtiene es que la perturbación de pequeñas distancias reduce el ajuste de ADDTREE y la perturbación de distancias grandes tiene un efecto sensiblemente inferior, mientras que en KYST parece observarse el resultado contrario.

En consecuencia, el estudio y trabajos citados anteriormente nos inducen a pensar en la posibilidad de que la distribución de la disimilitud observada entre los objetos sirva para poder decidir entre un modelo espacial o un modelo en árbol.

Sobre el mismo particular, Tversky (1977), llamando μ a la media del conjunto de distancias y $\lambda = \frac{1}{2} \max \{ d(x,y); x,y \in S \}$ obtiene los siguientes resultados:

- 1) Si las distancias provienen de un árbol verificando ciertas condiciones de regularidad se cumple: $\mu > \lambda$ (sesgado a la izquierda) .
- 2) En un conjunto convexo y acotado del plano euclídeo con la distribución uniforme se cumple: $\mu < \lambda$ (sesgado a la derecha) .

Este análisis sugiere que una configuración convexa de puntos en un plano tiende a generar muchas distancias pequeñas y pocas grandes, mientras en un árbol se produce el fenómeno contrario. Por tanto, esto parece indicar que el sesgo (momento central de tercer orden stand.) puede ser una medida para diagnosticar si el conjunto de datos obtenidos es más propenso a ser representado por un plano o por un árbol.

En la misma línea Pruzansky et al. (1982) proponen considerar además otra medida, la elongación, que consiste en la proporción de ternas i, j, k para las cuales si $d_{ij} \leq d_{jk} \leq d_{ik}$ nos resulta $2 d_{jk} \geq d_{ij} + d_{ik}$. La proporción de " triángulos elongados" tiende a ser mayor en árboles que en una representación espacial. Así, en una distancia ultramétrica todas las ternas verifican la condición anterior por lo que la elongación es 1.

La combinación de ambas medidas puede resultar eficaz a fin de decidir entre ambos tipos de representaciones. La utilización de árboles tiende a producir sesgo negativo y mayor proporción de triángulos elongados, mientras que la representación en un plano conlleva menor elongación y sesgo positivo.

Realizan entonces un análisis a través de 20 conjuntos de datos reales obtenidos por otros autores que soporta la teoría anterior

tal y como se observa en el siguiente gráfico (Fig.5) que resume los resultados.

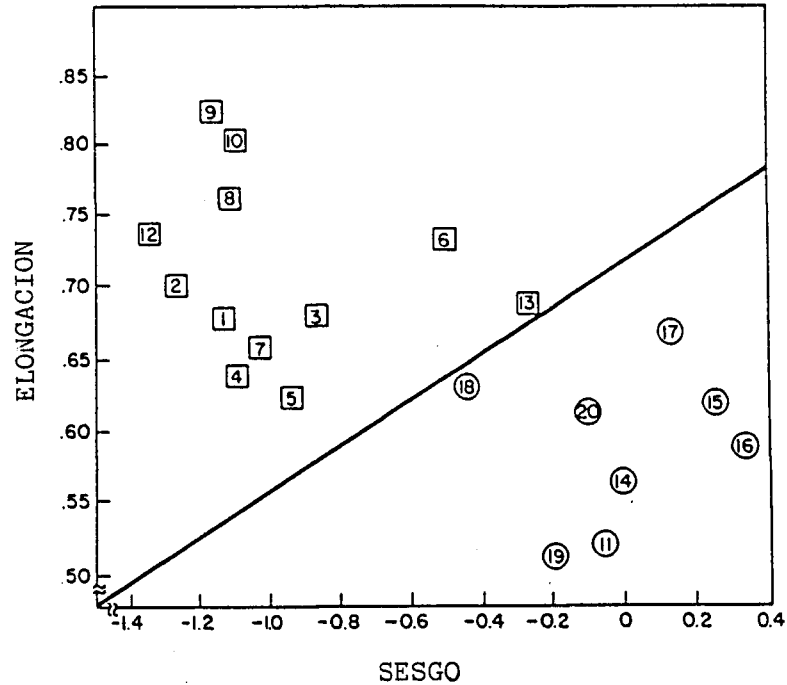


Figura 5: Un círculo indica que la solución KYST da mejor ajuste para el coeficiente de correlación, mientras un cuadrado indica que ADDTREE proporciona mejor ajuste.

También los resultados de este estudio concuerdan con las hipótesis de Tversky y Cunningham puesto que conjuntos de datos como vehículos, animales, profesiones,... que dejan entrever una estructura jerárquica son representados mejor a través de un árbol, mientras que conjuntos de colores, sonidos, con una estructura factorial clara (por ejemplo sonidos en términos de intensidad y frecuencias) son representados mejor a través de un modelo espacial.

1.4.- OBJETIVOS DE LA MEMORIA

En los apartados anteriores así como en el capítulo II presentamos y analizamos los principales resultados relativos a los métodos de representación de datos como una parte del análisis multivariante. Queda así centrado el estudio de las representaciones mediante árboles aditivos, objetivo final de esta memoria.

Debido al carácter eminentemente práctico de las técnicas de representación, la mayor parte de los trabajos realizados (por ejemplo, el trabajo antes comentado de Pruzansky et al.) están dirigidos a estudios empíricos, como por ejemplo la búsqueda y comparación de algoritmos por técnicas de simulación o la adecuación a problemas con datos reales. Así resulta necesario un desarrollo teórico paralelo del tema pues, debido quizás a lo reciente del mismo y al desarrollo informático, ha sido dejado un tanto de lado, siendo escasos los estudios que lo tratan, destacando en este sentido el trabajo de Buneman (1971).

Así pues, en esta memoria pretendemos contribuir al estudio de las representaciones mediante árboles aditivos desarrollando aspectos teóricos de los mismos. En primer lugar analizamos las principales propiedades de los árboles aditivos a través de una formalización propia comparándolo con el estudio de P. Buneman (1971). Después dotamos al conjunto de distancias aditivas (asociadas a árboles aditivos) de una estructura matemática adecuada a fin de tratar problemas relativos a contrastes, algoritmos y representación de distancias asociadas a árboles por modelos continuos.

2 . SOBRE LOS DISTINTOS METODOS DE REPRESENTACION
DE DATOS MULTIDIMENSIONALES

Resumen: En el presente capítulo se realiza una exposición de las principales propiedades y características de los métodos de representación más utilizados.

Sumario:

- 2.1.- Introducción.
- 2.2.- Disimilaridades y distancias.
- 2.3.- Propiedades generales del MDS.
- 2.4.- Representación mediante árboles ultramétricos.
- 2.5.- Representación mediante grafos piramidales.
- 2.6.- Ventajas de los árboles aditivos frente los ultramétricos.

2.1.- INTRODUCCION

En este capítulo se exponen algunos resultados generales sobre los distintos modelos de representación que son de especial interés para el posterior desarrollo del tema, sirviéndonos además para presentar de un modo más detallado los diferentes métodos a que hemos hecho referencia en el capítulo anterior.

En la primera parte se analizan las principales propiedades sobre los distintos tipos de disimilaridades y distancias (singular, métrica, ultramétrica, aditiva) que juegan un papel fundamental en las técnicas de representación por árboles, obteniendo así una serie de resultados básicos que utilizamos en el trabajo.

En la segunda parte presentamos detalladamente los distintos modelos de representación que se han citado en el capítulo anterior (Ap. 1.3), pretendiendo en este caso incidir tanto en propiedades generales propias del modelo que estudiamos como en observar los procedimientos más importantes relativos al mismo.

Finalmente argumentamos las ventajas que supone la utilización de árboles aditivos frente árboles ultramétricos. De este modo, pensamos que queda cubierto un primer objetivo como es reflejar el papel de los árboles aditivos como métodos de representación del análisis multivariante en relación con los demás métodos de representación.

2.2.- DISIMILARIDADES Y DISTANCIAS

En este apartado analizamos los principales resultados sobre disimilaridades y distancias que utilizamos en esta memoria. Es importante resaltar las relaciones entre los distintos tipos de distancias que introducimos.

Definición 2.1

Dado S un conjunto finito, a una función $d: S \times S \rightarrow R$ se la llama disimilaridad sobre S si verifica las siguientes condiciones:

$$\begin{array}{ll}
 \text{D1) } d(x,y) \geq 0 & \forall x,y \in S \\
 \text{D2) } d(x,y) = d(y,x) & \forall x,y \in S \\
 \text{D3) } d(x,x) = 0 & \forall x \in S
 \end{array} \quad (1)$$

Definimos a continuación los tipos de disimilaridades que nos interesa relacionar:

Definición 2.2

Una disimilaridad d es métrica si verifica las condiciones (1) y la desigualdad triangular

$$\text{D4) } d(x,y) \leq d(x,z) + d(z,y) \quad \forall x,y,z \in S \quad (2)$$

Una disimilaridad d es ultramétrica si verifica las condiciones (1) y la desigualdad ultramétrica

$$\text{D5) } d(x,y) \leq \max \{d(x,z), d(y,z)\} \quad \forall x,y,z \in S \quad (3)$$

Si verifica las condiciones (1) y además se cumple la condición del cuarto punto, es decir

$$D6) d(x,y) + d(z,t) \leq \max \{ d(x,z) + d(y,t), d(x,t) + d(y,z) \} \quad (4)$$

para todo $x,y,z,t \in S$, decimos que d es aditiva.

Definición 2.3

Si d es una disimilaridad sobre S que cumple

$$D7) d(x,y) = 0 \quad \text{si y sólo si } x = y \quad (5)$$

decimos que es una disimilaridad definida.

Es inmediato pues que una disimilaridad métrica definida es una distancia métrica o distancia sobre S .

De las definiciones anteriores se deducen los siguientes resultados:

Proposición 2.1

Si d es una disimilaridad ultramétrica y existen x,y,z de S tales que

$$d(x,y) \leq d(x,z) \leq d(y,z)$$

entonces

$$d(x,z) = d(y,z) \quad (6)$$

Demostración:

Al ser d ultramétrica, utilizando las hipótesis

$$d(y,z) \leq \max \{ d(y,x), d(x,z) \} = d(x,z)$$

con lo que queda demostrado (6).

Recíprocamente, si dados cualesquiera tres elementos x, y, z de S podemos escoger un par (x, y) para el cual

$$d(x,y) \leq d(x,z) = d(y,z) \quad (7)$$

nos resulta que d es ultramétrica.

Proposición 2.2

Si d es una disimilaridad aditiva y existen x, y, z, t de S tales que

$$d(x,y) + d(z,t) \leq d(x,z) + d(y,t) \leq d(x,t) + d(y,z)$$

entonces

$$d(x,z) + d(y,t) = d(x,t) + d(y,z) \quad (8)$$

La demostración es idéntica a la anterior, utilizando (4).

Recíprocamente, si dados cualesquiera cuatro elementos x, y, z, t de S podemos escoger dos pares para los cuales

$$d(x,y) + d(z,t) \leq d(x,z) + d(y,t) = d(x,t) + d(y,z) \quad (9)$$

es inmediato comprobar que d es aditiva.

Intuitivamente, dada una disimilaridad ultramétrica, el "triángulo" que forman tres puntos es isósceles. Del mismo modo, dada una disimilaridad aditiva y considerando el tetraedro que forman cuatro puntos, existen dos pares de aristas opuestas cuya suma es la misma.

Las relaciones entre los distintos tipos de disimilaridad pueden resumirse en las siguientes propiedades:

Proposición 2.3

Si d es una disimilaridad ultramétrica entonces es disimilaridad métrica.

Demostración:

Dados x, y, t de S ,

$$d(x,y) \leq \max \{ d(x,t), d(y,t) \}$$

de donde se sigue inmediatamente

$$d(x,y) \leq d(x,t) + d(y,t)$$

Proposición 2.4

Si d es una disimilaridad aditiva entonces es también disimilaridad métrica.

Demostración:

Dados $x, y, z \in S$, a través de la definición 2.2

$$d(x,y) + d(z,z) \leq \max \{ d(x,z) + d(y,z), d(x,z) + d(y,z) \}$$

y puesto que $d(z,z) = 0$, obtenemos

$$d(x,y) \leq d(x,z) + d(y,z) \quad \text{para } x, y, z \in S$$

Proposición 2.5

Si d es disimilaridad ultramétrica entonces es disimilaridad aditiva.

Demostración:

Sean x, y, z, t elementos de S y supongamos que

$$d(x, y) \leq \min \{d(x, z), d(x, t), d(y, z), d(y, t), d(z, t)\}$$

Por ser d una disimilaridad ultramétrica se verifica que

$$d(x, y) \leq d(x, z) = d(y, z) \quad (10)$$

y

$$d(x, y) \leq d(x, t) = d(y, t) \quad (11)$$

con lo que

$$d(x, z) + d(y, t) = d(y, z) + d(x, t) \quad (12)$$

Por otro lado

$$d(z, t) \leq \max \{d(y, z), d(y, t)\}$$

por lo que puede ocurrir que

$$d(z, t) \leq d(y, z) \quad (13)$$

ó

$$d(z, t) \leq d(y, t) \quad (14)$$

Si estamos en la situación (13), a partir de (11)

$$d(x, y) + d(z, t) \leq d(x, t) + d(y, z) = d(x, z) + d(y, t)$$

Del mismo modo si estamos en (14), a partir de (10)

$$d(x, y) + d(z, t) \leq d(x, z) + d(y, t) = d(x, t) + d(y, z)$$

con lo cual dados cuatro elementos siempre podemos elegir dos pares tales que cumplan la condición (4). Así pues, utilizando (9) obtenemos que d es una disimilaridad aditiva.

A partir de la proposición 2.3 se deduce que una disimilaridad ultramétrica definida y una disimilaridad aditiva definida son distancias, y en lo sucesivo para referirnos a ellas las llamaremos distancias ultramétricas y distancias aditivas.

Cabe resaltar que las proposiciones 2.1, 2.2, 2.4, y 2.5 se cumplen también para distancias.

Podríamos resumir las proposiciones 2.3, 2.4, y 2.5 en el siguiente cuadro de implicaciones:

$$\begin{array}{ccc}
 & \Rightarrow & \text{D4)} \\
 \text{D5)} & & \uparrow \\
 & \Rightarrow & \text{D6)}
 \end{array}$$

Definición 2.4

A una función $d: S \times S \rightarrow R$ tal que existe $f: S \rightarrow R$ de manera que

$$d(x,y) = \begin{cases} f(x) + f(y) & \text{si } x \neq y, x,y \in S \\ 0 & \text{si } x = y \end{cases}$$

la llamamos una función singular definida sobre S . Si además $d(x,y) \geq 0$, $\forall x,y \in S$, diremos que d es una disimilaridad singular.

Proposición 2.6

Si d es distancia aditiva y d_s disimilaridad singular entonces $d + d_s$ es una distancia aditiva.

Demostración:

$$(d+d_s)(x,y) + (d+d_s)(z,t) = d(x,y) + d(z,t) + f(x)+f(y)+f(z)+f(t)$$

y por ser d aditiva

$$d(x,y) + d(z,t) \leq \max \{ d(x,z) + d(y,t), d(x,t) + d(y,z) \}$$

de donde

$$(d+d_s)(x,y)+(d+d_s)(z,t) \leq \max \{ d(x,z)+d(y,t)+f(x)+f(y)+f(z)+f(t), \\ d(x,t)+d(y,z)+f(x)+f(y)+f(z)+f(t) \}$$

es decir

$$(d+d_s)(x,y)+(d+d_s)(z,t) \leq \max \{ (d+d_s)(x,t)+(d+d_s)(y,z), \\ (d+d_s)(x,z)+(d+d_s)(y,t) \}$$

Corolario:

Si d es ultramétrica y d_s disimilaridad singular entonces $d+d_s$ es aditiva.

Proposición 2.7

Si d es distancia ultramétrica tiene a lo sumo $n-1$ valores distintos siendo n el cardinal de S ($|S| = n$).

Demostraciones distintas pueden encontrarse en Lerman (1981) y Arcas (1983).

En Arcas (1983) y Salicrú (1983) se explica como se puede dotar al conjunto de disimilaridades y distancias de una estructura de espacio métrico; asimismo se explica la conveniencia de utilizar en mu---

chos casos una distancia como un vector $\binom{n}{2}$ -dimensional siendo $|S| = n$ en vez de la notación en forma matricial.

Otro resultado particularmente interesante es:

Proposición 2.8

Si d es distancia ultramétrica y $S = \{s_1, \dots, s_n\}$ podemos obtener una ordenación de la matriz (d_{ij}) con $d_{ij} = d(s_i, s_j)$ asociada a d en la forma

$$a) \quad d_{ij} \leq d_{ij+1} \quad \text{y} \quad d_{i+1 j} \leq d_{ij} \quad (15)$$

$$b) \quad \forall k, \text{ de } d_{kk+1} = \dots = d_{kk+s+1} < d_{kk+s+2}$$

se deduce

$$d_{k+1j} = d_{kj} \quad \text{para } j > k+s+1 \quad \text{siendo } s \geq 0$$

que nos permite observar una forma general para las matrices asociadas a una distancia ultramétrica (Lerman, 1981), (Salicrú, 1983).

Tal como se ha expuesto anteriormente, el objetivo de esta memoria es el estudio de la representación en modelos de redes. Sin embargo, es interesante hacer alguna referencia a los modelos espaciales para no perder la visión de conjunto en el estudio de las técnicas de representación. Así en el apartado siguiente damos una relación de los resultados fundamentales de MDS.

2.3.- PROPIEDADES GENERALES DEL MDS

Como hemos expuesto en el capítulo I, Multidimensional Scaling (MDS) es un término general para un conjunto de métodos que tienen por objetivo la búsqueda de una configuración de puntos para representar un conjunto en un espacio R^p a partir de una disimilaridad definida sobre el mismo. Nos propondremos en este apartado analizar algunas de las principales técnicas MDS.

2.3.1.- Modelo MDS para una distancia euclídea

Dado (S, δ) siendo S un conjunto de objetos y δ una disimilaridad definida sobre S , si existe

$$\phi : (S, \delta) \longrightarrow (R^p, || \quad ||)$$

siendo $|| \quad ||$ la norma euclídea, de modo que

$$|| \phi(x) - \phi(y) || = \delta(x,y)$$

decimos que δ es una distancia euclídea y obtenemos en este caso una solución MDS en R^p que reproduce exactamente la disimilaridad inicial.

Así pues, el problema principal radica en conocer las condiciones que debe verificar una distancia $\delta_{ij} = \delta(s_i, s_j)$ para ser euclídea, pudiendo citarse en este sentido los siguientes resultados.

Teorema 2.1

Sea (δ_{ij}) una distancia sobre el conjunto de objetos S con n elementos. Si consideramos las matrices H y A definidas como

$$h_{ij} = \begin{cases} 1 - 1/n & \text{si } i = j \\ -1/n & \text{si } i \neq j \end{cases}$$

$$a_{ij} = -\frac{1}{2} \cdot \delta_{ij}^2$$

y

$$B = H.A.H$$

se verifica que δ es euclídea si y sólo si B es semidefinida positiva.

Así pues, si diagonalizamos B tenemos

$$B = T.D.T'$$

con T matriz ortogonal y D matriz diagonal formada por los valores propios en la diagonal principal y ceros en el resto, obteniéndose

$$B = X . X'$$

siendo X una matriz $n \times p$, p el número de valores propios distintos de cero y representando las filas las coordenadas de los n puntos en R^p . Estas coordenadas reciben el nombre de "coordenadas principales".

Teorema 2.2

Dado S conjunto finito de cardinal igual a n y δ una disimilaridad definida sobre S , si consideramos

$$A = \left(- \frac{1}{2} \cdot \delta_{ij}^2 \right) \quad \text{y} \quad e' = (1, \dots, 1)$$

se verifica que δ es euclídea si y sólo si se verifica:

$$\left. \begin{array}{l} \forall s \in \mathbb{R}^n \text{ tal que } s'e = 1 \\ s'A \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} B_S = (I - es').A.(I - se') \\ \text{es semidefinida positiva} \end{array} \quad (16)$$

La demostración de este teorema (Gower, 1982) nos conduce a interesantes propiedades de la configuración. En este sentido si Y es una matriz que nos define una configuración euclídea en S

$$\delta_{ij}^2 = \sum_{k=1}^m (y_{ik} - y_{jk})^2 \quad i, j \in \{1, \dots, n\}$$

y consideramos una traslación de la forma

$$Z = Y - e \cdot s' \cdot Y$$

verificándose

$$s' \cdot e = 1$$

$$s' \cdot A \neq 0$$

La matriz de productos escalares $B = Z.Z^t$ ($\delta_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$) es igual a B_s ,

$$B = B_s = (I - es').A.(I - se')$$

Observamos que esta propiedad es independiente de la configuración de puntos Y que disponemos de entrada. Esto permite estudiar algunas cuestiones geométricas, como es el hecho de poder conseguir una traslación de modo que el origen de coordenadas verificara ser algún punto especial. En este sentido se pueden demostrar los siguientes resultados:

- a) Si $s' = (0, \dots, 0, 1, 0, \dots, 0)$ tenemos que el origen de coordenadas es P_i (coordenadas del i -ésimo punto).
- b) Si $s' = (1/n, \dots, 1/n)$ el origen de coordenadas es el baricentro de la configuración. Así, en el caso de representación geométrica utilizando coordenadas principales tenemos como una característica geométrica de la representación que el origen de coordenadas es justamente el baricentro.
- c) Si existe A^{-1} para $s = \frac{1}{e'D^{-1}e} . A^{-1} . e$ el origen de coordenadas es el circuncentro de la configuración, con lo cual el radio de la hiperesfera en que estarían situados los puntos sería $r = (1/ e'.A^{-1}.e)^{1/2}$.

Teorema 2.3

δ se puede representar en R^p si cualesquiera $p+3$ puntos de S se pueden representar en R^p (De Leeuw, 1982).

2.3.2.- Caso no euclídeo

Método de las coordenadas principales (Torgersson, 1952, 1958)

Es el método más clásico del MDS. Si (δ_{ij}) no es euclídea y la matriz B asociada tiene por valores propios

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s \geq 0 \geq \lambda_{s+2} \geq \dots \geq \lambda_n$$

tomamos la matriz X de orden $n \times s$ de modo que las columnas están formadas por los vectores propios de valores propios $\lambda_1, \dots, \lambda_s$ normalizados y consideramos

$$B^* = X.X'$$

que será semidefinida positiva de rango s , minimizando $D = \text{traz}(B - B^*)^2$ para la matriz B fijada y B^* semidefinida positiva, según resultado general de Eckart y Young(1936). Se verifica que para $B^* = X.X'$, $D = \lambda_{s+2}^2 + \dots + \lambda_n^2$ y para una representación en dimensión $m < s$ el mínimo es

$$D = \sum_{i=m+1}^s \lambda_i^2 + \sum_{k=s+2}^n \lambda_k^2$$

valor que se puede considerar como una medida de distorsión.

Método de Shepard (1962)

Shepard (1962) desarrolló la idea de la construcción de soluciones de modo que la representación euclídea de los objetos respeten

las relaciones de orden en el vector de distancias observado. Es decir, si consideramos la preordenación en $S \times S$

$$(i,j) \leq (i',j') \quad \Leftrightarrow \quad \delta_{ij} \leq \delta_{i'j'}$$

la solución MDS (d_{ij}) es aconsejable que respete dicha preordenación de una forma exacta o aproximada.

El método de Shepard constituye el primer algoritmo MDS no métrico y de este modo fue el primero en mostrar como soluciones métricas se podían obtener a partir de datos ordinales. El método empieza suponiendo los n objetos como vértices de un símplex $n-1$ dimensional, y en una primera fase deforma el mismo en base a hacer crecer las distancias grandes y decrecer las pequeñas. En una segunda fase se proyecta el poliedro resultante en un espacio con el número de dimensiones deseado y mediante un proceso iterativo se pretenden ajustar los datos iniciales con las distancias espaciales resultantes. El proceso consiste en construir para cada punto $n-1$ vectores con dirección los demás puntos, sentido distinto según la distancia sea mayor o menor que la observada y magnitud proporcional al desajuste, moviendo el punto en la dirección de la resultante de los $n-1$ puntos una distancia proporcional a la magnitud; se reitera el proceso hasta llegar al equilibrio. Podemos ver en Kruskal (1977) la comprobación de la convergencia del método.

Transformaciones monótonas

Existen algunos métodos basados en la transformación de la matriz de disimilaridad por una función monótona de modo que nos resulte una distancia euclídea.

Como resultados interesantes en este sentido cabría citar los siguientes:

Teorema 2.4

Si δ no es euclídea, existe una constante α tal que,

$$\delta_{ij}^* = \sqrt{\delta_{ij}^2 + \alpha(1 - \delta^{ij})} \quad (17)$$

siendo δ^{ij} los deltas de Kronecker, es una distancia euclídea pudiendo inducirse en un espacio euclídeo de dimensión $n-2$.

Una demostración del teorema sirvió a Lingóes(1971) para obtener una solución MDS que realizaba exactamente la preordenación sobre el conjunto de disimilaridades inicial, considerando como constante α el doble del valor absoluto del menor valor propio de la matriz B.

Sin embargo, existe el problema del aumento de la distorsión D al disminuir la dimensión. En el caso de la solución de Lingoes la distorsión es

$$D = (n - 1) \cdot \lambda_n^2$$

y en dimensión reducida tomando las m primeras coordenadas

$$D = \sum_{i=m+1}^{n-2} \lambda_i^2 + (m+1) \cdot \lambda_n^2$$

(Cuadras, 1981)

Mardia (1978) utilizando la misma idea propone una solución a fin de disminuir la distorsión. Puede adaptarse a una dimensión inferior esta solución, aunque a costa de que la solución no respete exactamente la preordenación inicial.

Otros métodos parecidos pueden verse en Cooper (1972) y Cailliez (1983).

No está resuelto el problema del cálculo de una función f tal que $f(\delta_{ij})$ sea euclídea en dimensión mínima.

Método de Kruskal

Según De Leeuw(1982) uno de los principales efectos que produjeron los trabajos de Shepard fue el interés que Joseph Kruskal dedicó al MDS en un afán de mejorar los primeros. Realizó una mejora de los procedimientos de Shepard siguiendo sus mismas ideas. En este sentido se introduce :

a) La noción de optimización de una medida de ajuste explícitamente definida.

b) Utilización de un procedimiento numérico.

La idea general del método consiste en ajustar una distancia euclídea d_{ij} a la transformación $f(\delta_{ij}) = \hat{d}_{ij}$ donde f es una función monótona tal que la medida de ajuste que la llama STRESS

$$\Lambda = \frac{\sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2} \quad (18)$$

sea mínima.

Al no conocerse una solución exacta al problema, este ha sido tratado por métodos numéricos. Así, Kruskal propone como algoritmo general el siguiente: partir de una configuración euclídea inicial obtenida de alguna forma (por ejemplo, coordenadas principales) en una dimensión determinada, obteniendo una transformación monótona de la matriz de distancias inicial que minimice el STRESS (Regresión monótona). De este modo volvemos a obtener una configuración euclídea asociada a la minimización anterior, siguiendo el proceso iterativo hasta que la función converge a un mínimo (que puede ser local).

El algoritmo de Kruskal fue originalmente implementado en un programa llamado MDSCAL. Recientemente algunas características de otros programas llamados TORSCA se combinaron con MDSCAL para producir un programa llamado KYST (el nombre está constituido por las iniciales de los 4 principales contribuidores a la teoría del MDS, Kruskal, Young, Shepard, Torgersen).

Otros procedimientos MDS basados en ideas parecidas con el nombre general de Smallest Space Analysis (SSA) fueron independientemente desarrollados por Guttman and Lingoes. Una comparación de SSA con ADDTREE puede verse en Tversky (1977).

Un problema abierto muy importante es decidir el número de dimensiones apropiado. Algunos criterios en este sentido son desarrollados en Wagenaar & Padmos (1971), Tuckey (1977), Kruskal y Wish (1978).

Tal y como hemos comentado anteriormente las principales técnicas de representación mediante modelos de redes son: árboles aditivos, árboles ultramétricos y grafos piramidales. Puesto que esta memoria trata en profundidad la representación mediante árboles aditivos, trataremos en los siguientes apartados sobre las principales propiedades relativas a árboles ultramétricos y grafos piramidales.

2.4.- REPRESENTACION MEDIANTE ARBOLES ULTRAMETRICOS

La idea general es la deformación de la disimilaridad dada sobre el conjunto de objetos hasta conseguir una distancia ultramétrica que se ajuste a la primera desde algún punto de vista. Estas técnicas se conocen con el nombre de taxonomía numérica.

Definición 2.9

Dado S conjunto finito, un par (J, i) siendo $J \subset P(S)$ e

$$i : J \longrightarrow \mathbb{R}^+$$

diremos que es una jerarquía indexada si cumple las siguientes condiciones:

- a) Si $A, B \in J$ entonces $A \cap B \in \{A, B, \emptyset\}$
- b) Si $A \in J$ entonces $\bigcup \{C \mid C \in J, C \subsetneq A\} \in \{A, \emptyset\}$ (19)
- c) $i(\{s\}) = 0$ si $\{s\} \in J$
- d) Si $A \subset B$ entonces $i(A) \leq i(B)$

La obtención de una distancia ultramétrica d_u sobre el conjunto de objetos equivale a la obtención de una jerarquía indexada de partes (J,i) sobre los mismos (Cuadras, 1981) que queda definida como el conjunto de partes de S verificando la siguiente condición: Un subconjunto A de S pertenece a la jerarquía si y sólo si

$$\exists r \in \mathbb{R} \quad | \quad \forall x,y \in A \quad , \quad d_u(x,y) \leq r \quad (20)$$

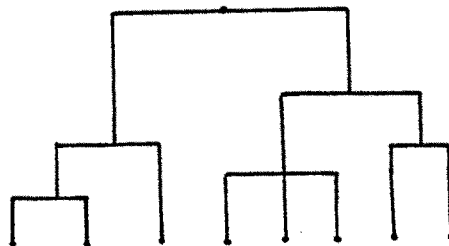
siendo el valor del índice de la jerarquía para el conjunto A el ínfimo de los valores reales r que verifican la condición anterior.

Para todo r podemos considerar la relación

$$x R_r y \Leftrightarrow d_u(x,y) \leq r$$

que es una relación de equivalencia por ser d_u una distancia ultramétrica, obteniéndose pues una partición del conjunto S para cada nivel r .

Así pues, la representación natural asociada a una jerarquía indexada es un dendograma que puede interpretarse como un grafo simplemente conexo sin ciclos con un nudo interno equidistante de los extremos (fig.1)



El concepto de jerarquía indexada es equivalente al de clasificación jerárquica en el sentido de considerarla como una función

$$\psi : \mathbb{R}^+ \longrightarrow P$$

siendo P el conjunto de particiones de S , verificando las siguientes

condiciones:

- 1) $\psi(0) \supset \{ \{s_1\}, \dots, \{s_n\} \} \quad \forall s_i \in S$
- 2) Si $r < r'$ entonces $\psi(r) \subset \psi(r')$ (más fina) (21)
- 3) $\exists r \in R^+$ tal que $\psi(r) = S$

Puede formalizarse la taxonomía numérica a través de la utilización de clasificaciones jerárquicas (Arcas, 1983). Podemos observar como la taxonomía numérica tendría por objetivo la construcción de una colección de clases naturales que podría interpretarse en un sentido evolutivo y utilizarse para la formación de clasificaciones razonables de los objetos.

Ejemplos de la utilización práctica de este método pueden verse en Sokal (1963), Cuadras (1981), Salicrú (1983) etc... El proceso de deformación de la disimilaridad inicial puede venir dado de muy diversas maneras. Un procedimiento general viene dado a través de lo que se conoce como algoritmo fundamental de clasificaciones jerárquicas que puede describirse mediante la siguiente recurrencia (Arcas y Salicrú, 1984):

$$\text{Si } l_1 = \min \{ \delta(s_i, s_j) \mid s_i \neq s_j, s_i, s_j \in S \} = \delta(s_{i0}, s_{j0})$$

es la mínima distancia entre los elementos de S , definimos

$$\psi(l_1) = \{ \{s_1\}, \dots, \{s_{i0}\} \cup \{s_{j0}\}, \dots, \{s_n\} \}$$

y consideramos una función $f_1: R^3 \longrightarrow R$ de forma que

$$f_1(\delta(s_{i0}, s_{j0}), \delta(s_{i0}, s_k), \delta(s_{j0}, s_k)) \geq l_1 \quad \forall s_k \in S$$

A $\psi(l_1)$ le asociamos la disimilaridad definida como sigue:

$$\delta_1(\{s_i\}, \{s_j\}) = \delta(s_i, s_j) \quad \text{si} \quad \{s_i, s_j\} \cap \{s_{i0}, s_{j0}\} = \emptyset$$

$$\delta_1(\{s_{i0}\} \cup \{s_{j0}\}, \{s_k\}) = f_1(\delta(s_{i0}, s_{j0}), \delta(s_{i0}, s_k), \delta(s_{j0}, s_k)) \geq l_1$$

de forma que la distancia entre los elementos que no se han juntado se mantiene invariante y la que separa la clase formada por los dos elementos que se han juntado con un tercer elemento viene determinada por la función f_1 . Podemos pasar entonces del par (S, δ) al par $(\psi(l_1), \delta_1)$ y de forma análoga a $\psi(l_2), \dots, \psi(l_s)$ considerando sucesivamente funciones f_2, \dots, f_{s-1} . Podemos entonces construir la clasificación jerárquica ψ en la forma

$$\text{a) } \psi(\alpha) = \psi(l_{m-1}) \quad \text{si} \quad \alpha \in [l_{m-1}, l_m)$$

$$\text{b) } \psi(\alpha) = \{\{s_1\}, \dots, \{s_n\}\} \quad \text{si} \quad l_1 \neq 0 \quad \text{y} \quad \alpha \in [0, l_1)$$

$$\text{c) } \psi(\alpha) = \{S\} \quad \text{si} \quad \alpha \geq l_s$$

Veamos algunos procedimientos que pueden ser construidos a través de este proceso.

Método del mínimo (máximo)

(Johnson, 1967)

Se considera

$$f_n(\alpha, \beta, \gamma) = \min \{ \beta, \gamma \} \quad (\max \{ \beta, \gamma \}) \quad (22)$$

de este modo, si

$$\psi(l_m) = \{ A_1, \dots, A_k \}$$

y

$$l_{m+1} = \delta_m(A_{i_0}, A_{j_0}) \text{ m\u00ednimo valor para } \delta_m$$

resulta

$$\psi(l_{m+1}) = \{ A_1, \dots, A_{i_0} \cup A_{j_0}, \dots, A_k \}$$

verific\u00e1ndose que

$$\begin{aligned} \delta_{m+1}(A_{i_0} \cup A_{j_0}, A_k) &= \min \{ \delta_m(A_{i_0}, A_k), \delta_m(A_{j_0}, A_k) \} \\ & \quad (\max \{ \delta_m(A_{i_0}, A_k), \delta_m(A_{j_0}, A_k) \}) \quad (23) \end{aligned}$$

$$\delta_{m+1}(A_i, A_j) = \delta_m(A_i, A_j) \quad \text{si } [i, j] \cap [i_0, j_0] = \emptyset$$

La idea geom\u00e9trica consiste en la deformaci\u00f3n de un tri\u00e1ngulo hasta obtener dos lados iguales que coincidan con el menor (mayor) de los lados que no son base.

Método UPGMA (Sokal y Michener, 1958)

$$\text{Si } \psi(l_m) = \{A_1, \dots, A_k\}$$

y $l_{m+1} = \delta_m(A_{i0}, A_{j0})$ mínimo valor para δ_m , resulta que

$$\psi(l_{m+1}) = \{A_1, \dots, A_{i0} \cup A_{j0}, \dots, A_k\}$$

verificándose que δ_m inducirá δ_{m+1} de la forma:

$$\delta_{m+1}(A_{i0} \cup A_{j0}, A_k) = \frac{N_i}{N_i + N_j} \cdot \delta_m(A_{i0}, A_k) + \frac{N_j}{N_i + N_j} \cdot \delta_m(A_{j0}, A_k) \quad (24)$$

siendo $N_i = |A_{i0}|$ y $N_j = |A_{j0}|$

La idea geométrica es la deformación de un triángulo hasta obtener dos lados iguales que coinciden con la media de los lados mayores ponderada respecto los cardinales de las clases.

Este método tiene importantes propiedades relativas al coeficiente de correlación cofenética (medida de ajuste entre la disimilitud inicial y la ultramétrica), tal y como se puede observar en Farris (1969) y Arcas (1983). También es el método más utilizado en las aplicaciones prácticas. Ejemplos de estas aplicaciones las tenemos en Dallot (1972), Rios (1985).

Otros resultados referentes a: la obtención de la disimilitud δ_m en términos de δ , la continuidad de los métodos expuestos, la distancia que separa dos ultramétricas asociadas a la misma disimilitud inicial pueden verse en Arcas y Salicrú (1984).

Otros métodos jerárquicos interesantes son: de la media (Sokal y Michener, 1958), de la mediana (Gower, 1967), del centroide (Sokal y Michener, 1958). También han sido estudiados métodos de deformación basados en otros puntos de vista. Cabría citar por ejemplo los métodos B_k, B_k^C y D_u (Jardine & Sibson, 1971), (Matula, 1977); métodos de tipo iterativo no jerárquico como puede ser ISODATA (Ball & Hall, 1967), k-medias.

Un problema abierto muy interesante en taxonomía numérica es la determinación del número de clases realmente significativas. No se ha llegado a resultados muy concretos en este sentido, aunque algunas técnicas se han desarrollado para resolver este problema. Cabría citar los trabajos de Rohlf (1970), Anderson (1985), Hubert & Arabie (1985).

2.5 - REPRESENTACION POR GRAFOS PIRAMIDALES

La representación mediante grafos piramidales es una extensión de los métodos de taxonomía numérica. Es una técnica desarrollada recientemente por Bertrand & Diday (1985) que parte de un conjunto S de objetos, un orden fijado sobre el mismo y una disimilaridad δ , la cual está basada en la deformación de dicha disimilaridad observada sobre el conjunto de objetos hasta obtener una distancia d que conserve el orden en el siguiente sentido:

Si $s_1 \leq s_2 \leq s_3$ entonces

$$d(s_1, s_3) \geq d(s_1, s_2)$$

y

$$d(s_1, s_3) \geq d(s_2, s_3)$$

(25)

Definición 2.10

Una disimilaridad definida que cumple (25) se la denomina índice piramidal. Podemos observar que si d es ultramétrica existe una ordenación de los elementos tal que d conserva este orden.

Definición 2.11

Dado un orden sobre S , diremos que una parte C es conexa si el conjunto formado por todos los elementos entre cualesquiera dos de C es de C . Así, un orden es compatible con un conjunto $P \subset P(S)$ si cada elemento de P es conexo con respecto al mismo.

Definición 2.12

Dado S conjunto y $P \subset P(S)$, P será una pirámide si:

- 1) $S \in P$
- 2) $\forall \omega \in S, \{\omega\} \in P$
- 3) $\forall \{A, B\} \subset P, A \cap B = \emptyset \text{ ó } A \cap B \in P$ (26)
- 4) Existe un orden compatible con P
- 5) $\exists f: P \longrightarrow \mathbb{R}^+$ tal que
 - a) $f(p) = 0$ si p tiene un solo elemento.
 - b) $\forall p, p' \in P, p \subset p' \Rightarrow f(p) \leq f(p')$

A la representación la llamaremos grafo piramidal indexado.

Definición 2.13

Dados $p, p' \in P$ diremos que p' es un predecesor de p si $p \subset p'$ y no existe ningún p'' tal que $p \subset p'' \subset p'$. Los principales resultados que se obtienen son:

Proposición 2.9

El conjunto de jerarquías indexadas está incluido dentro de las pirámides.

Proposición 2.10

Cada elemento de P no tiene más de 2 predecesores.

Proposición 2.11

Existe una biyección entre los índices piramidales y los grafos piramidales indexados (i.e. pirámides).

Por ejemplo si d es una disimilaridad con matriz asociada

$$\begin{pmatrix} 0 & 4 & 4 & 5 \\ & 0 & 2 & 5 \\ & & 0 & 3 \\ & & & 0 \end{pmatrix}$$

podemos representarla mediante la "pirámide" (ver fig.2)

{ {1}, {2}, {3}, {4}, {2,3}, {1,2,3}, {3,4}, {1,2,3,4} }

con

$f(\{2,3\}) = 2$; $f(\{1,2,3\}) = 4$; $f(\{3,4\}) = 3$ y $f(\{1,2,3,4\}) = 5$

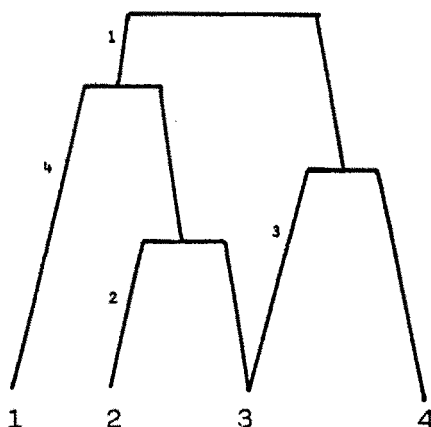


Figura 2.- Representación gráfica de una pirámide.

2.6.- VENTAJAS DE LOS ARBOLES ADITIVOS FRENTE LOS ULTRAMÉTRICOS

La representación mediante árboles ultramétricos produce en general restricciones muy importantes tales como: a) la distancia ultramétrica entre elementos de dos clases distintas a cualquier nivel debe ser la misma. b) Distancias entre elementos de una misma clase no puede superar la distancia entre elementos de esta y cualquier otra.

Estas restricciones no se dan en árboles aditivos, con lo cual la representación se puede adaptar de una forma sustancialmente mejor a los datos originales, pudiéndose por otro lado interpretar más fácilmente el árbol resultante. Por ejemplo, Salicrú (1983) en un estudio sobre la distribución de *Drosophila Subobscura* demuestra que las barreras geográficas

cas son barreras genéticas, tomando como distancia genética (Prevosti, 1974).

$$\delta(a,b) = \frac{1}{2r} \cdot \sum_{j=1}^r \sum_{h=1}^{s_j} |p_{ajh} - p_{bjh}| \quad (27)$$

siendo:

r = número de cromosomas distintos.

s_j = número de ordenaciones distintas en el cromosoma j .

p_{ajh} = proporción de la ordenación del cromosoma j en la población a .

p_{bjh} = idem. en la población b .

En este sentido, si intentamos probar que el estrecho de Drover es una barrera genética, se consideran las poblaciones de Heriot (H), Dalkeith (D), Gröningen (G), Viena (V), Zurich (Z), Drobak (Dr) que tienen por matriz de disimilaridades asociadas

| | H | D | G | V | Z | Dr |
|----|---|-------|-------|-------|-------|-------|
| H | 0 | 0.083 | 0.290 | 0.399 | 0.331 | 0.307 |
| D | | 0 | 0.276 | 0.370 | 0.3 | 0.307 |
| G | | | 0 | 0.187 | 0.112 | 0.152 |
| V | | | | 0 | 0.128 | 0.260 |
| Z | | | | | 0 | 0.235 |
| Dr | | | | | | 0 |

(28)

utilizando el método UPGMA se obtiene la clasificación jerárquica

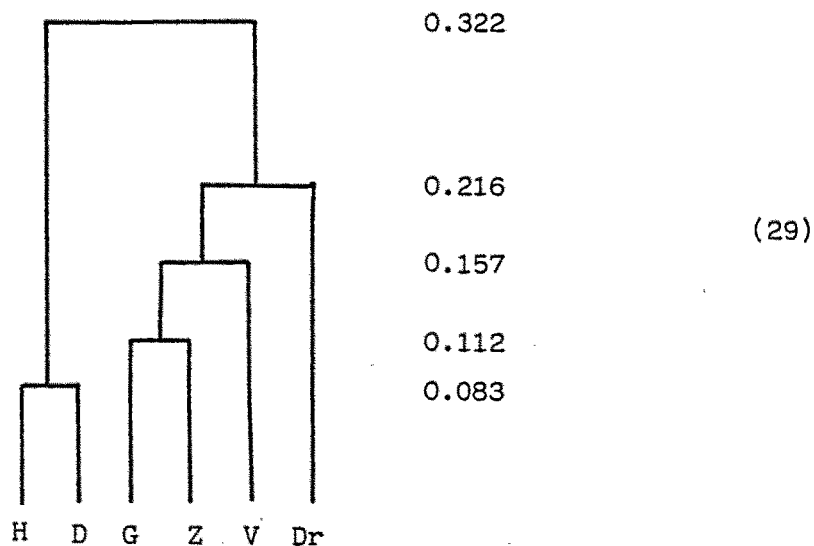


Figura 3

a partir de la cual se justifica la hipótesis que se formulaba. Sin embargo, observamos que la distancia ultramétrica u de la población Dr a las poblaciones de G, Z y V es la misma:

$$u(Dr, G) = u(Dr, Z) = u(Dr, V) = 0.216$$

lo cual no está excesivamente de acuerdo con los datos obtenidos a través de δ , puesto que

$$\delta(Dr, G) = 0.152$$

$$\delta(Dr, Z) = 0.235$$

$$\delta(Dr, V) = 0.26$$

(30)

siendo la primera sensiblemente distinta a las otras dos. Por otro lado,

la distancia observada entre Groningen y Viena es mayor que aquella entre Drobak y Groningen, y sin embargo para la ultramétrica resultante

$$u(G, V) = 0.157$$

$$u(G, Dr) = 0.216$$

observando aquí un aspecto contradictorio entre la disimilaridad inicial y la distancia ultramétrica obtenida.

Estos problemas surgen por las restricciones propias de las distancias ultramétricas descritas en a) y b).

Por otro lado el coeficiente de correlación entre la disimilaridad inicial y la estimada es 0.91778.

Utilizando árboles aditivos, mediante el algoritmo introducido por Tversky (1977) obtenemos la representación:

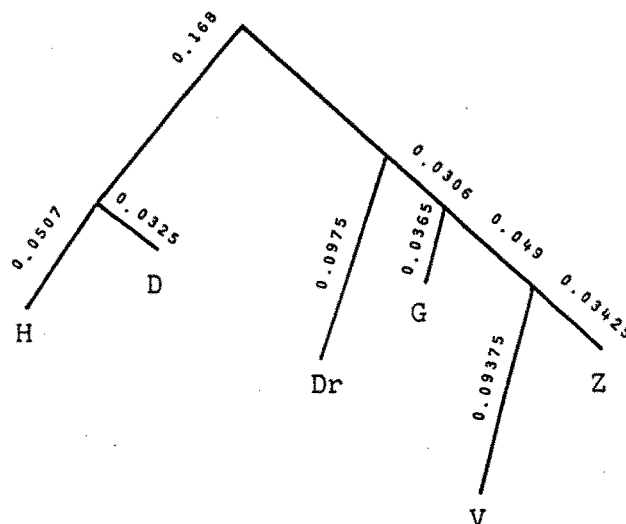


Figura 4

Los problemas descritos anteriormente quedan resueltos pues si d es la distancia aditiva resultante:

$$d(\text{Dr}, \text{G}) = 0.164$$

$$d(\text{Dr}, \text{Z}) = 0.211$$

$$d(\text{Dr}, \text{V}) = 0.270$$

lo que está muy de acuerdo con los datos (30).

Además

$$d(\text{Dr}, \text{G}) = 0.164$$

$$d(\text{G}, \text{V}) = 0.179$$

y observamos como en un árbol aditivo la distancia entre elementos de una clase puede superar la distancia entre elementos de esta clase y externos a la misma.

Además el coeficiente de correlación al cuadrado entre d y δ es 0.9996, lo que supone una mejora sustancial del mismo.

Así, hemos podido observar las importantes ventajas que pueden suponer los árboles aditivos frente los árboles ultramétricos.

Cabe por último señalar que los árboles aditivos se podrían considerar como un paso intermedio entre los clásicos métodos de MDS y la representación por medio de dendogramas puesto que siendo propiamente representaciones en árbol, no están sometidos a las restricciones ultramétricas.

3. REPRESENTACION MEDIANTE ARBOLES ADITIVOS.

Resumen: En el presente capítulo se desarrolla una formalización de la representación asociada a un conjunto sobre el que tenemos definida una distancia que verifica el axioma del cuarto punto.

Sumario:

- 3.1 - Introducción.
- 3.2 - Formalización de P. Buneman.
- 3.3 - Formalización de la representación de un conjunto asociada a una distancia aditiva.
- 3.4 - Relaciones con las distancias ultramétricas.

3.1.- INTRODUCCION

El objetivo que nos proponemos en este capítulo es el estudio específico de la representación de un conjunto de objetos mediante árboles. Para ello estudiamos la relación existente entre una distancia verificando el axioma del 4º punto (distancia aditiva)

$$\forall x,y,z,t \in S$$

$$d(x,y) + d(z,t) \leq \max \{ d(x,z) + d(y,t), d(x,t) + d(y,z) \} \quad (1)$$

y su representación asociada (árbol aditivo) que intuitivamente lo entenderemos como cierto tipo de grafo conexo sin ciclos.

También observamos las relaciones existentes entre las distancias aditivas y las distancias ultramétricas a través de sus representaciones asociadas (árboles aditivos, dendogramas respectivamente) vistas desde la perspectiva de la teoría de grafos. Dejamos para otros capítulos su estudio desde otros puntos de vista.

Hemos dividido el capítulo en dos partes. En la primera tratamos la relación entre una distancia aditiva y su representación, mientras que la segunda se centra en el estudio de las relaciones entre árboles aditivos y dendogramas, obteniendo conclusiones sobre las distancias asociadas.

Buneman (1971) advierte del problema que surge al querer utilizar los árboles aditivos desde un punto de vista eminentemente práctico, en el sentido de situar el interés en la búsqueda de algoritmos más o

menos eficaces para lograr árboles que representen mejor a alguna disimilitud que tengamos dada de entrada. Así pues, centra su estudio en una nueva definición del concepto de árbol que se podría considerar como una caracterización de las relaciones que imponen los nudos y aristas con respecto los elementos del conjunto, dando un método a fin de asociar una distancia sobre el árbol. Describe un algoritmo mediante el cual asigna un árbol aditivo, entendido como el par formado por el árbol y una distancia definida sobre el mismo, a una disimilitud dada sobre el conjunto a representar de forma que la distancia sobre el árbol coincide con la disimilitud inicial si y sólo si es aditiva.

El problema básico de esta formalización estriba en el alejamiento de unos planteamientos intuitivos si pensamos únicamente en la relación distancia aditiva-árbol aditivo, pues en este estudio dicha relación no se plantea como un objetivo en sí, ya que el mismo es el planteo del algoritmo con unas propiedades teóricas interesantes (continuidad, unicidad) en un sentido parecido a como podemos observar el método del mínimo (Johnson, 1967) para representaciones de distancias ultramétricas (Jardine & Sibson, 1971). Dado, sin embargo, el interés de esta formalización, muy citada en trabajos posteriores (Cunningham, 1978; Tversky, 1977; Waterman, 1977, etc) ofrecemos una síntesis de la misma en el presente capítulo.

En líneas generales estudiamos los árboles aditivos desde el punto de vista de su representación como grafo, y no atendiendo tanto a las relaciones que sus elementos (nudos, aristas) producen sobre los elementos del conjunto. Con este planteamiento se podrán tratar cuestiones relativas a la forma del grafo de interés para el estudio y análisis de la representación. Siguiendo esta misma línea se tratan las relaciones entre las distancias aditivas y distancias ultramétricas.

3.2 - FORMALIZACION DE P. BUNEMAN

Tal y como se ha indicado anteriormente comenzaremos con un pequeño repaso al interesante estudio de P. Buneman (1971) que en una de sus partes caracteriza la representación de una distancia aditiva.

Inicialmente se trata de caracterizar un árbol por un procedimiento conjuntista basado en las relaciones que imponen sus aristas y a partir de dichas relaciones definir el concepto de nudo y distancia sobre el árbol obteniendo seguidamente propiedades interesantes sobre la representación del mismo.

Si S es el conjunto a representar, se define:

Definición 3.1

σ es una división de S si y sólo si $\sigma = \{P_1, P_2\}$ con
 $P_1 \cup P_2 = S$ y $P_1 \cap P_2 = \emptyset$ (2)

En este sentido dos elementos de S tales que pertenezcan a conjuntos distintos de una división se llamarán elementos separados por σ .

Definición 3.2

Dadas dos divisiones de S , diremos que son compatibles si alguna de las cuatro intersecciones entre los conjuntos que las forman es vacía.

Es fácil observar la relación existente entre el concepto de división y el de arista. También, y bajo este punto de vista, divisiones asociadas a dos aristas de un árbol serían compatibles. Define entonces el concepto de árbol.

Definición 3.3

Un árbol asociado a S es un conjunto de divisiones de S dos a dos compatibles

$$A = \{\sigma_1, \dots, \sigma_m\} \quad (3)$$

Intuitivamente, un árbol queda dividido en dos partes por una arista, de forma que un nudo pertenece sólo a una de las partes.

En este sentido, y formalizando este concepto, se define un nudo como

Definición 3.4

N es un nudo del árbol $A = \{\sigma_1, \dots, \sigma_m\}$ con $\sigma_k = \{P_k^1, P_k^2\}$ e $i_k \in \{1, 2\}$ si

$$N = \{P_1^{i_1}, \dots, P_m^{i_m}\}$$

de modo que

$$P_k^{i_k} \in \sigma_k$$

y

$$P_k^{i_k} \cap P_j^{i_j} \neq \emptyset \quad \text{si } k \neq j \quad (4)$$

Definición 3.5

Dado el árbol $A = \{\sigma_1, \dots, \sigma_m\}$ se dice que los nudos N_1 y N_2 están conectados por σ_k si son de la forma

$$\begin{aligned} N_1 &= \{P_1^{i_1}, \dots, P_m^{i_m}\} \\ N_2 &= \{P_1^{j_1}, \dots, P_m^{j_m}\} \end{aligned} \quad (5)$$

con $i_s = j_s$ para $s \neq k$ e $i_k \neq j_k$

Tiene entonces sentido hablar de camino como un conjunto de nudos conectados. Podemos también hablar de nudo terminal como aquel nudo que sólo admite una conexión.

Se demuestra que dos nudos de un árbol siempre están conectados por un camino. Asimismo se demuestran otras propiedades de los árboles.

Hasta aquí está caracterizado el árbol en cuanto a su "forma". Es evidente que el estudio formal en sí, planteado de esta manera, se aleja del verdadero sentido intuitivo si bien el pensar en la identificación entre división y arista puede clarificar en gran medida su interpretación.

Ahora se trata de definir una distancia sobre el conjunto S a través de un árbol asociado al mismo.

Definición 3.6

Δ es una distancia definida sobre S asociada al árbol $A = \{\sigma_1, \dots, \sigma_m\}$ si y sólo si

$$\Delta = \sum_{\sigma \in A} \alpha_{\sigma} \cdot \delta_{\sigma} \quad (6)$$

siendo $\alpha_1, \dots, \alpha_m$ números reales y δ_{σ} la pseudométrica definida como

$$\delta_{\sigma}(x,y) = \begin{cases} 1 & \text{si } x,y \in S \text{ están separados por } \sigma \\ 0 & \text{si } x,y \in S \text{ no están separados por } \sigma \end{cases}$$

A partir de este momento el problema queda centrado en la búsqueda de un árbol dotado de una distancia definida como 3.6, asociado a una disimilaridad d dada sobre el conjunto S . Para ello, si

$$\sigma = \{P_1, P_2\}$$

es una división cualesquiera de S , se define

$$\mu_{\sigma} = \frac{1}{2} \cdot \min \{d(x,z) + d(y,t) - d(x,y) - d(z,t)\} \quad (7)$$

siendo $x,y \in P_1$ y $z,t \in P_2$

que intuitivamente, dada la configuración del árbol, sería justamente la longitud de la arista que representa σ si la distancia real del árbol fuera d .

De este modo se considera el conjunto de divisiones

$$A_d = \{ \sigma \mid \mu_\sigma > 0 \} \quad (8)$$

el cual resulta ser un árbol.

A partir de dicho árbol se define la distancia

$$\Delta_d = \sum_{\sigma \in A_d} \mu_\sigma \cdot \delta_\sigma \quad (9)$$

obteniendo el algoritmo mediante el cual a cualquier disimilaridad d se le hace corresponder la representación con árbol A_d y distancia asociada Δ_d .

Se demuestra que dicha representación es única y además verifica

$$\Delta_d \leq d \quad (10)$$

llegándose entonces a

Proposición 3.1

$\Delta_d = d$ si y sólo si d verifica la condición del 4º punto. Con esta proposición queda demostrado que la representación natural asociada a un conjunto sobre el que tenemos definida una distancia verificando la condición del 4º punto es un árbol definido según 3.3 con una distancia asociada en el sentido 3.6.

3.3.- FORMALIZACION DE LA REPRESENTACION DE UN CONJUNTO ASOCIADA A UNA DISTANCIA ADITIVA

En este apartado analizamos la representación de un conjunto de un modo distinto a la formalización anterior. Las diferencias fundamentales entre ambos estudios radican en la definición de árbol y en el objetivo pretendido, puesto que si bien P. Buneman intenta lograr un algoritmo de interés teórico, en nuestro estudio se analizan las relaciones entre las distancias y representaciones como finalidad en sí, cuidando de obtener una formalización que respete las nociones intuitivas sin perder el rigor necesario.

Comenzaremos definiendo el concepto de representación asociada a un conjunto.

Si S es el conjunto de objetos a representar, d una distancia definida sobre S y R otro conjunto finito tal que $S \cap R = \emptyset$, consideramos el conjunto $U = S \cup R$ al que llamamos conjunto de nudos y la existencia de $G_u \subset U \times U$ verificando las siguientes condiciones:

R1.- G_u grafo conexo sin ciclos (Bergé, 1973)

R2.- Si $(x,y) \in G_u$ entonces $(y,x) \in G_u$

R3.- Si $(x,y) \in G_u$ y $(x,z) \notin G_u \quad \forall z \in U \mid z \neq y$
entonces $x \in S$.

Supongamos además dada d^* distancia sobre U tal que verifique:

$$R4.- d^*(x,y) = \min \left\{ \sum_{i=1}^{m-1} d^*(x_i, x_{i+1}) \mid x_1=x, x_m=y, (x_i, x_{i+1}) \in G_u \right\}$$

para $i=1, \dots, m-1$

$$R5.- d^* \Big|_S = d$$

Definición 3.7

Diremos que (G_u, d^*) es una representación asociada al par (S, d) .

Definición 3.8

Dado un grafo G_u verificando las propiedades R1, R2, R3, R4 diremos que es una representación asociada a S .

Definición 3.9

Un camino que une dos elementos x e y de U es un conjunto C de la forma

$$C = \{(x, u_1), (u_1, u_2), \dots, (u_m, y)\}; \text{ de modo que } (x, u_1), (u_i, u_{i+1}), (u_m, y) \in G_u$$

para $i=1, \dots, m-1$ }

Definición 3.10

Dado $x \in U$ que cumple

$$\begin{aligned} 1) \quad & \exists u \in U \quad \text{con} \quad (x,u) \in G_u \\ 2) \quad & \forall z \in U \quad z \neq u \quad (x,z) \notin G_u \end{aligned} \quad (12)$$

diremos que es un nudo terminal.

Definición 3.11

Diremos que una representación (G_u, d^*) asociada a (S, d) es un árbol aditivo de tipo 1 si se verifican las siguientes condiciones:

$$\text{AD 1.} \quad \exists \phi : U \longrightarrow \mathbb{R}^2 \text{ inyectiva}$$

Para $(x,y) \in G_u$, si $\phi = (\phi_1, \phi_2)$ y $\phi_1(x) \leq \phi_1(y)$, consideramos la existencia de

$$\psi : [\phi_1(x), \phi_1(y)] \longrightarrow \mathbb{R}$$

función continua y monótona (si $\phi_1(y) < \phi_1(x)$ consideramos

$$\psi : [\phi_1(y), \phi_1(x)] \longrightarrow \mathbb{R}$$

tal que

$$\psi(\phi_1(x)) = \phi_2(x)$$

$$\psi(\phi_1(y)) = \phi_2(y)$$

y denominaremos por P_{xy} al arco de curva así construido entre las imágenes de x a y ,

$$P_{xy} = \{ (z, \psi(z)) \mid z \in [\phi_1(x), \phi_1(y)] \}$$

y suponemos además en $\bigcup_{(x,y) \in G_u} P_{xy}$ la preordenación

$$(a_1, a_2) \leq (b_1, b_2) \Leftrightarrow a_2 \leq b_2$$

Deberá entonces cumplirse

AD 2.- SI x es nudo terminal y $(x,u) \in G_u$ entonces $\phi(x) \leq \phi(u)$

AD 3.- $\bigcup_{(x,y) \in G_u} P_{xy}$ es un conjunto simplemente conexo.

AD 4.- Para $(x,y) \in G_u$, se debe verificar

$$d^*(x,y) = \text{long}(P_{xy})$$

siendo $\text{long } P_{xy}$ la longitud del arco de curva P_{xy} .

Podemos definir una distancia d' en $\bigcup P_{xy}$ como la longitud del arco que une dos puntos, la cual verificará por la condición anterior:

$$d^*(x,y) = d'(\phi(x), \phi(y))$$

A efectos de notación podemos expresar un árbol aditivo de tipo 1 como

$$AD_1(G_u, d) = \left(\bigcup_{(x,y) \in G_u} P_{xy}, d' \right)$$

Definición 3.12

Sea un árbol aditivo de tipo 1 asociado a un conjunto S y tres elementos x, y, z de S . Podemos considerar el nudo que pertenece a la vez a los caminos xy , yz y xz . Tal nudo es único, pues en caso contrario el grafo no sería simplemente conexo. A un nudo verificando la condición anterior lo llamamos nudo principal.

Vamos a demostrar que las únicas distancias compatibles con árboles aditivos de tipo 1 son las que verifican el axioma del 4º punto.

Lema 3.1

Si S es un conjunto finito y d es una distancia aditiva definida sobre S , se cumple que:

$$\exists (x_0, y_0) \in S \times S \quad \text{tal que}$$

$$d(x_0, y_0) + d(z, t) \leq d(x_0, z) + d(y_0, t) = d(x_0, t) + d(y_0, z)$$

$$\forall (z, t) \in S \times S \quad \text{de modo que } \{x_0, y_0\} \cap \{z, t\} = \emptyset$$

Demostración:

La demostración la efectuamos por inducción sobre el cardinal de S ($|S| = n$), resultando inmediato para el caso en que el mismo sea $n = 4$. Si suponemos cierto el lema para un conjunto de cardinal $n-1$, para un conjunto S de cardinal n , consideramos

$$S' = S - \{y\}$$

siendo y un elemento cualquiera de S .

Por hipótesis de inducción podemos encontrar $(x_0, y_0) \in S' \times S'$ tales que

$$\forall (z, t) \in S' \times S' \quad \text{verificando} \quad \{x_0, y_0\} \cap \{z, t\} = \emptyset$$

$$d(x_0, y_0) + d(z, t) \leq d(x_0, z) + d(y_0, t) = d(x_0, t) + d(y_0, z) \quad (13)$$

Si

$$d(x_0, y_0) + d(y, t) \leq d(x_0, y) + d(y_0, t) = d(x_0, t) + d(y_0, y) \quad (14)$$

$\forall t \in S'$ tal que $t \notin \{x_0, y_0\}$, la proposición queda probada, siendo (x_0, y_0) la pareja buscada satisfaciendo las condiciones del lema.

Si por el contrario existe $z_0 \in S'$ tal que

$$d(x_0, y) + d(y_0, z_0) < d(x_0, y_0) + d(y, z_0) = d(x_0, z_0) + d(y, y_0) \quad (15)$$

vamos a probar primero que

$$d(x_0, y) + d(y_0, z) \leq d(x_0, y_0) + d(y, z) = d(x_0, z) + d(y, y_0) \quad z \notin \{x_0, y\} \quad (16)$$

y a partir del mismo el resultado más general

$$d(x_0, y) + d(z_1, z_2) \leq d(x_0, z_1) + d(y, z_2) = d(x_0, z_2) + d(y, z_1) \quad (17)$$

$$\forall (z_1, z_2) \in S \times S \quad \text{tal que} \quad \{z_1, z_2\} \cap \{x_0, y\} = \emptyset$$

con lo cual quedará probado el lema, siendo (x_0, y) la pareja que cumple las condiciones del mismo.

(Si existe $z_0 \in S'$ tal que

$$d(y_0, y) + d(x_0, z_0) < d(x_0, y_0) + d(z_0, y) = d(y_0, z_0) + d(y, x_0))$$

la demostración es paralela, resultando (y_0, y) la pareja en las condiciones del lema.)

Demostramos primero (16) por reducción al absurdo.

A partir de la cuaterna $\{x_0, y, y_0, z\}$ y suponiendo que no se verifica (16), por el hecho de ser d una distancia aditiva nos quedan dos posibilidades:

$$a) \quad d(x_0, y_0) + d(y, z) < d(x_0, y) + d(y_0, z) = d(x_0, z) + d(y_0, y) \quad (18)$$

$$b) \quad d(x_0, z) + d(y, y_0) < d(x_0, y) + d(z, y_0) = d(x_0, y_0) + d(z, y) \quad (19)$$

Analicemos la suposición a)

$$d(x_0, y_0) + d(y, z) < d(x_0, y) + d(y_0, z) = d(x_0, z) + d(y_0, y)$$

Obtenemos

$$d(y, y_0) = d(x_0, y) + d(y_0, z) - d(x_0, z)$$

y de (15)

$$d(y, y_0) = d(x_0, y_0) + d(y, z_0) - d(x_0, z_0)$$

por lo que

$$d(x_0, y) + d(y_0, z) - d(x_0, z) = d(x_0, y_0) + d(y, z_0) - d(x_0, z_0) \quad (20)$$

considerando ahora la cuaterna $\{x_0, y_0, z, z_0\}$, a partir de (13)

$$d(y_0, z) = d(x_0, z) + d(y_0, z_0) - d(x_0, z_0)$$

y sustituyendo en (20) resulta

$$d(x_0, y) + d(y_0, z_0) = d(x_0, y_0) + d(y, z_0)$$

por lo que si consideramos $\{x_0, y, y_0, z_0\}$, al ser d una distancia aditiva obtenemos

$$d(x_0, z_0) + d(y, y_0) \leq d(x_0, y) + d(y_0, z_0) = d(x_0, y_0) + d(y, z_0)$$

que está en contradicción con (15).

Por tanto el caso (18) descrito en a) no se puede dar.

Analicemos la suposición b),

$$d(x_0, z) + d(y, y_0) < d(x_0, y) + d(y_0, z) = d(x_0, y_0) + d(y, z) \quad (21)$$

Puesto que a partir de $\{x_0, y_0, z, z_0\}$ y de (13)

$$d(x_0, z) + d(y_0, z_0) = d(x_0, z_0) + d(z, y_0)$$

sustituyendo $d(z, y_0)$ en (21), se tiene

$$d(x_0, y) + d(x_0, z) + d(y_0, z_0) - d(x_0, z_0) = d(x_0, y_0) + d(y, z)$$

es decir

$$d(x_0, y) + d(y_0, z_0) = d(x_0, y_0) + d(y, z) + d(x_0, z_0) - d(x_0, z)$$

resultado de la igualdad anterior y de (15)

$$d(x_0, y_0) + d(y, z) + d(x_0, z_0) - d(x_0, z) < d(x_0, z_0) + d(y, y_0)$$

es decir

$$d(x_0, y_0) + d(y, z) < d(y, y_0) + d(x_0, z)$$

con lo que encontramos una contradicción con (21).

Queda pues probado (16) , es decir

$$d(x_0, y) + d(y_0, z) \leq d(x_0, y_0) + d(y, z) = d(x_0, z) + d(y, y_0) \quad \forall z \notin [x_0, y]$$

Probaremos ahora (17).

Puesto que por (13), para $(z_1, z_2) \in S' \times S'$

$$d(x_0, y_0) + d(z_1, z_2) \leq d(x_0, z_1) + d(y_0, z_2) = d(x_0, z_2) + d(y_0, z_1) \quad (22)$$

nos resulta

$$d(x_0, y) + d(z_1, z_2) \leq d(x_0, y) + d(x_0, z_1) + d(y_0, z_2) - d(x_0, y_0)$$

y como por (16)

$$d(x_0, y) + d(y_0, z_2) \leq d(x_0, y_0) + d(y, z_2)$$

obtenemos

$$d(x_0, y) + d(z_1, z_2) \leq d(x_0, z_1) + d(y, z_2) \quad (23)$$

Por otro lado, a partir de (22)

$$d(x_0, y) + d(z_1, z_2) \leq d(x_0, y) + d(x_0, z_2) + d(y_0, z_1) - d(x_0, y_0)$$

y como por (16)

$$d(x_0, y) + d(y_0, z_1) \leq d(x_0, y_0) + d(y, z_1)$$

obtenemos

$$d(x_0, y) + d(z_1, z_2) \leq d(y, z_1) + d(x_0, z_2) \quad (24)$$

Así de (23) y (24),

$$d(x_0, y) + d(z_1, z_2) \leq \min \{ d(x_0, z_1) + d(y, z_2), d(y, z_1) + d(x_0, z_2) \}$$

por lo que al ser d distancia aditiva, resulta

$$d(x_0, y) + d(z_1, z_2) \leq d(x_0, z_1) + d(y, z_2) = d(x_0, z_2) + d(y, z_1) \quad (25)$$

quedando probado (17).

De este modo la pareja (x_0, y) verifica las condiciones del lema puesto que para cualquier $(z, t) \in S \times S$ tal que $\{x_0, y\} \cap \{z, t\} = \emptyset$

$$d(x_0, y) + d(z, t) \leq d(x_0, z) + d(y, t) = d(x_0, t) + d(y, z)$$

Teorema 3.1

d es una distancia aditiva sobre S si y sólo si existe un árbol aditivo de tipo 1 asociado a una representación del par (S, d) .

Demostración:

Veamos en primer lugar que si existe un árbol aditivo de tipo 1 asociado a una representación del par (S, d) , entonces d es una distancia aditiva.

Sean $x, y, z, t \in S$, y supongamos que se verifica

$$d(x, y) + d(z, t) \leq d(x, z) + d(y, t) \leq d(x, t) + d(y, z) \quad (26)$$

Consideremos

$$U_1 = \{ u \in U \mid \phi(u) \text{ pertenece al camino entre } \phi(x) \text{ y } \phi(y) \}$$

$$U_2 = \{ u \in U \mid \phi(u) \text{ pertenece al camino entre } \phi(z) \text{ y } \phi(t) \}$$

siendo U el conjunto de nudos del árbol aditivo.

1) Veamos primero que x e y no pueden pertenecer ambos a U_2 .

Si suponemos $x, y \in U_2$, como U_2 es un subgrafo con dos vértices terminales $\phi(z)$ y $\phi(t)$ tenemos

$$d(z, t) = d^*(z, t) = d'(\phi(z), \phi(x)) + d'(\phi(x), \phi(y)) + d'(\phi(y), \phi(t))$$

o bien

$$d(z, t) = d^*(z, t) = d'(\phi(z), \phi(y)) + d'(\phi(y), \phi(x)) + d'(\phi(x), \phi(t))$$

resultando en el primer caso

$$d(z, t) > d(x, z) + d(y, t)$$

y en el segundo

$$d(z, t) > d(y, z) + d(x, t)$$

siendo ambos resultados contradictorios con (26).

2) Si suponemos que $x \notin U_2$ e $y \in U_2$,

entonces

$$d(z, t) = d^*(z, t) = d(z, y) + d(y, t)$$

y sumando $d(x,y)$ en ambos términos

$$d(x,y) + d(z,t) = d(x,y) + d(z,y) + d(y,t)$$

de donde

$$d(x,y) + d(z,t) \geq d(z,y) + d(x,t) \quad (27)$$

por lo que se alcanza la igualdad en (26),

$$d(x,y) + d(z,t) = d(x,z) + d(y,t) = d(x,t) + d(y,z)$$

verificándose el axioma del cuarto punto.

3) Sólo resta tratar el caso $x,y \notin U_2$.

Si suponemos que $U_1 \cap U_2 \neq \emptyset$ y existe más de un nudo perteneciente a $U_1 \cap U_2$ llegamos fácilmente a un absurdo puesto que se nos forma un ciclo en el grafo, contradiciendo de este modo la propiedad R1. Así, $U_1 \cap U_2$ sólo puede contener un único elemento, es decir

$$U_1 \cap U_2 = \{u\}$$

En este caso también debe ser notado que si existe algún camino que conecte dos nudos pertenecientes respectivamente a U_1, U_2 sin pasar por u , obtenemos también un ciclo. Por lo tanto no debemos considerar esta posibilidad y llegamos a que

$$d(x,t)+d(y,z)=d'(\phi(x),\phi(u))+d'(\phi(u),\phi(t))+d'(\phi(y),\phi(u))+d'(\phi(u),\phi(z))$$

y

$$d(x,z)+d(y,t)=d'(\phi(x),\phi(u))+d'(\phi(u),\phi(z))+d'(\phi(y),\phi(u))+d'(\phi(u),\phi(t))$$

verificándose

$$d(x,t) + d(y,z) = d(x,z) + d(y,t)$$

y con ello el axioma del cuarto punto.

Nos resta finalmente considerar $U_1 \cap U_2 = \emptyset$

Escogemos u_1, u_2 tales que $u_1 \in U_1, u_2 \in U_2$ de manera que no exista $u \in U_1 \cup U_2$ que pertenezca al camino entre u_1 y u_2 .

Así ,

$$d(x,t)+d(y,z) = d'(\phi(x),\phi(u_1))+d'(\phi(u_1),\phi(u_2))+d'(\phi(u_2),\phi(t))+$$

$$d'(\phi(y),\phi(u_2))+d'(\phi(u_2),\phi(u_1))+d'(\phi(u_1),\phi(z))$$

y

$$d(x,z)+d(y,t) = d'(\phi(x),\phi(u_1))+d'(\phi(u_1),\phi(u_2))+d'(\phi(u_2),\phi(z))+$$

$$d'(\phi(y),\phi(u_2))+d'(\phi(u_1),\phi(u_2))+d'(\phi(u_2),\phi(t))$$

verificándose también la igualdad

$$d(x,t) + d(y,z) = d(x,z) + d(y,t)$$

Queda pues demostrado que d es una distancia aditiva.

Demostremos el recíproco por inducción sobre el cardinal de S .

Caso 1) $n = 4$, siendo n el cardinal de S ($|S| = n$).

Si $x,y,z,t \in S$ y suponemos

$$d(x,y) + d(z,t) \leq d(x,t) + d(y,z) = d(x,z) + d(y,t) \quad (28)$$

veamos que existe un árbol aditivo de tipo 1 (Fig.1) asociado a (S,d) .

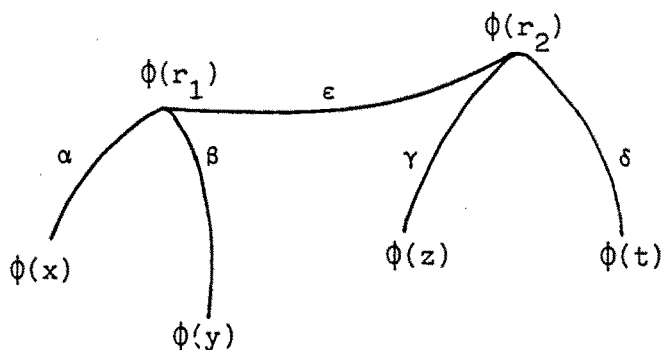


Figura 1

Para ello planteamos el sistema:

$$\left\{ \begin{array}{l} d(x,y) = \alpha + \beta \\ d(x,z) = \alpha + \epsilon + \gamma \\ d(x,t) = \alpha + \epsilon + \delta \\ d(y,z) = \beta + \epsilon + \gamma \\ d(y,t) = \beta + \epsilon + \delta \\ d(z,t) = \gamma + \delta \end{array} \right. \quad (29)$$

A partir de (28) y de las ecuaciones 3ª, 4ª, 5ª, obtenemos

$$d(x,z) = d(x,t) + d(y,z) - d(y,t) = \alpha + \gamma + \epsilon$$

por lo que la segunda ecuación depende linealmente de las demás. Es inmediato entonces comprobar que el sistema tiene rango máximo obteniendo como soluciones

$$\alpha = \frac{1}{2}(d(x,y) + d(x,t) - d(y,t))$$

$$\beta = \frac{1}{2}(d(x,y) + d(y,t) - d(x,t))$$

$$\gamma = \frac{1}{2}(d(y,z) + d(z,t) - d(y,t))$$

$$\delta = \frac{1}{2}(d(z,t) + d(y,t) - d(y,z))$$

$$\epsilon = \frac{1}{2}(d(y,z) + d(x,t) - d(x,y) - d(z,t))$$

que a partir de (28) son positivas. La construcción del árbol resulta ahora inmediata.

Caso 2) $n > 4$.

Veamos que si es cierto para $n-1$ también lo es para n .

Sea una pareja (x,y) que verifique

$$d(x,y) + d(z,t) \leq d(x,z) + d(y,t) = d(x,t) + d(y,z) \quad (30)$$

$\forall (z,t) \in S \times S$ tal que $\{x,y\} \cap \{z,t\} = \emptyset$, cuya existencia se ha demostrado en el lema.

Consideremos

$$S' = S - \{y\}$$

y la distancia $d_1 = d|_{S'}$ que evidentemente verifica la propiedad del cuarto punto.

A partir de la hipótesis de inducción podemos asegurar la existencia de un árbol aditivo de tipo 1 que representa al par (S', d_1) . Se trata ahora de efectuar una "ampliación" del árbol (fig.2), introduciendo ahora $\phi(y)$. Elegimos $z \in S'$ y vemos si existen

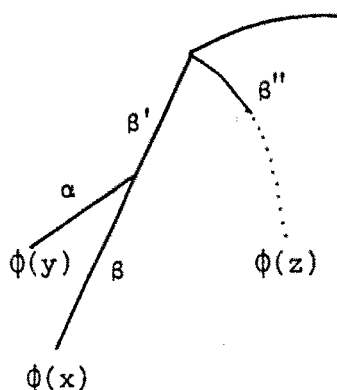


Figura 2

valores α, β, β' dispuestos como en la figura 2, verificando que la distancia asociada a este árbol es justamente d .

Para ello, planteamos el sistema

$$\begin{cases} d(y,z) = \alpha + \beta' + \beta'' \\ d(x,y) = \beta + \alpha \\ d(x,z) = \beta + \beta' + \beta'' \end{cases} \quad (31)$$

Las soluciones a este sistema son:

$$\begin{aligned} \beta' &= \frac{1}{2}(d(y,z) + d(x,z) - d(x,y) - 2\beta'') \\ \beta &= \frac{1}{2}(d(x,z) + d(x,y) - d(y,z)) \\ \alpha &= \frac{1}{2}(d(y,z) + d(x,y) - d(x,z)) \end{aligned} \quad (32)$$

comprobándose fácilmente a partir del lema que las soluciones son no negativas.

La distancia a través del árbol entre "y" y otro elemento $z' \in S'$ es:

$$\alpha + \beta' + d(x, z') - (\beta + \beta')$$

y de sustituir en (32) se tiene

$$\begin{aligned} \alpha + \beta' + d(x, z') - \beta - \beta' &= \frac{1}{2}(d(y, z) + d(x, y) - d(x, z)) + d(x, z') - \frac{1}{2}(d(x, z) + d(x, y) - d(y, z)) \\ &= d(x, z') + d(y, z) - d(x, z) \end{aligned}$$

y puesto que

$$d(x, y) + d(z, z') \leq d(x, z) + d(y, z') = d(x, z') + d(y, z)$$

por la elección del par (x, y)

$$d(x, z') + d(y, z) - d(x, z) = d(y, z')$$

comprobándose que la distancia a través del árbol construido coincide con la distancia original d .

De este modo queda demostrado el teorema, y con ello caracterizadas las representaciones asociadas a una distancia aditiva.

Definición 3.13

Dos árboles aditivos definidos sobre S , (G_{U_1}, d_1^*) y (G_{U_2}, d_2^*) con conjuntos de nudos principales P_1 y P_2 diremos que son equivalentes si la aplicación

$$\Psi: P_1 \longrightarrow P_2$$

que transforma el nudo principal definido por una terna s_1, s_2, s_3 de U_1 en el nudo también definido por s_1, s_2, s_3 en U_2 es biyectiva, y además

$$a) \quad d_1^*(u, v) = d_2^*(\Psi(u), \Psi(v)) \quad \forall (u, v) \in P_1 \times P_1$$

$$d_1^*(x, y) = d_2^*(x, y) \quad \forall (x, y) \in S \times S$$

- b) Si N_{uv} es el conjunto de nudos principales en el camino entre u y v , $\Psi(N_{uv})$ es el conjunto de nudos en el camino entre $\Psi(u)$ y $\Psi(v)$.

Vamos a demostrar la unicidad de la representación asociada a una distancia aditiva en el sentido de la definición anterior.

Teorema 3.2

Dos árboles aditivos asociados a (S, d) , siendo d distancia aditiva, son equivalentes.

Demostración:

Probemos el teorema por inducción sobre el cardinal del conjunto S ($|S| = n$).

Para el caso $n=4$, el resultado es inmediato.

Si suponemos cierto el teorema para $|S| = n-1$, veamos que también se verifica para $|S| = n$

Sea $(x,y) \in S \times S$ tal que

$$d(x,y) + d(z,t) \leq d(x,z) + d(y,t) = d(x,t) + d(y,z) \quad (33)$$

$\forall (z,t) \in S \times S$ con $\{x,y\} \cap \{z,t\} = \emptyset$, que existe en virtud del lema 3.1

Es inmediato comprobar que si x ó y son nudos terminales (no terminales) para un árbol también lo son para el otro, resultando al menos uno de ellos ser nudo terminal. Es por ello que analizaremos los siguientes casos:

a) x nudo terminal e y nudo no terminal.

b) x e y nudos terminales.

Caso a) Si P_1 y P_2 son los nudos principales asociados a los dos árboles con representaciones (G_{U_1}, d_1^*) , (G_{U_2}, d_2^*) entonces $y \in P_1$ e $y \in P_2$.

Si tomamos $S' = S - \{y\}$, podemos considerar (G_{U_1}, d_1^*) como una representación de S' considerando al elemento y como un elemento del conjunto R_1 que cumple

$$U_1 = S' \cup R_1$$

Análogamente para (G_{U_2}, d_2^*) .

Puede ocurrir que en la representación de S' , y sea o no sea nudo principal (Fig. 3)

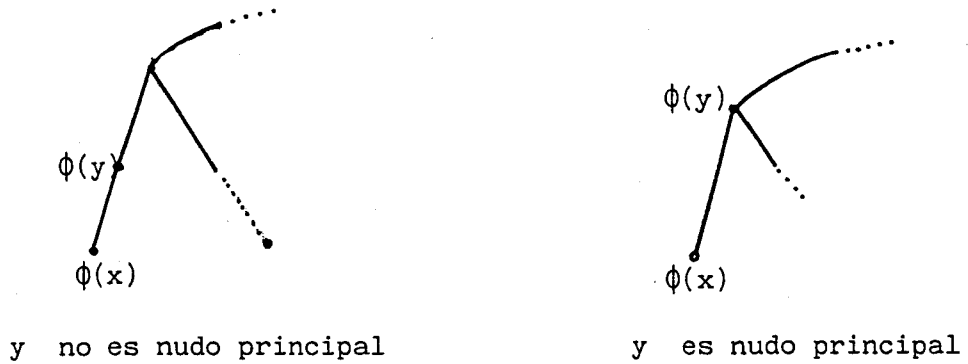


Figura 3

Si y es nudo principal, a partir de la hipótesis de inducción sobre S' queda probado directamente la equivalencia entre ambos árboles. En caso de no ser nudo principal se puede definir

$$\psi' : P_1 \longrightarrow P_2$$

de modo que

$$\psi'(u) = \psi(u) \quad \text{para} \quad u \in P_1^* = P_1 - \{y\}$$

siendo ψ la función, $\psi : P_1^* \rightarrow P_2^*$, que existe por la hipótesis de inducción sobre S' , y $\psi'(y) = y$, función que verifica las condiciones establecidas en la definición 3.7. Así, ambos árboles resultan equivalentes.

Caso b) Tomando

$$S' = S - \{y\}$$

$$G_1 = G_{U_1} - \{(y, u_1) ; (y, u_1) \in G_{U_1}\}$$

$$G_2 = G_{U_2} - \{(y, u_2) ; (y, u_2) \in G_{U_2}\}$$

y considerando $(G_1, d_1^*|_{U_1 - \{y\}})$ y $(G_2, d_2^*|_{U_2 - \{y\}})$, árboles aditivos asociados a $(S', d|_{S'})$ con conjuntos de nudos principales respectivamente P_1 y P_2 que son equivalentes por hipótesis de inducción,

existe

$$\psi : P_1 \longrightarrow P_2$$

cumpliendo las condiciones de la definición 3.13.

Podemos ahora construir

$$\psi' : P_1 \cup \{u_1\} \longrightarrow P_2 \cup \{u_2\} \quad \text{de modo que}$$

$$\psi'(u) = \psi(u) \quad \text{para } u \in P_1$$

$$\psi'(u_1) = u_2$$

deduciéndose a partir del teorema 3.1 y de las soluciones del sistema (31) que (G_{U_1}, d_1^*) y (G_{U_2}, d_2^*) son equivalentes.

Definición 3.14

Una representación (G_U, d^*) asociada a (S, d) es árbol aditivo de tipo 2 si se verifica:

a) Todas las condiciones de AD_1 .

b) $\forall (x, y) \in G_U$, si

$$\phi(x) = (\phi_1(x), \phi_2(x))$$

$$\phi(y) = (\phi_1(y), \phi_2(y))$$

y

$$\phi_2(x) \leq \phi_2(y)$$

podemos descomponer P_{xy} en la forma

$$P_{xy} = P_1 \cup P_2$$

siendo

$$P_1 = \{ (\phi_1(x), \alpha) \mid \phi_2(x) \leq \alpha \leq \phi_2(y) \}$$

$$P_2 = \{ (\alpha, \phi_2(y)) \mid \phi_1(x) \leq \alpha \leq \phi_1(y) \}$$

ó

$$\phi_1(y) \leq \alpha \leq \phi_1(x) \}$$

Teorema 3.3

d es una distancia aditiva definida sobre S si y sólo si existe un árbol aditivo de tipo 2 (AD_2) asociado a una representación del par (S,d) .

La demostración es idéntica a la del teorema 3.1.

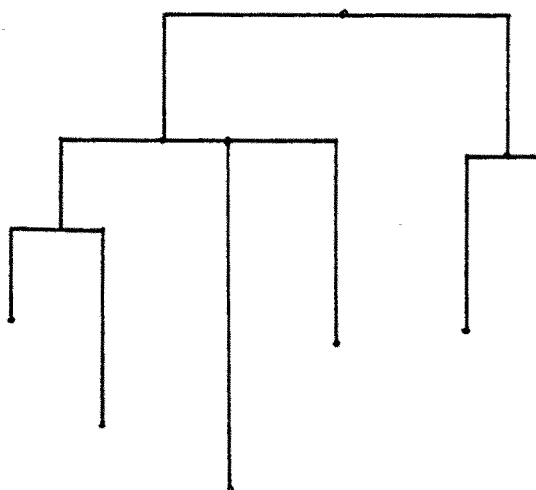


Figura 4.- Representación de un árbol aditivo de tipo 2.

Definición 3.15

Una representación (G_U, d^*) es un árbol aditivo de tipo 3 si:

- Se verifican las condiciones AD_1 , AD_2 , AD_3
- Se cumple la condición b) de la definición 3.14.
- $\forall (x,y) \in G_U$ se verifica

$$d^*(x,y) = | \phi_2(y) - \phi_2(x) | \quad (34)$$

induciendo la disimilaridad d' en $(x,y) \in G_U$ definida como valor absoluto de la diferencia entre las coordenadas segundas, de donde $d^*(x,y) = d'(\phi(x), \phi(y))$

Intuitivamente se trata de una representación como la anterior, solo que para el cálculo de la distancia entre dos nudos se tiene en cuenta únicamente la longitud de arista vertical.

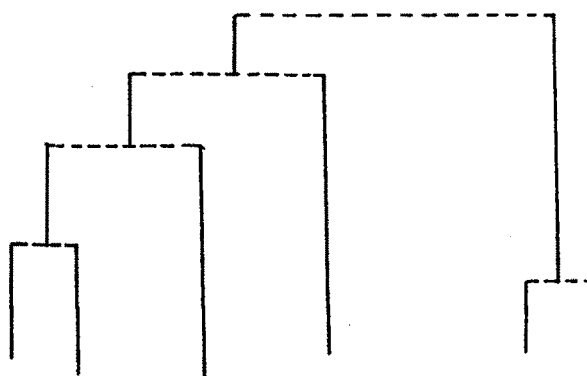


Figura 5. Representación de un árbol aditivo de tipo 3. Las longitudes en trazo continuo son las únicas tenidas en cuenta para el cálculo de la distancia.

Teorema 3.4

d es una distancia aditiva definida sobre S si y sólo si existe un árbol aditivo de tipo 3 asociado a una representación del par (S,d) .

La demostración es idéntica a la del teorema 3.1.

Definición 3.16

A un árbol aditivo de tipo 3 lo llamaremos dendograma aditivo.

3.4. RELACIONES CON LAS DISTANCIAS ULTRAMÉTRICAS

Estudiaremos en esta parte las relaciones entre las representaciones asociadas a una distancia aditiva que hemos visto en el apartado anterior y las representaciones asociadas a una distancia ultramétrica.

Proposición 3.2

Una distancia ultramétrica se puede representar como un árbol aditivo de tipo 3 en el cual $\phi_2(s) = \phi_2(s')$ para todo $s, s' \in S$.

Demostración:

A partir del algoritmo fundamental de clasificaciones sabemos que la representación asociada a una distancia ultramétrica es un dendograma (Cuadras, 1981).

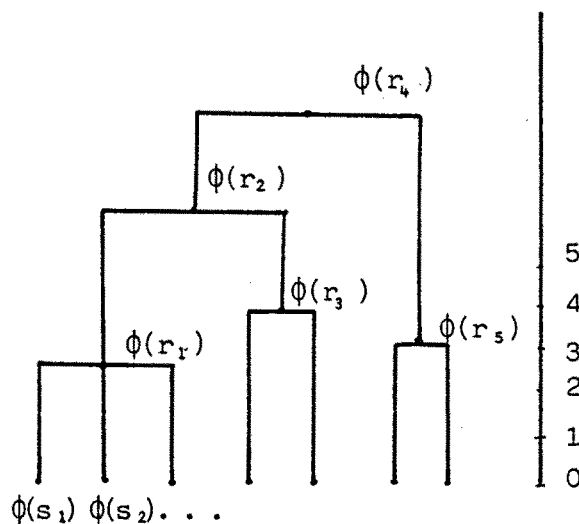


Figura 6

Si d_u es la distancia ultramétrica, construyamos el dendograma asociado a $d'_u = \frac{1}{2} \cdot d_u$, y consideremos como R al conjunto de nudos que no representan los elementos de S .

Sea $\phi(r) = (\beta_r^1, \beta_r^2)$ la representación en el plano de un nudo r , según la escala que fija el índice de la jerarquía. De este modo podemos considerar sin perder generalidad que para todo $s \in S$, $\phi(s) = (\alpha_s, 0)$. Resulta entonces evidente por (34) que

$$d^*(s_1, s_2) = |\phi_2(r_0) - \phi_2(s_1)| + |\phi_2(r_0) - \phi_2(s_2)| \quad (35)$$

siendo $r_0 \in R$ un nudo perteneciente al camino entre s_1 y s_2 verificando

$$\phi(r) \leq \phi(r_0)$$

para cualquier otro $r \in R$ del camino; a partir de (35) puesto que la segunda componente de $\phi(s_1)$ y $\phi(s_2)$ es cero,

$$d^*(s_1, s_2) = 2\phi_2(r_0) = 2 \cdot \frac{d_u(s_1, s_2)}{2} = d_u(s_1, s_2)$$

Las demás condiciones son verificadas por el árbol de modo evidente.

Proposición 3.3

Todo árbol aditivo de tipo 3 en que $\{\phi(s); s \in S\}$ tiene la segunda componente constante, considerando S como el conjunto de nudos terminales, está asociado a una distancia ultramétrica.

Demostración:

Sean $s_1, s_2, s_3 \in S$ y

$$d(s_1, s_2) \leq d(s_2, s_3) \leq d(s_1, s_3) \quad (36)$$

siendo d la distancia a la que está asociada el árbol. A partir de las hipótesis sobre el árbol, existen r_1, r_2, r_3 nudos que verifican

$$\phi(r_1) \leq \phi(r_2) \leq \phi(r_3) \quad (37)$$

y tienen por segundas componentes r_1'', r_2'', r_3'' cumpliendo

$$d^*(s_1, s_2) = 2 \cdot (r_1'' - \alpha)$$

$$d^*(s_2, s_3) = 2 \cdot (r_2'' - \alpha)$$

$$d^*(s_1, s_3) = 2 \cdot (r_3'' - \alpha)$$

siendo α la segunda componente de las imágenes por ϕ de los elementos de S .

Puesto que

$$d^*(s_1, s_3) \leq d^*(s_1, r_1) + d^*(r_1, r_2) + d^*(r_2, s_3)$$

entonces

$$2 \cdot (r_3'' - \alpha) \leq (r_1'' - \alpha) + (r_2'' - r_1'') + (r_2'' - \alpha)$$

es decir

$$r_3'' \leq r_2''$$

Puesto que en (37) teníamos $r_2'' \leq r_3''$ queda demostrado que

$$d(s_2, s_3) = d(s_1, s_3)$$

resultando que d es una distancia ultramétrica, tal y como queríamos demostrar.

Proposición 3.4

Supongamos un árbol aditivo de tipo 3 verificando la siguiente condición:

- a) $\max \{ \phi_2(s) \mid s \in S \} \leq \min \{ \phi_2(r) \mid r \in R \}$
- b) El conjunto de nudos terminales coinciden con los elementos de S .

En estas condiciones se verificará que la distancia asociada al mismo se puede descomponer como suma de una distancia ultramétrica d_u y una disimilitud singular d_s , es decir

$$d = d_u + d_s$$

Demostración:

Sea

$$\max \{ \phi_2(s) \mid s \in S \} = \phi_2(s') = \alpha_2$$

Construyamos

$$\omega : S \longrightarrow \mathbb{R}^+$$

definida como

$$\omega(x) = \alpha_2 - \phi_2(x)$$

Ahora podemos definir

$$\phi^* : U \longrightarrow \mathbb{R}^2$$

En la forma

$$\phi^*(s) = (\phi_1(s), \alpha_2) \quad \text{si } s \in S$$

$$\phi^*(r) = \phi(r) \quad \text{si } r \in R$$

A partir de los teoremas anteriores tenemos que utilizando ϕ^* en la representación, el árbol está asociado a una distancia ultramétrica que viene definida por

$$d_u(x,y) = d(x,y) - \omega(x) - \omega(y)$$

Si definimos

$$d_s(x,y) = \begin{cases} \omega(x) + \omega(y) & \text{si } x \neq y \\ 0 & \text{si } x = y \end{cases}$$

obtenemos una disimilaridad singular, por lo que

$$d = d_u + d_s$$

siendo d_u ultramétrica y d_s singular .

Así en determinadas condiciones hemos comprobado que una distancia aditiva la podemos descomponer en la suma de una distancia ultramétrica y una disimilaridad singular.

Con un razonamiento parecido al anterior se puede probar ,

Proposición 3.5

Si d es distancia aditiva , existe d_u distancia ultramétrica (no única) y f función de S en R de modo que

$$d(x_i, x_j) = d_u(x_i, x_j) + f(x_i) + f(x_j)$$

Estos resultados ilustran las relaciones existentes entre distancias ultramétricas y aditivas.

Es interesante comentar que los resultados anteriores pueden servir de punto de partida para plantear la representación espacial de distancias aditivas en la línea de los trabajos de Ohsumi y Nakamura (1981) y Cuadras (1985) referentes básicamente al mismo problema para distancias ultramétricas .

4.- ESTRUCTURA DE VARIEDAD EN EL CONJUNTO DE DISTANCIAS ADITIVAS.

Resumen: En el presente capítulo se dota de una estructura de variedad con frontera al conjunto de distancias aditivas, adecuada para el estudio formal de diferentes aspectos relacionados con este método de representación sobre todo dentro del campo de la inferencia sobre árboles y del tratamiento de algoritmos.

Sumario:

- 4.1.- Introducción.
- 4.2.- Definiciones previas.
- 4.3.- Estructura del conjunto de distancias aditivas.
- 4.4.- Determinación de la frontera de la variedad.
- 4.5.- Estructura del conjunto de distancias ultramétricas dentro de la variedad.
- 4.6.- Ajuste por mínimos cuadrados en una carta de la variedad.
- 4.7.- Representación espacial de distancias aditivas.

4.1.- INTRODUCCION

En este capítulo nos proponemos dotar al conjunto de distancias aditivas con configuración de una estructura matemática adecuada. Por su interés en el tratamiento de algoritmos y para el estudio de cuestiones relativas a inferencia sobre árboles, estructuraremos el conjunto anterior como una variedad diferenciable con frontera. En una segunda parte, estudiaremos aspectos relacionados con algunos problemas generales que surgen en la utilización de la representación por árboles aditivos.

Otros estudios relativos a la búsqueda de estructuras de conjuntos parecidos han sido desarrollados en otros trabajos. Gondran(1976) introduce una estructura algebraica sobre el conjunto de las clasificaciones jerárquicas a fin de poder interpretar resultados sobre las mismas de una manera análoga al análisis factorial. Rao(1945) dota al conjunto de poblaciones sobre el que hemos definido k variables aleatorias de una estructura de variedad diferenciable a partir de una clase paramétrica de funciones de densidad, para obtener una distancia entre poblaciones invariante frente a las transformaciones admisibles de parámetros.

Hemos dividido la primera parte del capítulo en tres apartados. En el primero dotamos propiamente al conjunto de distancias aditivas con configuración de la estructura citada. En el segundo caracterizamos la frontera de la variedad. En el último estudiamos la estructura del conjunto de distancias ultramétricas dentro de la variedad. Es de notar el interés que esta estructura podría tener en el estudio de propiedades teóricas relativas a los algoritmos de transformación en tratamientos pareci

dos a los que se dan en Jardine & Sibson (1971), Salicrú (1983), Arcas (1983) respecto a métodos de taxonomía numérica.

En la segunda parte estudiamos primeramente el ajuste por mínimos cuadrados en una carta de la variedad y después observamos algunas propiedades sobre la representación espacial, utilizando como técnica MDS el método de las coordenadas principales, asociada a una distancia aditiva.

4.2.- DEFINICIONES PREVIAS

Definición 4.1

Un espacio métrico M se dice que es una variedad con frontera si para todo $x \in M$ existe un U entorno de x tal que U es homeomorfo a \mathbb{R}^n ó H^n ($H^n = \{ (x_1, \dots, x_n) \in \mathbb{R}^n; x_i \geq 0 \ \forall i \in \{1, \dots, n\} \}$). Los elementos cuyo entorno puede ser únicamente homeomorfo a H^n los llamaremos de la frontera.

Definición 4.2

Un par (ϕ, U) es una carta local definida en un abierto U de M si $\phi: U \rightarrow \mathbb{R}^n$ es un homeomorfismo. Así en U se inducen las coordenadas de \mathbb{R}^n .

Definición 4.3

Una carta local (ϕ, U) es C^r respecto a (ψ, V) si las aplicaciones $\phi \circ \psi^{-1}$ y $\psi \circ \phi^{-1}$ son C^r en $\psi(U \cap V)$ y $\phi(U \cap V)$ respectivamente.

Gráficamente

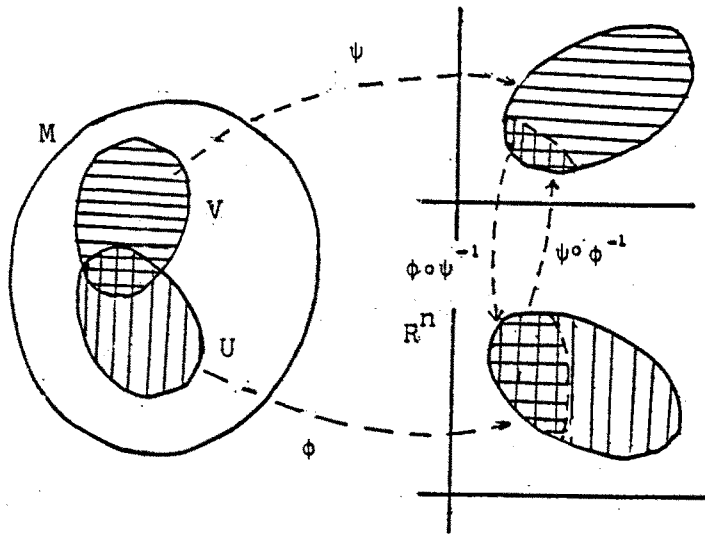


Figura 2

Un atlas C^r será una familia de homeomorfismos mutuamente C^r cuyos dominios cubran M .

Definición 4.4

A un par $(M ; T)$ donde T es un atlas maximal C^r para M se le llama variedad diferenciable.

Intuitivamente, una variedad diferenciable será un espacio métrico con coordenadas locales de forma que cuando en un punto se tengan dos coordenadas el paso de una a otra sea diferenciable. Lógicamente, una variedad diferenciable con frontera se definirá de forma análoga aceptando en las cartas locales homeomorfismos $\phi : U \rightarrow H^n$.

Definición 4.5

Un subconjunto $M_1 \subset M$ es una subvariedad de M si la inclusión

$$i: M_1 \longrightarrow M$$

es una inmersión.

4.3.- ESTRUCTURACION DEL CONJUNTO DE DISTANCIAS ADITIVAS

Sea A el conjunto de las distancias aditivas sobre un conjunto S de cardinal n , C el conjunto de las posibles configuraciones en árbol de S con $2n-3$ parámetros y $|C|=m$ (Consideramos las configuraciones "modelo" como árboles con los elementos de S como nudos terminales y nudos principales de grado 3. De este modo dependen de $2n-3$ parámetros, resultando las demás configuraciones como casos particulares de las primeras a través de la obtención de aristas de longitud cero. En Bergé(1973) se estudia el valor de m para diferentes tipos de árboles). Sea $A_c \subset \{(d,i) ; d \in A, i \in \{1, \dots, m\}\}$ representando el conjunto formado por los pares ordenados cuya primera componente es una distancia aditiva y cuya segunda componente representa la configuración.

Podemos definir en A_c la métrica

$$\Delta((d,i),(d',j)) = \sqrt{\sum_{\substack{k,l=1,\dots,n \\ k < l}} (d(x_k, x_l) - d'(x_k, x_l))^2} + \delta^{ij} \quad (1)$$

donde δ^{ij} son los deltas de Kronecker, y la parametrización

$$d \longrightarrow (\beta_1, \dots, \beta_{2n-3}) \quad (2)$$

siendo β_t la longitud de la arista t una vez ordenadas las aristas en cada configuración, es decir, el vector β es la solución del sistema de ecuaciones

$$d = X \cdot \beta \quad (3)$$

siendo X la matriz $\binom{n}{2} \times (2n-3)$ de la forma

$$x_{lkj} = \begin{cases} 1 & \text{si la arista } j \text{ conecta en el camino entre } l \text{ y } k \\ 0 & \text{en caso contrario} \end{cases}$$

con $l=1, \dots, n$; $k=1, \dots, n$; $l < k$; $j=1, \dots, 2n-3$ y d el vector de distancias, $d = (d_{12}, \dots, d_{n-1n})$

Sea $(d_0, i) \in A_c$ fijado con parametrización $\beta^0 = (\beta_1^0, \dots, \beta_{2n-3}^0)$ y sea X la matriz asociada a la configuración i . Tomemos la aplicación lineal $\psi: R^{\binom{n}{2}} \rightarrow R^{2n-3}$ con matriz asociada

$$Z = (X' \cdot X)^{-1} \cdot X'$$

siendo X' matriz traspuesta de X .

Es evidente que $\psi(d_0) = \beta^0$, y además al ser ψ lineal también es continua.

Si suponemos que $\exists j$ tal que $\beta_j^0 = 0$, consideramos $\epsilon = \frac{1}{2} \cdot \min \{ \beta_i^0 ; \beta_i^0 \neq 0 \}$ y tomamos

$$V = [\beta_{1-}^0 - \epsilon, \beta_{1+}^0 + \epsilon) \times \dots \times [\beta_{j-1-}^0 - \epsilon, \beta_{j-1+}^0 + \epsilon) \times [0, \epsilon) \times [\beta_{j+1-}^0 - \epsilon, \beta_{j+1+}^0 + \epsilon) \times \dots \times [\beta_{2n-3-}^0 - \epsilon, \beta_{2n-3+}^0 + \epsilon)$$

$$U_{d_0} = \psi^{-1}(V) \quad \text{y} \quad V_{d_0} = \{(d, i); d \in U_{d_0}, i \text{ fijo}\} \subset A_c$$

(d aditiva con configuración i fijada)

Se trata de comprobar que V_{d_0} es un entorno de (d_0, i) y que es homeomorfo a V .

Al ser ψ continua, tenemos que para el ϵ tomado $\exists \delta > 0$ tal que

$$\psi(B(d_0, \delta)) \subset B(\beta^0, \epsilon)$$

y tomando $\delta' = \frac{1}{2} \cdot \min\{1, \delta\}$ se tendrá :

$$B((d_0, i), \delta') \subset V_{d_0} \quad (4)$$

En efecto, si $(d, j) \in B((d_0, i), \delta')$, entonces

$$\Delta((d_0, i), (d, j)) < \delta'$$

y por ser $\delta' < 1$, obtenemos $i=j$; además, al ser $\delta' < \delta$, nos resulta $\|d - d_0\| < \delta$ por lo que $\psi(d) \in B(\beta^0, \varepsilon)$ con todas las componentes de la parametrización no negativas al ser d aditiva con configuración i . Así, $d \in U_{d_0}$ y $(d, j) \in V_{d_0}$ probando pues (4) .

Si definimos ahora el homeomorfismo

$$\begin{aligned} \phi : V_{d_0} &\longrightarrow V \\ (d, i) &\longrightarrow \psi(d) \end{aligned}$$

queda probado que (d_0, i) es un elemento de la frontera de A_c , pues V_{d_0} es homeomorfo a H^{2n-3} no existiendo ningún entorno de (d_0, i) homeomorfo a R^n para $n \in N$.

Análogamente, tomando

$$V = (\beta_{1-\varepsilon}^0, \beta_{1+\varepsilon}^0) \times \dots \times (\beta_{2n-3-\varepsilon}^0, \beta_{2n-3+\varepsilon}^0)$$

para $\beta^0 = (\beta_1^0, \dots, \beta_{2n-3}^0)$ con $\beta_i^0 > 0$ para todo i , se tendrá que existe V_{d_0} entorno de d_0 homeomorfo a R^{2n-3} .

Queda así probado que A_c es una variedad con frontera, formada esta última por las distancias aditivas con configuración en la - cual alguna arista es cero. Es inmediato comprobar que es una variedad C^∞ .

4.4.- DETERMINACION DE LA FRONTERA DE LA VARIEDAD

En este apartado buscamos las condiciones que debe verificar una distancia aditiva para que pertenezca a la frontera de la variedad.

Por notación, indicaremos por $\alpha_1, \dots, \alpha_n$ las longitudes de las aristas terminales (limitadas por algún nudo terminal) y por $\beta_1, \dots, \beta_{n-3}$ las longitudes de las aristas internas al haber fijado la configuración modelo asociada a la distancia en el sentido al que nos hemos referido en el apartado anterior. Consideramos por tanto de entrada que los nudos terminales coinciden con los elementos de S , aunque al existir la posibilidad de aristas de longitud cero volvemos a estar en el caso general. Notamos por R_{ij} al conjunto de aristas internas de longitud positiva en el camino entre los elementos i y j , y para simplificar, escribiremos d_{ij} en vez de $d(s_i, s_j)$ siendo $s_i, s_j \in S$. Así pues

$$d_{ij} = \alpha_i + \alpha_j + \sum_{R_{ij}} \beta_k \quad (5)$$

Supongamos que existe $j \in \{1, \dots, n\}$ tal que $\alpha_j = 0$. En este caso resulta que existen $i, k \in \{1, \dots, n\} - \{j\}$ tales que

$$d_{ij} + d_{jk} = d_{ik} \quad (6)$$

En efecto, es suficiente con tomar i, k tales que los conjuntos de aristas R_{ij} y R_{jk} tengan intersección vacía. Recíprocamente si $d_{ij} + d_{jk} = d_{ik}$, los conjuntos R_{ij} y R_{jk} deberán tener intersección vacía, y de ahí deducimos que $\alpha_j = 0$

En efecto, a partir de (5) y (6)

$$\alpha_i + \alpha_j + \sum_{R_{ij}} \beta_r + \alpha_j + \alpha_k + \sum_{R_{jk}} \beta_r = \alpha_i + \alpha_k + \sum_{R_{ik}} \beta_r \quad (7)$$

y puesto que $R_{ik} \subset R_{ij} \cup R_{jk}$

es inmediato a partir de (7) que

$$\sum_{R_{ik}} \beta_r = \sum_{R_{ij}} \beta_r + \sum_{R_{jk}} \beta_r$$

y $\alpha_j = 0$

Queda así establecido el siguiente resultado:

Proposición 4.1

El conjunto de distancias aditivas para las cuales existen elementos i, j, k tales que

$$d_{ij} + d_{jk} = d_{ik}$$

pertencen a la frontera de la variedad.

En el mismo sentido anterior demostramos

Proposición 4.2

Si existen cuatro puntos para los cuales

$$d_{ij} + d_{kl} = d_{ik} + d_{jl} = d_{il} + d_{kj} \quad (8)$$

entonces d pertenece a la frontera de la variedad

Demostración:

A través del teorema 3.1 (cap.3) demostramos que dados cuatro puntos i, j, k, l tales que

$$d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}$$

la representación an árbol es única y en la forma que observamos en la figura 2

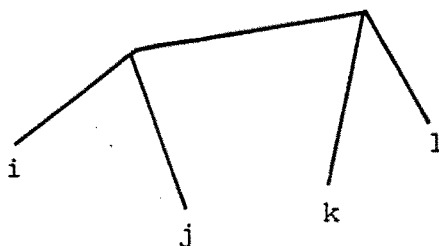


Figura 2

Si consideramos en el árbol aditivo asociado a d el subgrafo determinado por los elementos i, j, k, l en las condiciones (8), obtenemos que todas las aristas pertenecientes al camino entre el nudo principal determinado por i, j, k y el determinado por k, l, j son iguales a cero. Es decir,

$$(R_{ij} \cap R_{kl}) \cup (R_{ik} \cap R_{jl}) \cup (R_{il} \cap R_{jk})$$

se reduce al conjunto vacío.

Por lo tanto queda probado que el par formado por d y la configuración asociada pertenece a la frontera. También deducimos que la distancia d restringida a estos cuatro puntos es singular.

Recíprocamente, si la longitud de alguna arista interna es cero siendo las aristas terminales distintas de cero, encontramos un nudo principal de grado mayor o igual que 4, es decir, con al menos cuatro aristas adyacentes al mismo. Basta entonces considerar cuatro elementos de S que sean nudos terminales de los caminos a los que pertenecen estas cuatro aristas, verificándose para los mismos la propiedad (8).

Así pues, podemos enunciar

Proposición 4.3

La frontera de la variedad A_c viene dada por

$$\partial A_c = \{(d, i) \in A_c \mid d \text{ verifica las propiedades (6) o (8)}\}$$

Así un ejemplo de distancia cumpliendo la hipótesis de la proposición 4.2 sería

$$d = \begin{pmatrix} 0 & 2 & 4 & 3.5 & 5.6 & 5.8 & 6 \\ & 0 & 4 & 3.5 & 5.6 & 5.8 & 6 \\ & & 0 & 1.5 & 3.6 & 3.8 & 4 \\ & & & 0 & 3.1 & 3.3 & 3.5 \\ & & & & 0 & 1.4 & 3.6 \\ & & & & & 0 & 3.8 \\ & & & & & & 0 \end{pmatrix}$$

la cual tiene una representación asociada (Figura 3)

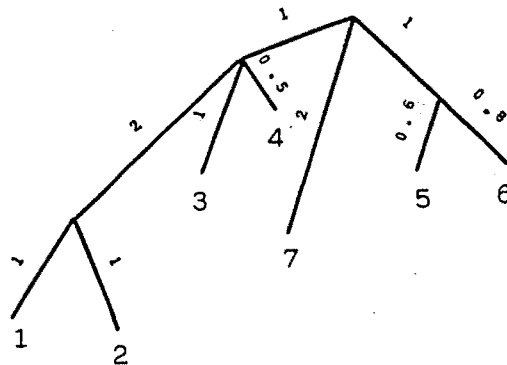


Figura 3

resultando las aristas terminales con longitudes (1,1,1,0.5,2,0.6,0.8) y las aristas internas (2,0,1,1), obteniendo pues un elemento de la frontera.

Un ejemplo de distancia cumpliendo las hipótesis de la proposición 4.1 sería

$$d = \begin{pmatrix} 0 & 1 & 3.5 & 4.6 & 6 \\ & 0 & 2.5 & 3.6 & 5 \\ & & 0 & 2.1 & 3.5 \\ & & & 0 & 2.6 \\ & & & & 0 \end{pmatrix}$$

con representación asociada (Fig. 4)

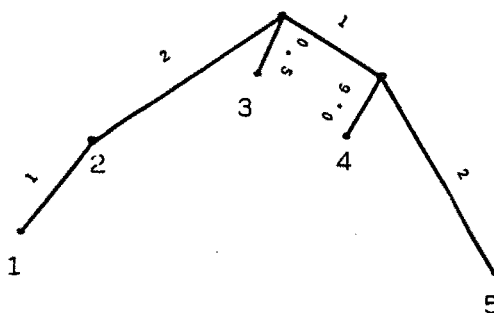


Figura 4

resultando la parametrización $(\underbrace{1, 0, 0.5, 0.6}_{\text{arist. term.}}, \underbrace{2, 2, 1}_{\text{arist. int.}})$

En la siguiente proposición trabajamos directamente sobre el árbol obteniendo conclusiones sobre el mismo, pero partiendo de propiedades de la distancia.

Proposición 4.4

Si $d \in A$ y suponemos la existencia de 6 elementos $\{1, 2, 3, 4, 5, 6\}$ para los que

$$d_{12} + d_{ij} \leq d_{1i} + d_{2j} = d_{1j} + d_{2i} \quad i, j \in \{3, 4, 5, 6\}$$

$$d_{34} + d_{ij} \leq d_{3i} + d_{4j} = d_{3j} + d_{4i} \quad i, j \in \{1, 2, 5, 6\}$$

$$d_{56} + d_{ij} \leq d_{5i} + d_{6j} = d_{5j} + d_{6i} \quad i, j \in \{1, 2, 3, 4\}$$

y

$$d_{34} = \frac{1}{2} (d_{13} + d_{24} + d_{35} + d_{46} - d_{15} - d_{26}) \quad (9)$$

se verifica que el conjunto de aristas que enlazan los elementos 3 y 4 con el nudo principal determinado por 1, 3 y 5 no pertenecientes al camino entre 3 y 4 son cero.

Demostración:

Es inmediato comprobar que si

$$d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk} \quad (10)$$

se verifica que

$$R_{ij} \cap R_{kl} = \emptyset$$

y además

$$R_{ik} - (R_{ij} \cup R_{kl}) = R_{jl} - (R_{ij} \cup R_{kl}) = R_{il} - (R_{ij} \cup R_{kl}) = R_{jk} - (R_{ij} \cup R_{kl}) \quad (11)$$

por la propiedad (5), a partir de (9)

$$\begin{aligned} \alpha_3 + \alpha_4 + \sum_{R_{34}} \beta_r &= \frac{1}{2} \left(\sum_{R_{13}} \beta_r + \alpha_1 + \alpha_3 + \sum_{R_{24}} \beta_r + \alpha_2 + \alpha_4 + \sum_{R_{35}} \beta_r + \alpha_3 + \alpha_5 \right. \\ &\quad \left. + \sum_{R_{46}} \beta_r + \alpha_4 + \alpha_6 - \sum_{R_{15}} \beta_r - \alpha_1 - \alpha_5 - \sum_{R_{26}} \beta_r - \alpha_2 - \alpha_6 \right) \end{aligned}$$

de donde

$$\sum_{R_{34}} \beta_r = \frac{1}{2} \left(\sum_{R_{13}} \beta_r + \sum_{R_{24}} \beta_r + \sum_{R_{35}} \beta_r + \sum_{R_{46}} \beta_r - \sum_{R_{15}} \beta_r - \sum_{R_{26}} \beta_r \right) \quad (12)$$

y a partir de (10) y (11) utilizando las hipótesis de la proposición nos resulta el segundo miembro de la igualdad (12) como

$$\begin{aligned}
& \frac{1}{2} (2(\sum_{R_{13}-(R_{12} \cup R_{34})} \beta_r) + 2(\sum_{R_{35}-(R_{34} \cup R_{56})} \beta_r) - 2(\sum_{R_{15}-(R_{12} \cup R_{56})} \beta_r) \\
& + \sum_{R_{13} \cap (R_{12} \cup R_{34})} \beta_r + \sum_{R_{24} \cap (R_{12} \cup R_{34})} \beta_r + \sum_{R_{35} \cap (R_{34} \cup R_{56})} \beta_r + \sum_{R_{46} \cap (R_{34} \cup R_{56})} \beta_r \\
& - \sum_{R_{15} \cap (R_{12} \cup R_{56})} \beta_r - \sum_{R_{26} \cap (R_{12} \cup R_{56})} \beta_r)
\end{aligned}$$

que operando nos queda como

$$\sum_{R_{13}-(R_{12} \cup R_{34})} \beta_r + \sum_{R_{35}-(R_{34} \cup R_{56})} \beta_r - \sum_{R_{15}-(R_{12} \cup R_{56})} \beta_r + \sum_{R_{34}} \beta_r \quad (13)$$

y puesto que $R_{15}-(R_{12} \cup R_{56})$ podemos descomponerlo en la forma

$$(R_{15} \cap (R_{12} \cup R_{56})^c \cap R_{13}) \cup (R_{15} \cap (R_{12} \cup R_{56})^c \cap (R_{35} - R_{13})) =$$

$$(R_{15} \cap R_{12}^c \cap R_{56}^c \cap R_{13}) \cup (R_{15} \cap R_{12}^c \cap R_{56}^c \cap (R_{35} - R_{13}))$$

y al ser

$$R_{15} \cap R_{12}^c \cap R_{56}^c \cap R_{13} \subset R_{13} \cap (R_{12} \cup R_{34})^c$$

$$R_{15} \cap R_{12}^c \cap R_{56}^c \cap (R_{35} - R_{13}) \subset R_{35} \cap R_{34}^c \cap R_{56}^c$$

La expresión (12) nos queda como

$$\sum_{R_{34}} \beta_r = \sum_{R_1} \beta_r + \sum_{R_2} \beta_r + \sum_{R_{34}} \beta_r \quad (14)$$

siendo

$$R_1 = R_{13} \cap R_{12}^c \cap R_{34}^c \cap (R_{15}^c \cup R_{12} \cup R_{56} \cup R_{13}^c)$$

$$R_2 = R_{35} \cap R_{34}^c \cap R_{56}^c \cap (R_{15}^c \cup R_{12} \cup R_{56} \cup (R_{35} - R_{13})^c)$$

resultando a partir de las hipótesis del teorema

$$R_1 = R_{13} \cap R_{12}^c \cap R_{34}^c \cap R_{15}^c$$

$$R_2 = R_{35} \cap R_{34}^c \cap R_{56}^c \cap R_{15}^c$$

de donde utilizando la igualdad (14)

$$\sum_{R_{13} \cap R_{34}^c \cap (R_{12}^c \cup R_{56}^c) \cap R_{15}^c} \beta_r = 0$$

es decir, que las aristas que conectan el elemento 3 y 4 con el nudo principal determinado por 1, 3 y 5 que no pertenecen al camino entre 3 y 4 son todas de longitud cero, tal y como queríamos demostrar.

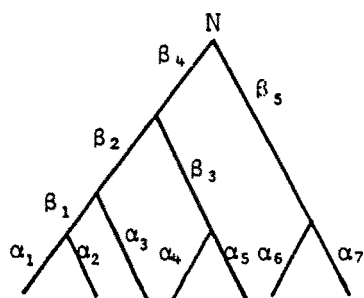
Estudiaremos en el siguiente apartado la relación entre distancias ultramétricas y aditivas desde la perspectiva de la estructura estudiada. Así demostraremos que el conjunto de distancias ultramétricas con configuración resulta ser una subvariedad de A_c con frontera.

4.5.- ESTRUCTURA DEL CONJUNTO DE DISTANCIAS ULTRAMÉTRICAS DENTRO DE LA VARIEDAD

Puesto que una distancia ultramétrica puede representarse como un árbol tal que cualquier nudo interno es equidistante a todos los nudos terminales que tienen al anterior en el camino que los une, podríamos pensar en una parametrización como (Fig. 3)

$$u \longleftrightarrow (\beta_1, \dots, \beta_{n-2}, \alpha_1) \quad (15)$$

es decir, tomando como parámetros las longitudes de las aristas internas y la longitud de la arista externa mínima para una configuración i asociada a la distancia ultramétrica u .



$$\alpha_2 = \alpha_1$$

$$\alpha_3 = \beta_1 + \alpha_1$$

$$\alpha_4 = \alpha_1 + \beta_1 + \beta_2 - \beta_3$$

$$\alpha_5 = \alpha_1 + \beta_1 + \beta_2 - \beta_3$$

$$\alpha_6 = \alpha_7 = \alpha_1 + \beta_1 + \beta_2 + \beta_4 - \beta_5$$

Figura 5: La distancia de un nudo interno a los terminales es constante.

Obtenemos pues

$$\alpha_i = \alpha_1 + \sum_{R_{iN}} \beta_r - \sum_{R_{iN}} \beta_r \quad (16)$$

siendo R_{iN} el conjunto de aristas que pertenecen al camino que une el nudo terminal i y el nudo N equidistante a todos los nodos terminales.

De modo idéntico al desarrollo efectuado para las distancias aditivas obtenemos una estructura de variedad diferenciable con frontera, asociada a las distancias ultramétricas con configuración. Demostraremos que se trata de una subvariedad de A_c de dimensión $n-1$ siendo $|S| = n$.

Si (u, i) es un par formado por una distancia ultramétrica u con su configuración i con parametrización

$$(\beta_1, \dots, \beta_{n-2}, \alpha_1)$$

a través de relaciones (16) y de

$$\begin{cases} \beta'_{n-3} = \beta_{n-3} + \beta_{n-2} \\ \beta'_i = \beta_i \quad \text{para } i \in \{1, \dots, n-4\} \end{cases} \quad (17)$$

obtenemos la parametrización de u como distancia aditiva en el sentido (2)

$$(\beta'_1, \dots, \beta'_{n-3}, \alpha_1, \dots, \alpha_n)$$

eligiendo la configuración derivada de modo natural de la obtenida a través de la configuración ultramétrica (Fig. 6').

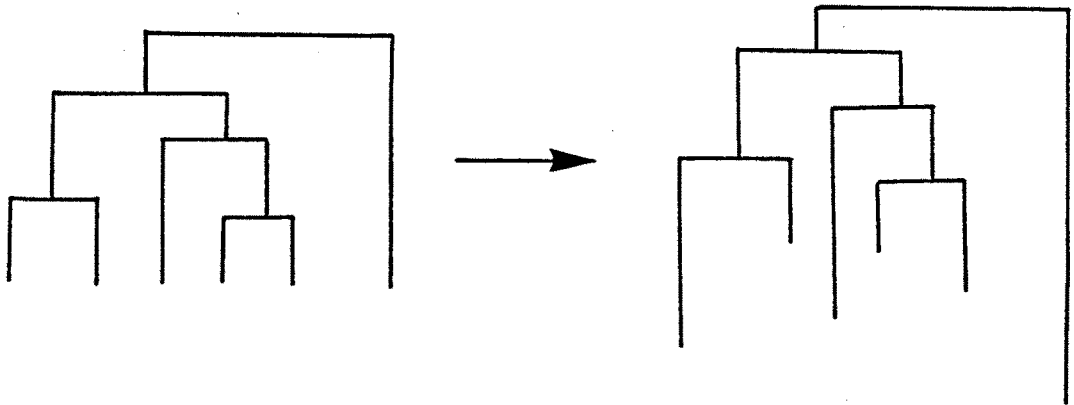


Figura 6: Configuración ultramétrica.

Conf. aditiva derivada.

A través de las relaciones (16) y (17) es inmediato definir una inmersión del conjunto de las ultramétricas con configuración (U_c) en el conjunto de las aditivas con configuración, de donde

Proposición 4.5

U_c es una subvariedad de A_c .

Vamos a caracterizar la frontera de la subvariedad. Observemos que si existe $i \in \{1, \dots, n\}$ tal que $\alpha_i = 0$ entonces existe $j \in \{1, \dots, n\} - \{i\}$ tal que $d_{ij} = 0$, con lo que d no sería una distancia, por lo que no puede darse este caso. Si tenemos una arista interna con longitud igual a cero, es inmediata la existencia de al menos tres elementos i, j, k tales que

$$d_{ij} = d_{ik} = d_{jk}$$

El recíproco es una consecuencia del algoritmo fundamental de clasificación (Cuadras, 1981). De este modo podemos enunciar

Proposición 4.6

La frontera de U está formada por los pares (d, c) siendo c la configuración asociada a d de modo que existan i, j, k tales que $d_{ij} = d_{ik} = d_{jk}$.

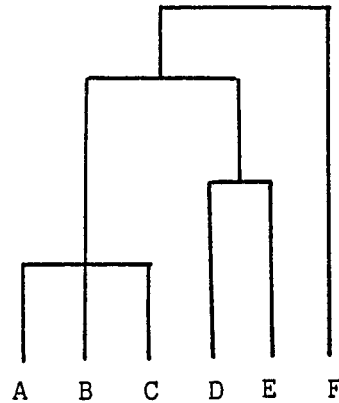


Figura 7: Representación de una ultramétrica perteneciente a la frontera.

4.6.- AJUSTE POR MINIMOS CUADRADOS EN UNA CARTA DE LA VARIEDAD

Sea X matriz $\binom{n}{2} \times (2n-3)$ asociada a una determinada configuración de árbol aditivo dentro de una carta de la variedad. Existen varios algoritmos (Cap. 6) que una vez fijada la configuración del árbol precisan de la estimación de las longitudes de las aristas utilizando el método de los mínimos cuadrados.

Si llamamos δ a la disimilaridad observada, el problema se traduce en calcular $\hat{\beta}$ de modo que

$$X'X\hat{\beta} = X'\delta$$

(Cuadras, 1981)

Veamos como queda la matriz de productos escalares $X' \cdot X$ (supongamos que las primeras n columnas de X corresponden a aristas terminales a_1, \dots, a_n y las restantes a aristas internas b_1, \dots, b_{n-3}). Es inmediato comprobar que el elemento ij de la matriz $X'X$ es el número de distancias entre dos vértices terminales de modo que las aristas i y j pertenecen al camino determinado por dichos vértices. Así, si consideramos un árbol de manera que b_i separe k y $n-k$ elementos y b_j separe s y $n-s$ elementos (Fig. 8) se verifica:

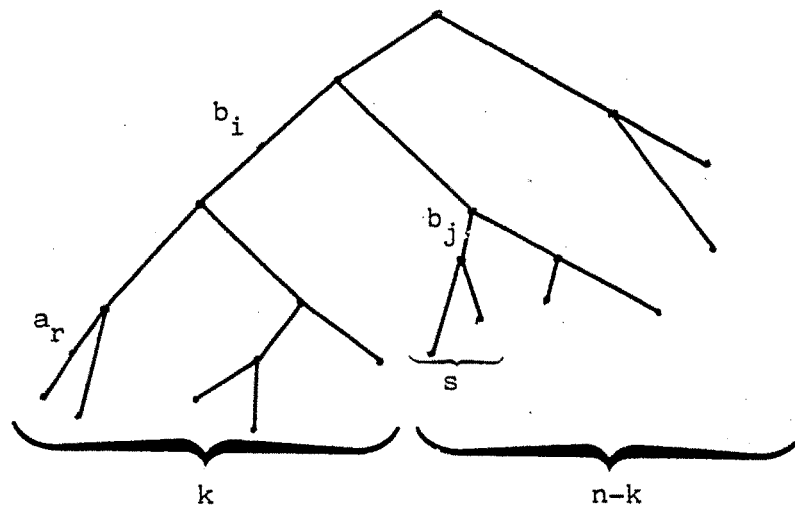


Figura 8

Proposición 4.7

$$a_i \cdot a_j = \begin{cases} 1 & \text{si } i \neq j \\ n-1 & \text{si } i = j \end{cases}$$

$$a_r \cdot b_i = n - k \tag{18}$$

$$b_i \cdot b_j = \begin{cases} k \cdot s & \text{si } i \neq j \\ k(n-k) & \text{si } i = j \end{cases}$$

A partir de este resultado es inmediato comprobar que

Proposición 4.8

La suma de distancias entre dos puntos de modo que los caminos que los conectan pasan por una determinada arista es igual a la suma de distancias observadas entre dichos puntos.

Considerando el subespacio formado por las columnas de X , estamos hallando la proyección del vector β en tal subespacio. En este sentido es interesante comprobar que

Proposición 4.9

El ángulo formado por a_r y a_s es mayor que el ángulo formado por a_r y b_i .

Demostración:

A partir de la proposición 4.7

$$\cos(a_r, a_s) = \frac{1}{n-1} \quad \text{y} \quad \cos(a_r, b_i) = \frac{n-k}{\sqrt{n-1} \cdot \sqrt{k(n-k)}} = \frac{\sqrt{n-k}}{\sqrt{k(n-1)}}$$

y podemos comprobar como

$$\frac{\sqrt{n-k}}{\sqrt{k(n-1)}} > \frac{1}{n-1} \quad (19)$$

En efecto: probar (19) equivale a probar

$$\sqrt{n-1} \cdot \sqrt{n-k} > \sqrt{k}$$

o equivalentemente

$$0 < n(n - (k + 1))$$

y esto es cierto puesto que k es menor o igual que $n-2$, con lo que queda demostrado el resultado enunciado.

Queda abierto en este punto el problema de la búsqueda de condiciones para la existencia de soluciones positivas en una determinada carta de la variedad.

4.7.- REPRESENTACION ESPACIAL DE DISTANCIAS ADITIVAS

Veremos en este apartado algunas propiedades de la representación espacial asociada a una distancia aditiva.

Tal y como hemos mencionado en 2.3.2 la representación MDS del par (S, δ) utilizando el método de las coordenadas principales se realiza a través de los vectores propios de la matriz

$$B = H.A.H$$

donde

$$h_{ij} = \begin{cases} 1 - \frac{1}{n} & \text{si } i \neq j \\ -\frac{1}{n} & \text{si } i = j \end{cases}$$

y

$$a_{ij} = -\frac{1}{2} \cdot d_{ij}^2$$

Dada d_{ij} distancia aditiva, definimos en S la relación de equivalencia

$$j \sim k \Leftrightarrow d_{jt} = d_{kt} \quad \forall t \in S$$

Esta relación de equivalencia induce la partición en S

$$S = S_1 \cup \dots \cup S_r$$

en la que cada S_i es una clase maximal respecto la inclusión formada por elementos equidistantes, es decir

$$d_{jh} = d_{jk} = d_{hk} \quad \text{para } j, k, h \in S_i$$

Puesto que las interdistancias entre los elementos de una misma clase son todos iguales, al valor de esta distancia para $|S_i| \geq 2$ la llamamos α_i .

Podemos entonces demostrar

Proposición 4.10

$\frac{1}{2} \cdot \alpha_i^2$ es un valor propio de la matriz B asociado a la clase S_i .

Demostración:

Es inmediato comprobar que $\frac{1}{2} \cdot a_i^2$ es valor propio asociado a la matriz A , obteniéndose como subespacio de vectores propios asociado

$$\begin{aligned} v_1 &= (0, \dots, 0, 1, -1, 0, \dots, 0) \\ &\vdots \\ v_{n_i} &= (0, \dots, 0, 1, 1, \dots, -(n_i-1), 0, \dots, 0) \end{aligned}$$

siendo $|S_i| = n_i$, con la ordenación de los elementos de S en la matriz A compatible con la partición de S obtenida.

Observando entonces que $H(v_j) = v_j \quad \forall j \in \{1, \dots, n_i\}$ se deduce inmediatamente el resultado propuesto.

Cuadras (1985) estudia algunas propiedades de los valores propios de la matriz B en el caso en que d sea ultramétrica. Obtenemos que el menor valor propio distinto de 0 es $\frac{a^2}{2}$, siendo $a = \min \{ d_{ij} \mid i, j \in S, i \neq j \}$, con lo cual se deduce además que una distancia ultramétrica es también una distancia euclídea.

Para las distancias aditivas no se verifica necesariamente la condición de euclidicidad. Así por ejemplo dada la distancia aditiva

$$d = \begin{pmatrix} 0 & 1 & 1.01 & 0.52 \\ & 0 & 1.01 & 0.52 \\ & & 0 & 0.51 \\ & & & 0 \end{pmatrix}$$

con árbol asociado (Fig. 9).

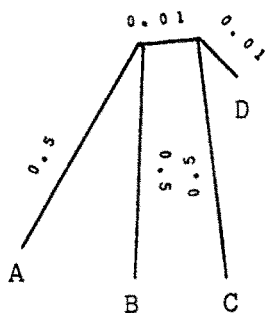


Figura 9

utilizando el resultado 2.1 obtenemos que la distancia d no resulta ser euclídea puesto que la matriz B tiene por valores propios

$$\lambda_1 = 0.513, \lambda_2 = 0.5, \lambda_3 = 0 \quad \text{y} \quad \lambda_4 = -0.053.$$

Dada $u = (u_{ij})$ distancia ultramétrica consideramos $|\xi_j|$ valores propios de B asociados a las clases S_i con $|S_i| > 2$ y $|\lambda_j|$ valores propios restantes. Definamos la distancia aditiva

$$d_{jk} = u_{jk} \quad \text{si} \quad k, j \neq i$$

$$d_{ik} = \bar{u}_{ik} + \epsilon$$

siendo ϵ una constante positiva (Fig. 8).

Veamos la variación de los valores propios de d con respecto los valores propios de u .

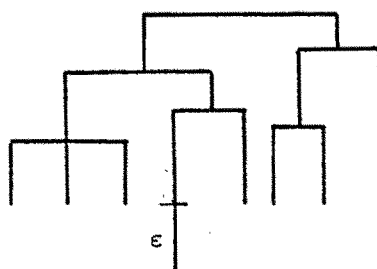


Figura 10

Si $i \in S_r$ con $|S_r| = 1$ resulta que todos los valores ξ_j son también valores propios para d . En caso contrario existe un valor propio ξ_r que no lo es para d . Notaremos por $\{\lambda'_j\}$ al conjunto de valores propios distintos de los ξ_j .

Puesto que

$$\sum_{j,k \neq i} u_{jk}^2 = n \left(\sum \xi_j + \sum \lambda_j \right) \quad (20)$$

(Cuadras, 1981)

y por otro lado

$$\sum_{j,k \neq i} u_{jk}^2 + \sum_t (u_{it} + \epsilon)^2 = n \left(\sum_{k \neq r} \xi_k + \sum \lambda'_j \right)$$

es decir

$$\sum_{j,k \neq i} u_{jk}^2 + \sum_t u_{it}^2 + 2\epsilon \sum_t u_{it} + (n-1)\epsilon^2 = n \left(\sum_{k \neq r} \xi_k + \sum \lambda'_j \right) \quad (21)$$

y restando las expresiones (21) y (20)

$$n \left(\sum \lambda'_j - \sum \lambda_j - \xi_r \right) = 2\epsilon \sum_t u_{it} + (n-1)\epsilon^2$$

es decir

$$\sum \lambda'_j - \sum \lambda_j - \epsilon_r = \frac{1}{n} (2 \cdot \epsilon \cdot \sum_t u_{it} + (n-1) \epsilon^2) \quad (22)$$

expresión que nos indica la variación de los valores propios para d y u . Obsérvese que si ϵ aumenta, crece también la diferencia entre los valores propios por lo que aumenta la variabilidad explicada por los valores λ_j .

Se puede también comprobar que $n-1$ valores propios de B asociados a d son mayores o iguales que cero.

5.- INFERENCIA EN ARBOLES ADITIVOS

Resumen: En este capítulo formulamos un modelo probabilístico para el vector de distancias a partir del cual analizamos algunos aspectos de inferencia en árboles aditivos.

Sumario:

- 5.1.- Introducción.
- 5.2.- Modelo probabilístico para el vector de distancias.
- 5.3.- Contrastes en árboles aditivos.

5.1.- INTRODUCCION

Distancias entre distribuciones de probabilidad han sido utilizadas en una gran variedad de trabajos sobre problemas de inferencia estadística y en aplicaciones prácticas para estudiar cuantitativamente las analogías y las diferencias entre un determinado conjunto de poblaciones en el análisis de datos biológicos, en economía, en sociología y muchos otros campos. Ver por ejemplo, Matusita (1957), Prevosti et al. (1975), Rao (1948, 1973, 1982). En todos estos casos la distancia puede ser interpretada como una medida de la información acerca de las analogías y diferencias entre las distribuciones comparadas.

Con frecuencia, en particular en el estudio de la evolución de poblaciones de seres vivos, resulta conveniente utilizar un modelo de representación basado en árboles aditivos, que presupone el haber definido previamente una distancia entre los objetos comparados, ya que este se ajusta bien a los conceptos biológicos de "Línea filogenética" y "divergencia evolutiva".

Estos estudios permiten esclarecer el fenómeno de la evolución ayudando a reconstruir los procesos históricos que dan origen a las poblaciones biológicas actuales, y en base a estos estudios efectuar además una clasificación de las mismas.

Por ejemplo, en Kimura (1984) puede verse, entre otros resultados, la construcción de un árbol filogenético de algunos vertebrados representativos basado en el cálculo de la distancia entre especies a partir de la composición de aminoácidos en las cadenas α de la hemoglobina, concordando los resultados con los previstos en la teoría neutralista de la Evolución.

Las distancias que observamos entre los objetos que comparamos, usualmente poblaciones, se calculan generalmente a partir de muestras aleatorias, por lo que aquellas debemos considerarlas como estimaciones de las "distancias reales" existentes entre los objetos comparados. Por tanto, cualquier modelo de clasificación basado en árboles aditivos obtenido por algún procedimiento de ajuste a partir de las estimaciones de las distancias, debe de ser considerado como una estimación del árbol aditivo que expresará las diferencias y analogías reales existentes entre los diferentes objetos estudiados.

Por todas estas consideraciones resulta conveniente desarrollar el tema de estimación y contrastes de hipótesis de árboles aditivos para poder cuantificar las probabilidades de error al aceptar o rechazar hipótesis sobre la configuración de los mismos. Por ejemplo, en el trabajo citado anteriormente de Kimura (1984), podría tener interés establecer una región confidencial sobre las longitudes de las aristas de algún árbol aditivo o controlar si los datos encajan bien con una determinada configuración que a su vez sería interpretable en términos evolutivos.

Para poder desarrollar el tema debemos introducir previamente un modelo probabilístico sobre la construcción de la distancia lo cual realizamos en el siguiente apartado.

5.2.- MODELO PROBABILISTICO PARA EL VECTOR DE DISTANCIAS

Empezaremos razonando brevemente el modelo probabilístico al que ajustamos el vector de distancias. Así, supondremos dadas k poblaciones estadísticas sobre las que hemos observado X_1, \dots, X_n variables aleatorias que admiten función de densidad conjunta cuando nos restringimos a cada una de las poblaciones, siendo las mismas caracterizadas por dicha densidad conjunta que la consideraremos perteneciente a una determinada clase de funciones de densidad paramétricas, $f(x_1, \dots, x_n, \theta_1, \dots, \theta_r)$. De este modo una población Π_i vendrá dada por una r -pla $(\theta_1^i, \dots, \theta_r^i)$ coordenadas de una determinada variedad paramétrica definida por

$$V = \{(\theta_1, \dots, \theta_r) \in R^r \mid f(x_1, \dots, x_n, \theta_1, \dots, \theta_r) \text{ es función de densidad}\}$$

Así, podemos interpretar la distancia real o teórica entre dos poblaciones como una función de los parámetros

$$d_{ij} = F(\theta^i, \theta^j) = F(\theta_1^i, \dots, \theta_r^i, \theta_1^j, \dots, \theta_r^j) \quad (1)$$

obteniendo el vector de distancias teórico,

$$d = (d_{11}, \dots, d_{k-1k})$$

aunque por comodidad de notación indicaremos

$$d = (d_1, \dots, d_m) \quad \text{siendo} \quad m = \binom{k}{2}$$

Una manera de obtener la distancia en la forma (1) podría ser a través de la definición de un campo tensorial covariante de segundo orden, simétrico y definido positivo, tomándolo como tensor métrico de la variedad y utilizando la métrica Riemanniana definida. Se podría utilizar por ejemplo la matriz de información de Fisher (Rao, 1948) en donde tomamos como tensor métrico

$$g_{\mu\nu} = E\left(\frac{1}{f^2} \cdot \frac{\partial f}{\partial \theta^\mu} \cdot \frac{\partial f}{\partial \theta^\nu}\right)$$

$$\mu, \nu = 1, \dots, r$$

La distancia observada la calcularemos utilizando las estimaciones máximo verosímiles de los parámetros θ^i a través de las muestras obtenidas de las variables en las poblaciones. Llamaremos $\hat{\theta}^i$ a las estimaciones máximo verosímiles de los parámetros θ^i . Para muestras suficientemente grandes, si la densidad cumple ciertas condiciones de regularidad y suponiendo independencia entre las poblaciones se verifica que $(\hat{\theta}_1^1, \dots, \hat{\theta}_r^1, \dots, \hat{\theta}_1^k, \dots, \hat{\theta}_r^k)$ converge asintóticamente a una distribución $N(\theta, \Sigma)$ (Mood-Graybill, 1969) siendo Σ una matriz de la forma

$$\Sigma = \begin{pmatrix} \Sigma_1 & & & \\ & \dots & & \\ & & 0 & \\ & & & \dots \\ & & 0 & & \Sigma_k \end{pmatrix} \quad (2)$$

Si suponemos que F admite desarrollo de Taylor de 2º orden y M_{ij} es la cota superior de la 2ª derivada en un entorno de $e^{ij} = (\theta_1^i, \dots, \theta_r^i, \theta_1^j, \dots, \theta_r^j)$ que contenga los segmentos $[\hat{\theta}^{ij}, \theta^{ij}]$ ó $[\theta^{ij}, \hat{\theta}^{ij}]$ nos resulta

$$| F(\hat{\theta}^i, \hat{\theta}^j) - F(\theta^i, \theta^j) - \sum_{\substack{s=i,j \\ k=1, \dots, r}} \frac{\partial F}{\partial x_k^s} \bigg|_{\theta^{ij}} (\hat{\theta}_k^s - \theta_k^s) | \leq M_{ij} \cdot \|\hat{\theta}^{ij} - \theta^{ij}\| \quad (3)$$

Si consideramos la matriz A de dimensiones $\binom{k}{2} \times r \cdot k$ siendo k el número de poblaciones, en donde en la fila correspondiente a la pareja (i, j) consideramos los elementos

$$\frac{\partial F}{\partial x_1^i} \bigg|_{\theta^{ij}}, \dots, \frac{\partial F}{\partial x_r^i} \bigg|_{\theta^{ij}}, \dots, \frac{\partial F}{\partial x_1^j} \bigg|_{\theta^{ij}}, \dots, \frac{\partial F}{\partial x_r^j} \bigg|_{\theta^{ij}}$$

en las columnas correspondientes a θ^i y θ^j , y cero en el resto, obtenemos que $A(\hat{\theta} - \theta)$ es asintóticamente normal por ser transformación lineal de $(\hat{\theta} - \theta)$ que también lo es.

Si $D = F(\hat{\theta}^i, \hat{\theta}^j)$ es la distancia observada, a partir de (3) es inmediato observar la existencia de una constante α tal que

$$\| \| D - d - A(\hat{\theta} - \theta) \| \| \leq \alpha \cdot \|\hat{\theta} - \theta\|$$

al converger $\hat{\theta} - \theta$ en probabilidad a cero, por la consistencia de la estimación máximo verosímil

$$D - d - A(\hat{\theta} - \theta) \xrightarrow{P} 0$$

y puesto que

$$A(\hat{\theta} - \theta) \xrightarrow{L} N(0, \Sigma_0)$$

por las propiedades de la convergencia de medidas de probabilidad en espacios de Banach (Billingsley, 1968)

$$D - d \xrightarrow{L} N(0, \Sigma_0)$$

es decir

$$D \xrightarrow{L} N(d, \Sigma_0)$$

y por lo tanto parece natural considerar como modelo probabilístico el de la normal multivariante de media la distancia teórica. Queda por tanto razonada la adecuación del modelo probabilístico (4) para el vector de distancias observado. Queda implícito a partir de la exposición anterior las rígidas condiciones a que sometemos el vector de distancias, pero que tienen bastante sentido si pensamos en una línea de distanciación de poblaciones basada en la parametrización de las mismas (Rao, 1945) , (Oller y Cuadras, 1985 a) .

2.- CONTRASTES EN ARBOLES ADITIVOS

Estudiamos en este apartado algunos contrastes que puedan servir de punto de partida sobre cuestiones relativas a inferencia sobre árboles. Consideramos que el vector de distancias observado sigue una normal multivariante de media la distancia teórica y matriz de covarianzas conocida Σ ,

$$D \sim N(d, \Sigma) \quad (4)$$

En la práctica podríamos considerar una estimación de Σ en el momento de formular el modelo a partir del desarrollo del apartado anterior.

Supongamos que queremos averiguar si nuestro vector de distancias se ajusta a una determinada distancia aditiva d_0 . Para resolver este contraste hallamos una región confidencial a un determinado nivel $1 - \epsilon$. Planteamos el contraste

$$H_0: d = d_0$$

$$H_1: d \neq d_0$$

Nos basamos en que bajo la hipótesis nula el estadístico

$$(D - d_0)' \cdot \Sigma^{-1} \cdot (D - d_0)$$

sigue una distribución ji-cuadrado con m grados de libertad (Andersson, 1958).

De este modo la región crítica viene dada por

$$W = \{ (D_1, \dots, D_m) \mid (D - d_0)' \cdot \Sigma^{-1} \cdot (D - d_0) \geq \chi_m^2(\epsilon) \} \quad (5)$$

También podríamos plantear este contraste considerando las distancias aditivas como puntos de una variedad paramétrica, tal y como observamos en 4.2. De este modo, si d_0 es aditiva no perteneciente a la frontera, existe alguna matriz X que caracteriza la configuración de orden $m \times (2k-3)$ ($k = n^{\circ}$ de poblaciones), de modo que:

$$d_0 = Xb_0$$

siendo

$$b_0 = (\beta_0^1, \dots, \beta_0^{2k-3})$$

parámetros que caracterizan a d_0 en la variedad.

Consideremos la transformación lineal

$$(X'X)^{-1}X'D = \beta \quad (6)$$

Bajo la hipótesis nula se verifica que

$$\beta \sim N [(X'X)^{-1}X'd_0, (X'X)^{-1}X'\Sigma X(X'X)^{-1}]$$

en donde $(X'X)^{-1}X'd_0 = b_0$

Indicando

$$(X'X)^{-1}X'\Sigma X(X'X)^{-1} = \Gamma$$

se tiene que

$$(\beta - b_0)' \Gamma^{-1} (\beta - b_0) \sim \chi^2_{2k-3} \quad (7)$$

y obtenemos una región confidencial que tendrá un gran interés si queda totalmente comprendida dentro de la carta que contiene a b_0 . Para tamaños muestrales suficientemente grandes utilizando una matriz Σ adecuada siempre podremos obtener una región confidencial en el sentido anteriormente citado.

Es interesante estudiar la función de potencia del contraste. Observemos que bajo la hipótesis alternativa

$$\beta \sim N[b, \Gamma]$$

$(b \neq b_0)$

de donde

$$\beta - b_0 \sim N [b - b_0, \Gamma]$$

Si diagonalizamos Γ ,

$$\Gamma = T \cdot \Lambda \cdot T' = T \cdot \Lambda^{1/2} \cdot (T \cdot \Lambda^{1/2})' = AA'$$

y efectuamos el cambio

$$Y = A^{-1} \cdot (\beta - b_0)$$

obtenemos que

$$Y \sim N(A^{-1}(b - b_0), I)$$

por lo que $Y'Y$ sigue una distribución ji-cuadrado no centrada con parámetro

$$\tau = \sqrt{(b - b_0)' \cdot (A^{-1})' \cdot A^{-1} \cdot (b - b_0)}$$

(Andersson, 1958)

Puesto que

$$Y'Y = (\beta - b_0)' \cdot (A^{-1})' \cdot A^{-1} \cdot (\beta - b_0) = (\beta - b_0)' \cdot \Gamma^{-1} \cdot (\beta - b_0)$$

(8)

$$P((\beta - b_0)' \cdot \Gamma^{-1} \cdot (\beta - b_0) \geq \chi_{2k-3}^2(\epsilon) | H_1) = P(Y'Y \geq \chi_{2k-3}^2(\epsilon))$$

resultando inmediato el cálculo de la potencia a través de la distribución de $Y'Y$. Recordemos que la función de densidad de $Y'Y$ es

$$f(v) = \frac{1}{\sqrt{2}^{2k-3}} \cdot e^{-\frac{1}{2}(\tau^2 + v)} \cdot v^{\frac{1}{2}(2k-3)-1} \cdot \sum_{\gamma=0}^{\infty} \frac{(\tau^2)^{\gamma} \cdot v^{\gamma}}{(\gamma)! \cdot \Gamma(\frac{1}{2}(2k-3) + \gamma)} \cdot 2^{2\gamma}$$

Podemos encontrarla tabulada entre otros en *Biometrika Tables for Statisticians* (1976)

De este modo podemos resumir los resultados obtenidos indicando que

- a) (5) determina una región crítica para resolver el contraste.
- b) A través de (7) obtenemos una región confidencial considerando la carta local de la variedad paramétrica que contiene d_0 .
- c) Mediante (8) calculamos la función de potencia del contraste.

Dado un vector de distancias, es interesante pensar en la adecuación del mismo a un modelo de representación como árbol ultramétrico versus un árbol aditivo. En este sentido vamos a estudiar un contraste que aborde este problema, aunque nos centraremos en considerar el contraste en una determinada carta de la variedad. Así, dado

$$D \sim N [d, \Sigma]$$

Siendo X la matriz que determina la configuración del árbol y E que sigue una distribución normal multivariante

$$E \sim N[0, \Sigma]$$

si

$$\Sigma = T \cdot A \cdot T' = (T \cdot A^{1/2})(T \cdot A^{1/2})' = AA'$$

efectuamos el cambio

$$Y = A^{-1} \cdot D$$

y nos resulta

$$Y = A^{-1} \cdot D \sim N[A^{-1} \cdot X \cdot \delta, I]$$

obteniendo el modelo

$$Y = A^{-1} \cdot X \cdot \delta + E_1$$

siendo

$$E_1 \sim N[0, I]. \text{ Indicando, } A^{-1} \cdot X = X_0$$

la estimación por mínimos cuadrados de δ resulta ser

$$\hat{\delta} = (X_0' X_0)^{-1} \cdot X_0' \cdot Y$$

y

$$R_0^2 = (Y - X_0 \hat{\delta})' \cdot (Y - X_0 \hat{\delta})$$

Bajo la hipótesis nula podemos considerar la matriz de diseño

$$\bar{X} = X_0 \cdot C$$

y en este caso

$$\hat{\gamma} = (\bar{X}'\bar{X})^{-1}\bar{X}'\cdot Y$$

y

$$R_1^2 = (Y - \bar{X}\cdot\hat{\gamma})'(Y - \bar{X}\cdot\hat{\gamma})$$

tenemos que $R_1^2 - R_0^2$ sigue una distribución ji-cuadrado con

$$2k-3-(k-1) = k-2 \quad \text{grados de libertad} \quad (10)$$

Otro contraste que tiene sumo interés es la comparación de dos modelos probabilísticos del tipo que estudiamos en que las distancias teóricas son distancias aditivas.

En este sentido supongamos

$$D_A \sim N[d_A, \Sigma]$$

$$D_B \sim N[d_B, \Sigma]$$

con Σ conocida, y d_A, d_B pertenecientes al interior de la misma carta local de la variedad.

Si X es la matriz asociada a la carta

$$d_A = X \cdot \beta_A$$

$$d_B = X \cdot \beta_B$$

Planteamos el contraste

$$\begin{aligned} H_0 &: d_A = d_B \\ H_1 &: d_A \neq d_B \end{aligned} \tag{11}$$

y consideramos previamente el cambio

$$Y_A = A^{-1}D_A$$

$$Y_B = A^{-1}D_B$$

siendo A matriz tal que $\Sigma = AA'$ obtenemos

$$Y_A \sim N[y_A, I]$$

$$Y_B \sim N[y_B, I]$$

siendo

$$y_A = A^{-1} \cdot d_A$$

$$y_B = A^{-1} \cdot d_B$$

quedando el contraste formulado como

$$H_0: y_A = y_B$$

$$H_1: y_A \neq y_B$$

Vamos a resolverlo utilizando técnicas de análisis de la varianza. Bajo la hipótesis alternativa podemos pensar en el diseño

$$Y_A = Z \cdot \beta_A + E$$

$$Y_B = Z \cdot \beta_B + E$$

siendo

$$Z = A^{-1} \cdot X \quad y \quad E \sim N [0, I]$$

De este modo la matriz de diseño nos resulta

$$M = \begin{pmatrix} Z & 0 \\ 0 & Z \end{pmatrix}$$

en donde las estimaciones de los parámetros son

$$\hat{\beta}_A = (Z'Z)^{-1} Z' \cdot Y_A$$

$$\hat{\beta}_B = (Z'Z)^{-1} Z' \cdot Y_B$$

y de este modo

$$R_0^2 = (Y_A - Z(Z'Z)^{-1} Z' Y_A)' (Y_A - Z(Z'Z)^{-1} Z' Y_A) +$$

$$(Y_B - Z(Z'Z)^{-1} Z' Y_B)' (Y_B - Z(Z'Z)^{-1} Z' Y_B) =$$

$$Y_A' Y_A - Y_A' Z (Z'Z)^{-1} Z' Y_A - Y_A' Z (Z'Z)^{-1} Z' Y_A + Y_A' \cdot Z (Z'Z)^{-1} \cdot Z' Z (Z'Z)^{-1} \cdot Z' Y_A +$$

$$+ Y_B' \cdot Y_B - Y_B' \cdot Z (Z'Z)^{-1} \cdot Z' Y_B - Y_B' Z (Z'Z)^{-1} \cdot Z' Y_B + Y_B' \cdot Z (Z'Z)^{-1} \cdot Z' Z (Z'Z)^{-1} \cdot Z' Y_B =$$

$$Y_A' Y_A + Y_B' Y_B - Y_A' \cdot Z (Z'Z)^{-1} \cdot Z' Y_A - Y_B' Z (Z'Z)^{-1} \cdot Z' Y_B$$

Bajo la hipótesis nula podemos tomar como matriz de diseño

$$\bar{M} = \begin{pmatrix} Z \\ Z \end{pmatrix}$$

resultándonos la estimación del parámetro

$$\hat{\beta} = (2Z'Z)^{-1} \cdot Z'(Y_A + Y_B) = \frac{1}{2}(Z'Z)^{-1} \cdot Z'(Y_A + Y_B)$$

y

$$\begin{aligned} R_1^2 &= (Y_A - \frac{1}{2}Z(Z'Z)^{-1} \cdot Z'(Y_A + Y_B))' \cdot (Y_A - \frac{1}{2}Z(Z'Z)^{-1} \cdot Z'(Y_A + Y_B)) + \\ &\quad (Y_B - \frac{1}{2}Z(Z'Z)^{-1} \cdot Z'(Y_A + Y_B))' \cdot (Y_B - \frac{1}{2}Z(Z'Z)^{-1} \cdot Z'(Y_A + Y_B)) = \\ &\quad Y_A' Y_A - \frac{1}{2}(Y_A + Y_B)' Z(Z'Z)^{-1} \cdot Z' Y_A - \frac{1}{2}(Y_A' Z(Z'Z)^{-1} \cdot Z'(Y_A + Y_B)) + \\ &\quad \frac{1}{2}(Y_A + Y_B)' Z(Z'Z)^{-1} \cdot Z' Z(Z'Z)^{-1} \cdot Z'(Y_A + Y_B) + (Y_B' Y_B) \\ &\quad - \frac{1}{2}(Y_A + Y_B)' \cdot Z(Z'Z)^{-1} \cdot Z' Y_B - \frac{1}{2} Y_B' Z(Z'Z)^{-1} \cdot Z'(Y_A + Y_B) = \\ &\quad Y_A' Y_A + Y_B' Y_B - \frac{1}{2}(Y_A + Y_B)' Z(Z'Z)^{-1} \cdot Z'(Y_A + Y_B) \end{aligned}$$

de donde

$$R_1^2 - R_0^2 = \frac{1}{2}(Y_A - Y_B)' Z(Z'Z)^{-1} \cdot Z'(Y_A - Y_B)$$

y obtenemos inmediatamente que $R_1^2 - R_0^2$ sigue una distribución χ^2 con $2k-3$ grados de libertad (Searle, 1971).

Queda abierta en este punto la posibilidad de realización de contrastes más generales. En este sentido sería interesante tener en cuenta la técnica propuesta por Oller(1983) y que se basa en el cálculo de la distancia estimada al cuadrado entre dos poblaciones normales , la cual multiplicada por una constante conveniente sigue una distribución ji-cuadrado. Para el cálculo de la distancia se deberá hallar primero el tensor métrico de la variedad a la que pertenecen las poblaciones, buscando después la distancia geodésica en la subvariedad en que realizamos el contraste. En Ríos (1984) se puede observar un ejemplo de la utilización de esta técnica.

6. ALGORITMOS DE CONSTRUCCION DE ARBOLES ADITIVOS

Resumen: En este capítulo se discute la utilización de diversos algoritmos de transformación de una disimilaridad dada sobre un conjunto en una distancia aditiva. Se concluye con un ejemplo práctico de la utilización de los árboles aditivos estudiando algunos aspectos de la localización europea de *Drosophila Subobscura*.

Sumario:

- 6.1 - Introducción.
- 6.2 - Principales algoritmos.
- 6.3 - Ejemplo.

6.1 - INTRODUCCION

Una disimilaridad observada sobre un conjunto de objetos no verifica, en general, la propiedad del cuarto punto (aditiva). Por lo tanto, si deseamos representar los objetos en un árbol aditivo debemos realizar una transformación de la misma en una distancia aditiva. De este modo consideraremos un algoritmo de construcción de árboles aditivos como un método que nos proporciona una distancia aditiva asociada a una disimilaridad sobre el conjunto, pudiéndose expresar en la forma

$$\begin{array}{l} \psi : D_S \longrightarrow D \\ \delta \longrightarrow d \end{array} \quad (1)$$

donde D_S es el conjunto de disimilaridades sobre S , D conjunto de distancias aditivas sobre S y $\psi(\delta) = d$ la distancia aditiva asociada a δ por el algoritmo en cuestión.

En este capítulo daremos una visión general de los principales algoritmos implementados hasta el momento comentando su eficiencia y los distintos puntos de vista en que están basados. También estudiaremos las líneas básicas para la confección de un algoritmo basado en la transformación de la disimilaridad en una distancia ultramétrica conveniente obteniendo a partir de la misma una distancia aditiva.

Pensamos que este capítulo puede ser de interés sobre todo para la utilización en problemas prácticos cuando hemos elegido como modelo de representación los árboles aditivos, puesto que la metodología a seguir en tales problemas se puede dividir en dos partes:

- 1ª) Cálculo de una disimilaridad adecuada entre los objetos (Sobre el particular cabe citar estudios como los de Rao (1945), Gower(1971) y Oller (1983)).
- 2ª) Una vez elegido el modelo de representación, se procede a la selección de un algoritmo de transformación adecuado.

6.2.- PRINCIPALES ALGORITMOS

6.2.1.- Algoritmo de Tversky (1977).

Este algoritmo fue elaborado por Tversky (1977) e implementado por J. Corter (1981) en un programa denominado ADDTREE, nombre con el que también es corriente referirse al algoritmo.

El algoritmo consta de dos partes:

- a) Construcción de la configuración del árbol.
- b) Estimación de la longitud de las aristas.

Describiremos a continuación estas dos partes:

- a) Si δ es la disimilaridad observada entre los objetos, considera para cada par $(x,y) \in S \times S$ el valor n_{xy} que representa el número de pares $(z,t) \in S \times S$ tales que

$$\delta(x,y) + \delta(z,t) \leq \min \{ \delta(z,x) + \delta(y,t), \delta(x,t) + \delta(y,z) \} \quad (2)$$

De este modo se toma el par (x_0, y_0) tal que

$$n_{x_0 y_0} = \max \{ n_{xy} \mid (x, y) \in S \times S \}$$

y consideramos unidos en un primer paso este par de objetos de S (Fig 1).



Figura 1

Se considera posteriormente una nueva disimilaridad en el conjunto

$$S' = (S - \{x_0, y_0\}) \cup \{u_0\}$$

definida en la forma

$$\begin{aligned} \delta'(x, y) &= \delta(x, y) & \text{si } x, y \neq u_0 \\ \delta'(x, u_0) &= \frac{1}{2}(\delta(x, x_0) + \delta(x, y_0)) \end{aligned} \quad (3)$$

y se aplica el mismo proceso anterior al par (S', δ') .

De este modo reiterando el proceso se llega a obtener la configuración del árbol (Fig.2).

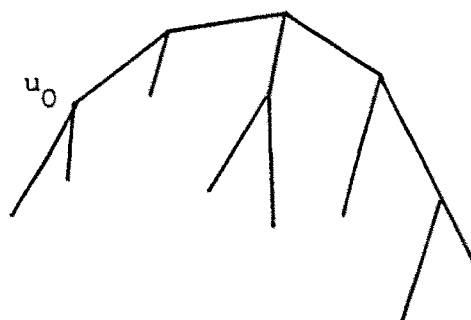


Figura 2

b) Una vez construida la forma del árbol se debe proceder a la estimación de la longitud de las aristas. Se considera la matriz X de orden $\binom{n}{2} \times (2n-3)$ representando la configuración del árbol, siendo n el número de objetos del conjunto S y $2n-3$ el número de aristas del árbol, por lo que $x_{ij} = 1$ si la distancia i -ésima contiene la arista j y $x_{ij} = 0$ en caso contrario, planteándose la estimación de las longitudes de las aristas $(\beta_1, \dots, \beta_{2n-3})$ por el método de los mínimos cuadrados, obteniendo como solución

$$\hat{\beta} = (X'X)^{-1} \cdot X' \cdot \delta$$

con lo que queda completado el algoritmo.

6.2.2.- Algoritmo de Cunningham (1978)

Este algoritmo, de corte muy parecido al anterior, tiene por objetivo la búsqueda de una distancia que verifique la condición del cuarto punto a partir de consideraciones sobre la forma del árbol según la disimilaridad inicial.

Si δ es la disimilaridad inicial y para cuatro elementos de S $\{i,j,k,l\}$ se verifica:

$$\delta_{ij} + \delta_{kl} < \delta_{ik} + \delta_{jl} < \delta_{il} + \delta_{kj}$$

se considera interesante que la distancia aditiva d buscada verifique

$$d_{ik} + d_{jl} = d_{il} + d_{kj}$$

De este modo se obtiene una colección de relaciones lineales entre las distancias, considerando sólo aquellas que resultan ser independientes. Estas relaciones se pueden expresar matricialmente como

$$C \cdot d = 0$$

donde C es una matriz de orden $r \times \binom{n}{2}$ siendo r el número de relaciones lineales impuestas, y de este modo si la i -ésima relación es

$$d_u + d_v = d_r + d_s$$

una vez numeradas las distancias (entre 1 y $\binom{n}{2}$) nos resulta

$$c_{ij} = \begin{cases} 1 & \text{si } i = u, v \\ -1 & \text{si } j = r, s \\ 0 & \text{resto} \end{cases}$$

Se procede seguidamente buscando la distancia aditiva d^* tal que

$$\| \delta - d^* \| = \min \{ \| \delta - d \| \mid C \cdot d = 0 \}$$

obteniendo como solución

$$d^* = \delta - C' \cdot (CC')^{-1} \cdot C\delta$$

Queda pues como último paso la reconstrucción del árbol, cuestión que ya hemos abordado en el capítulo 3.

En un intento de mejorar las posibilidades prácticas de aplicación del algoritmo anterior, De Soete (1983) propone un algoritmo basado en la minimización secuencial de

$$F(D,r) = \underbrace{\sum_{i,j} (d_{ij} - \delta_{ij})^2}_{L(D)} + r \underbrace{\sum_{\Omega} (d_{ik} + d_{jl} - d_{il} - d_{jk})^2}_{P(D)}$$

para una sucesión creciente de r y siendo

$$\Omega = \{ (i,j,k,l) \mid i,j,k,l \text{ distintos y } \delta_{ij} + \delta_{kl} \leq \min\{ \delta_{ik} + \delta_{jl}, \delta_{il} + \delta_{kj} \} \}$$

Este procedimiento, basado en la misma idea inicial del algoritmo de Cunningham, emplea sin embargo un algoritmo iterativo para la búsqueda de la distancia aditiva.

En concreto el algoritmo que propone comienza considerando

$$D^{(0)} = \{ \delta_{ij} + \epsilon_{ij} \mid i < j \}$$

siendo

$$\epsilon_{ij} \sim N(0, \sigma_{\epsilon}^2)$$

y

$$r^{(1)} = \frac{L(D^{(0)})}{P(D^{(0)})}$$

Se trata entonces de minimizar $F(D, r^{(q)})$ empezando por $D^{(q-1)}$ para obtener $D^{(q)}$, considerando $r^{(q)} = 10 \cdot r^{(q-1)}$

Se paran las evaluaciones cuando

$$\sum (d_{ij}^{(q)} - d_{ij}^{(q-1)})^2$$

es menor que una constante que hemos fijado de entrada.

En comparación con el algoritmo de Cunningham este método tiene la ventaja que se puede utilizar para conjuntos de datos con mayor número de elementos puesto que no requiere la inversión de matrices tan grandes como en el algoritmo anterior. En un estudio de simulación para evaluar la eficacia obtiene resultados parecidos a los obtenidos por ADDTREE.

6.2.3.- Algoritmo de Abdi (1985)

Al igual que los algoritmos anteriores H. Abdi (1985) propone un algoritmo dividido en dos fases: a) Determinación de la forma del árbol b) Estimación de la longitud de las aristas.

a) Forma del árbol.

Dada una disimilaridad δ sobre un conjunto S , se define la función

$$\phi_k : S \times S \longrightarrow N$$

de modo que

$$\phi_k(a,b) = n$$

si existen exactamente n subconjuntos U de $S - \{a,b\}$ tales que $|U| = k$ verificando

$$\delta(a,b) + \max_{x,y \in U} \{ \delta(x,y) \} < \max_{x,y \in U} \{ \delta(a,x) + \delta(b,y) \}$$

En caso de que δ fuera la distancia asociada a un árbol aditivo, el valor máximo de ϕ_k se encuentra precisamente entre los pares con un único nudo principal no terminal en el camino que los une (Figura 3).

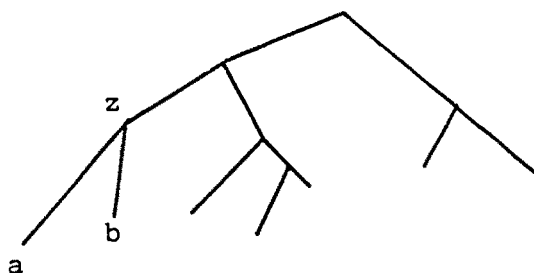


Figura 3

Así pues para la determinación de la forma del árbol, se unen en un primer paso los elementos a y b tales que $\phi_k(a,b)$ sea máximo y se considera entonces el elemento z representando al nudo constituido por $\{a,b\}$ y la nueva función sobre $(S - \{a,b\}) \cup \{z\} = S^{(1)}$

$$\phi_k^{(1)}(z,u) = \frac{1}{2}(\phi_k(a,u) + \phi_k(b,u))$$

$$\phi_k^{(1)}(x,y) = \phi_k(x,y) \quad \text{si} \quad \{x,y\} \cap \{a,b\} = \emptyset$$

reiterando entonces el proceso para $S^{(1)}$.

b) Estimación de las longitudes de las aristas.

Para la estimación de las longitudes de las aristas se propone un procedimiento heurístico basado en una estimación geométrica de las mismas. Observemos que para el árbol de la figura 3, si d es la distancia asociada al árbol se verifica que

$$d(z,u) = \frac{d(a,u) + d(b,u) - d(a,b)}{2}$$

para cualquier u nudo terminal.

Esta propiedad es la que sugiere el procedimiento de estimación a seguir.

Se considera el conjunto $S^* = S \cup \{g\}$ con la disimilaridad asociada δ^* definida como

$$\delta^*(u, u') = \delta(u, u') \quad \text{si } u, u' \neq g$$

$$\delta^*(g, u) = \sum_{u' \in S} \frac{\delta(u, u')}{n}$$

y se procede a la representación de (S^*, δ^*) en un espacio euclídeo (V, D)

$$\phi : (S^*, \delta^*) \longrightarrow (V, D)$$

considerando una nueva distancia sobre S^* definida en la forma

$$\bar{\delta}(x, y) = D(\phi(x), \phi(y))$$

Siguiendo paralelamente el proceso reiterativo de la construcción del árbol se procede para la estimación de la longitud de las aristas. Así si a y b se han unido en un nudo z , construimos el mismo de modo que la distancia entre z y g sea mínima. Posteriormente consideramos una nueva distancia sobre $S^{(1)}$ calculada a través de la geometría euclídea obtenida. Se reitera de este modo el proceso hasta la obtención de las estimaciones de las longitudes de las aristas.

En un estudio de simulación sobre la eficacia del método se obtienen resultados muy parecidos a los obtenidos para ADDTREE en Pruzansky et al.(1982).

6.2.4.- Algoritmo de Waterman (1978).

Waterman et al.(1978)desarrollan un algoritmo basado en métodos de programación lineal, efectuándose una construcción secuencial del árbol aditivo en que se añade un nuevo nudo terminal en cada paso minimizándose una función objetivo definida sobre las aristas.

Se consideran como restricciones: a) Las longitudes de las aristas deben ser no negativas. b) La distancia sobre el árbol entre dos objetos debe ser mayor o igual que la disimilaridad observada.

Así, si anotamos por e_1, \dots, e_m las aristas del camino entre los puntos i y j y δ_{ij} es la disimilaridad observada entre los mismos, tendremos que

$$\sum_{i=1}^m e_i > \delta_{ij}$$

$$e_i > 0 \quad \forall i, j \in S \quad (4)$$

obteniendo pues un poliedro convexo, por lo que se alcanza la solución al problema en un punto extremo del mismo (Roberts & Varberg, 1973).

Se sugiere tomar como función objetivo alguna de entre las siguientes

$$a) f_1(\{e_i\}) = \sum_{i=1}^{2n-3} e_i \quad (\text{simboliza la evolución total})$$

$$b) f_2(\{e_i\}) = \sum_{l,k} \frac{(\sum_{e_i \in C_{lk}} e_i) - \delta_{lk}}{\delta_{lk}}$$

(sugerida por Fitch & Margoliash, 1967)

(C_{lk} = aristas en el camino entre l y k)

$$c) f_3(\{e_i\}) = \sum_{l,k} \frac{(\sum e_i) - \delta_{lk}}{\frac{C_{lk}}{\delta_{lk}^2}}$$

El algoritmo consistirá en unir en cada paso un nuevo vértice minimizando la función según las restricciones (4). A fin de evitar el problema de la dependencia del orden en la unión de los vértices terminales del árbol se sugiere en una extensión del método unir en cada paso el vértice terminal que proporcionará un valor menor para la función objetivo.

No se obtiene sin embargo un mínimo global mediante este procedimiento, tal y como demuestra Waterman con un contraejemplo. Por la complejidad calculística del procedimiento se sugiere hallar árboles razonablemente válidos y a partir de los mismos estimar las aristas por el procedimiento descrito anteriormente.

6.2.5.-

En esta memoria sugerimos además la utilización de un algoritmo basado en la búsqueda a priori de una configuración adecuada del árbol. En este sentido aparte los métodos de Tversky, Cunningham y Abdi mencionados existe la posibilidad de buscar una clasificación taxonómica de los objetos mediante algún método de clasificación jerárquica; en particular el método UPGMA (con importantes propiedades con respecto al coeficiente de correlación cofenética) y realizar posteriormente una transformación del árbol ultramétrico hallado dentro de la misma carta de la variedad de modo que se ajuste según algún criterio de optimización a la disimilaridad observada. Así el algoritmo se podría entender como una aplicación definida en la forma:

$$\phi: D_S \longrightarrow D_u \longrightarrow D_A$$

$$\delta \longrightarrow d_u \longrightarrow d_a$$

siendo D_S = conjunto de disimilaridades sobre S.

D_u = " " " ultramétricas sobre S.

D_A = " " " aditivas sobre S.

El paso de la distancia ultramétrica a la distancia aditiva podría efectuarse de muchas maneras de entre las cuales sugerimos las siguientes:

a) Mediante el método de los mínimos cuadrados .

b) Calculando $\phi : S \longrightarrow R$ tal que

$$\sum_{\substack{i,j \\ i \neq j}} (\delta(i,j) - d_u(i,j) - \phi(i) - \phi(j))^2$$

sea mínima , considerando entonces la distancia aditiva

$$d_a(i,j) = d_u(i,j) + \phi(i) + \phi(j) \quad (5)$$

Si indicamos $y_{ij} = \delta(i,j) - d_u(i,j)$, $\phi(i) = \alpha_i$ se trata de minimizar

$$\sum_{i \neq j} (y_{ij} - \alpha_i - \alpha_j)^2$$

De este modo, derivando parcialmente con respecto α_i para todo i , obtenemos las ecuaciones

$$\sum_{\substack{j=1 \\ j \neq i}}^n (y_{ij} - \alpha_i - \alpha_j) = 0 \quad (6)$$

es decir

$$\sum_{\substack{j=1 \\ j \neq i}}^n y_{ij} - (n-1) \cdot \alpha_i - \sum_{\substack{j=1 \\ j \neq i}}^n \alpha_j = 0 \quad (7)$$

sumando las ecuaciones (6)

$$\sum_{i=1}^n \left(\sum_{\substack{j=1 \\ i \neq j}}^n (y_{ij} - \alpha_i - \alpha_j) \right) = 0$$

de donde

$$\sum_{i \neq j} y_{ij} - (2n-2) \cdot \sum_{k=1}^n \alpha_k = 0 \quad (8)$$

De (7) y (8) se deduce que

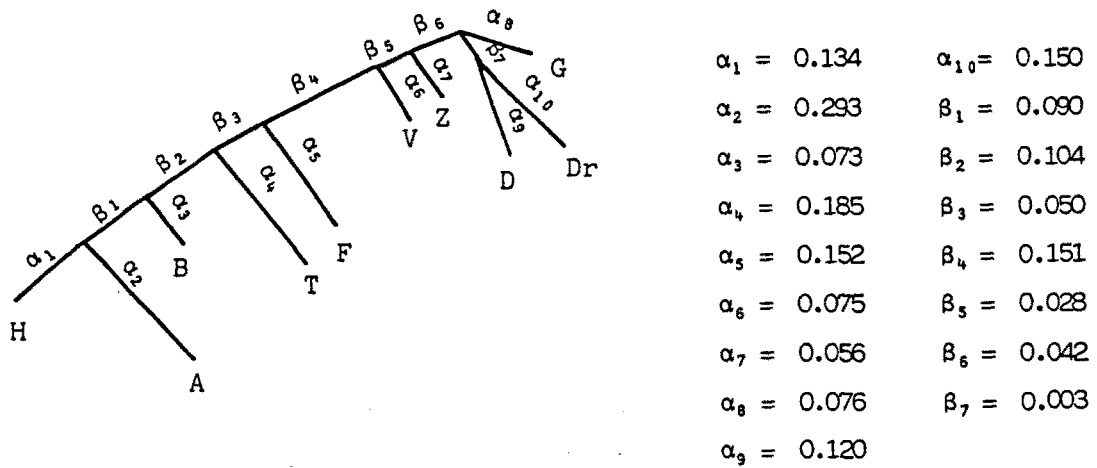
$$\alpha_i = \frac{(2n-2) \cdot \sum_{k=1}^n y_{ik} - \sum_{i \neq j} y_{ij}}{2 \cdot (n-2) \cdot (n-1)} \quad (9)$$

Así pues consideraremos que existe una solución mediante este algoritmo cuando podemos hallar d_u distancia ultramétrica y ϕ definida a través de (9) de modo que d_a construida en (5) resulta ser distancia. La ventaja de este método consiste en la posibilidad de relacionar una representación ultramétrica y un árbol aditivo asociados a δ .

6.3.- EJEMPLO

Este ejemplo tiene por objetivo realizar una ilustración de las representaciones mediante árboles aditivos. Se trata de una aplicación de las mismas al estudio de la distribución geográfica de *Drosophila Subobscura* a partir de la distancia genética propuesta por Prevosti (1974), la cual cuantifica la diferencia entre dos poblaciones a partir de las frecuencias de las ordenaciones cromosómicas producidas por inversiones y que ha sido descrita en 2.6.

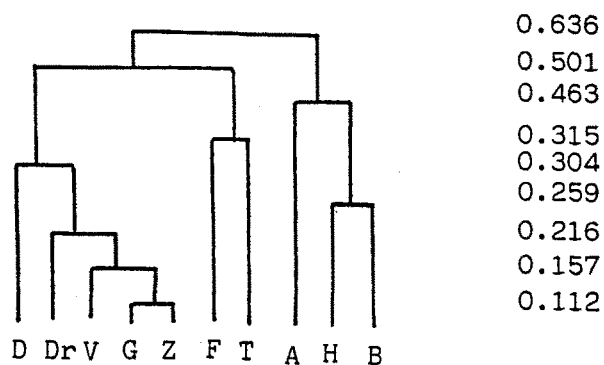
utilizando el algoritmo ADDTREE descrito en 6.2.1 obtenemos el árbol aditivo



Coeficiente correlación entre la distancia original y la distancia aditiva = 0.9965.

Figura 4

y mediante el método de clasificación jerárquica UPGMA descrito en 2.4 obtenemos el árbol ultramétrico



Coeficiente de correlación cofenética = 0.83238

Figura 5

Podemos comprobar que si bien mediante el dendograma obtenido utilizando el método UPGMA se pueden distinguir 3 clases significativas que se pueden entender como Sureste, Suroeste y Norte, coincidiendo pues con los resultados obtenidos por Alonso(1975) en que utilizaba otros métodos de representación, la distancia ultramétrica obtenida presenta en algunos casos los problemas comentados en 2.6, por lo que es preferible la representación utilizando árboles aditivos.

El árbol aditivo obtenido (Fig.4) se podría interpretar como la evolución de la especie en Europa según las poblaciones estudiadas, confirmándose a través del mismo la tendencia de expansión de Suroeste a Noreste y observándose como el Centro-Norte de Europa está más próximo al Sureste que al Suroeste, lo que induce a pensar en la hipótesis de una colonización Suroeste - Sureste - Norte. Esta hipótesis halla una justificación biológica en el hecho de que la colonización o recolonización del continente europeo tuvo lugar después de la última glaciación resultando ser la capacidad de dispersión de la especie mayor que la velocidad de retirada de los hielos. La elevada capacidad de dispersión se ha comprobado experimentalmente mediante el estudio de la colonización del continente americano por la especie europea *D.Subobscura*. En este análisis se ha observado que dicha especie ha ocupado Chile, desde La Serena (29° 55' Lat. Sur) hasta Coyhaique (45° 35' Lat.Sur) (3.000 Km. aprox.) en un periodo de tan sólo tres años.

CONCLUSIONES

En esta memoria se ha desarrollado una metodología para abordar problemas derivados de la representación de un conjunto mediante una clase especial de grafos llamados árboles aditivos y se ha realizado un estudio formal de algunos de los principales aspectos que surgen en la utilización de este tipo de representaciones.

Se ha situado en primer lugar este método de representación dentro de los métodos propios del análisis multivariante, resaltando las ventajas que tiene frente otros modelos de representación discretos como pueden ser las clasificaciones jerárquicas de la taxonomía numérica y las representaciones utilizando modelos continuos, explicitando los principales resultados y características de todos ellos. Se ha observado el interés que tiene el estudio de medidas para poder decidir objetivamente entre un modelo de representación continuo y un modelo de representación discreto, analizando los resultados obtenidos sobre el particular por Pruzansky et al. (1982) en que sugieren que el momento centrado de tercer orden y la proporción de triángulos elongados pueden ser efectivos para resolver el problema mencionado.

Se ha desarrollado después una nueva formalización sobre las relaciones entre una distancia verificando el axioma del cuarto punto y un árbol aditivo, hallando en términos de la misma formalización las relaciones entre distancias aditivas y ultramétricas. Las líneas seguidas en el estudio anterior han sido comparadas con las desa

rrolladas en Buneman (1971) donde se relacionan distancias aditivas y árboles aditivos desde otra perspectiva.

Se ha conseguido después dotar el conjunto de distancias aditivas con configuración de una estructura de variedad diferenciable con frontera, calculando explícitamente la misma y estudiando las distancias ultramétricas dentro de la variedad. Se advierte del interés que puede tener esta estructura en el estudio de problemas relativos a algoritmos de construcción, inferencias en árboles y representación de los mismos por modelos continuos.

A partir de la asociación de un modelo probabilístico a un vector de distancias obtenemos la resolución de diversos contrastes formulados, en general, en una carta de la variedad constituida por las distancias aditivas con configuración.

Se ha expuesto por último un análisis de los principales algoritmos de transformación de una disimilaridad definida sobre un conjunto en una distancia aditiva. Esta exposición se ha ilustrado con un ejemplo en que se muestra la utilización de los árboles aditivos como técnicas de representación. En el mismo se estudian algunos aspectos de la colonización europea de *Drosophila Subobscura* bajo la hipótesis histórica nosselectiva corroborando la teoría que la misma se ha producido en dirección Suroeste a Noroeste, pudiéndose observar algunas ventajas que supone esta técnica de representación frente a un dendograma.

Los temas desarrollados en esta memoria se podrían extender en las siguientes líneas:

- a) Estudio y comparaciones de aspectos teóricos y prácticos de los algoritmos de construcción de árboles aditivos.
- b) Estudiar las relaciones que existen entre los distintos modelos de representación.
- c) Desarrollar nuevas vías sobre la inferencia en árboles aditivos que puedan solucionar problemas tales como: significación de clases, elección de la configuración, comparación de modelos.
- d) Estudios de distancias asociadas a modelos discretos más comple--jos.

BIBLIOGRAFIA

- ABDI, H. (1985) . Tree representations of associative structures in semantic and episodic memory research. Trends in Mathematical Psychology,3-31 , Elsevier Science Publishers B.V. (North-Holland)
- ALONSO,G. (1975) . Estudio de la distribución geográfica del polimorfismo cromosómico en Drosophila Subobscura. Tesina, Fac.Biología,Univ.Barc.
- ANDERSON, J.J. (1985) . Normal mixtures and the number of clusters problem. Computational Statistics Quarterly,V.2, Issue1,3-14.
- ANDERSON,T.W. (1958) . An introduction to multivariate statistical analysis. John Wiley & Sons,Inc.,New York .
- ARCAS ,A. (1983) . Contribuciones a la construcción de clasificaciones estratificadas. Tesina, Fac.Matemáticas, Univ. Barcelona.
- ARCAS, A. (1984) . Sobre la no unicidad de clasificaciones jerárquicas asociadas a un método de clasificación taxonómico. Actas del XIV congreso nacional de estadística,inv. oper. e inform.
- ARCAS,A. y SALICRU,M. (1984). Sobre la no unicidad de la clasificación jerárquica asociada a una disimilaridad por los métodos del máximo y -UPGMA .Questió, V.8, Nº3,113-120.
- BALL,G.H. y HALL,D.J. (1967) . A clustering technique for summarizing multivariate data. Behavioral Science, 12(2),153-155.
- BARTLETT,M.S. (1948) . A note on the statistical estimation of demand and supply relations from time series. Econometrica, 16, p.323.
- BERGE, C. (1959) . Espaces topologiques.Fonctions multivoques.Dunod,Paris.
- BERGE, C. (1973) . Graphes et Hypergraphes. Dunod, Paris.
- BERTRAND,P. y Diday,E. (1985) . A Visual representation of the compatibility between an order and a dissimilarity index:the pyramids. Computational Statistics Quarterly, V.2,Issue 1,1985,31-41 .
- BHATTACHARYYA,A. (1942) . On a measure of divergence between two multinomial populations. Sanhkyā,7,401,406.
- BILLINGSLEY, P. (1968) . Convergence of Probability Measures.John Wiley , New York.
- BIOMETRIKA TABLES FOR STATICIANS VOL.2. Edited by E.Pearson and H.O.Hartley (1976) .
- BUNEMAN, P. (1971) . The recovery of trees from measures of dissimilarity . F.R. Hodson, D.G.Kendall,P.Tautu (Edit.)Mathematics in the Archeological and historical sciences. Edinburgh University Press.

- CAILLIEZ, F. (1983) . The analytical solution of the additive constant problem. Psychometrika, V.48, N°2, 305-308 .
- CARROLL, J.D. (1976). Spatial, non-spatial and hybrid models for scaling. Psychometrika, V.41, 4, 439-463.
- CHILLINGWORTH , D.R.H. (1976). Differential topology with a view to applications. Pitman Publishing, London.
- COOPER, C.H. (1972) . A new solution to the additive constant problem in metric multidimensional scaling. Psychometrika, 37, 311-322.
- CUADRAS, C.M. (1980) . Metodes de representació de dades i la seva aplicació en Biologia. Col.Soc.Catalana Biologia, 13, 95-133.
- CUADRAS, C.M. (1980) . Curso de Análisis de la Varianza. Publicaciones de Bioestadística y Biomatemática N°1, Barcelona.
- CUADRAS, C.M. (1981) . Métodos de análisis multivariante. Eunibar, Barcelona.
- CUADRAS, C.M. (1983) . Análisis algebraico sobre distancias ultramétricas. Actas 44 Per. de Sesiones del Inst.Intern. de Estadística, Madrid, V.2, - 554-557.
- CUADRAS, C.M. y Oller, J.M. (1984) . Geometría finita aplicada a la estadística. Actas XIV congreso nacional de estadística, inv.oper. e inform.
- CUADRAS, C.M. y Oller, J.M. (1985) . Eigenanalysis and metric multidimensional scaling on hierarchical structures. Sometido a Psychometrika.
- CUADRAS, C.M., Oller, J.M., Arcas, A. y Rios, M. (1986). Métodos geométricos de la estadística. Questiô . (En prensa)
- CUNNINGHAM, J.P. (1978). Free trees and bidirectional trees as representations of psychological distance . Journal of Mathematical Psychology, 17, 165-188.
- DALLOT , S. y Ibañez, F. (1972). Etude preliminaire de la morphologie et de l'evolution chez les chaetognates. Inv.Pesq. 36(1) , 31-41.
- DE SOETE, G. (1983) . A least square algorithm for fitting additive trees to proximity data. Psychometrika, 48(4), 621-626.
- DE LEEUW, J. y HEISER, W. (1982) . Theory of multidimensional scaling . Handbook of statistics V.2, North Holland Publising Company, 285-316.
- ECKART, C. y Young, G. (1936). The aproximation of one matrix by another of lower rank. Psychometrika, 1, 211-218.
- FARRIS, J.S. (1969) . On the cophenetic correlation coefficient. Syst.Zoology 18(3), 279-285 .

- FITCH, W. y Margoliash, E. (1967). Construction of phylogenetic trees. Science, V.155, 279-284 .
- GNANADESIKAN, R. (1977). Methods for statistical data analysis of multivariate observations. John Wiley, New York.
- GONDRAN, M. (1976). La structure algebraique des classifications hierarchiques. Annales de l'insee, N° 22-23, 181-190.
- GOWER, J.C. (1967). A comparison of some methods of cluster analysis. Biometrics, 23, 623-637.
- GOWER, J.C. (1982). Euclidean distance geometry. Math. Scientist, 7-14.
- HICKS, N.J. (1965). Notes on differential geometry. Van Nostran, Princeton.
- HOLMAN, E.W. (1972). The relation between hierarchical and euclidean models for psychological distances. Psychometrika, V.37, N°4, 417-423.
- HUBERT, L. y ARABIE, P. (1985) . Comparing Partitions . Com.Presentada en "Fourth european meeting of the Psychom. Soc. and the Class.Soc.", Cambridge.
- HYVER (1973) . Valeurs propres des systèmes de transformation representables par des graphes en arbres. J.Theoret.Biol., 42, 397-409.
- JARDINE, N. y SIBSON, R. (1971). Mathematical Taxonomy. John Wiley, New York.
- JOHNSON, S.C. (1967). Hierarchical clustering schemes. Psychometrika V.32, 241, 254.
- KIMURA, M. (1984) . Neutral evolution as an inevitable process of change at the molecular level. Darwin a Barcelona, P.P.U., 231-252.
- KOHNE, D.E., CHISCON, J.A. y HOYER, B.H. (1972). Evolution of primate DNA sequences. J.Human Evol. 1, 627-644.
- KRUSKAL, J.B. (1964) . Nonmetric multidimensional scaling: A numerical method. Psychometrika, 29, 115-129 .
- KRUSKAL, J.B., YOUNG, F.W., SEERY, J.B. (1973) . How to use KYST a very flexible program to do multidimensional scaling and unfolding. Bell Laboratories, Murray Hill.
- KRUSKAL, J.B. y WISH, M. (1978) . Multidimensional scaling. Sage Publication, Beverly Hills.
- LANG, S. (1962) . Introduction to differentiable manifolds. Interscience Publishers, New York.
- LERMAN, I.C. (1981) . Classification et analyse ordinale des données. Dunod, Paris.
- LINGOES, J.C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. Psychometrika, 36, 195-203.
- MARDIA, K.V. (1978). Some properties of classical multidimensional scaling. Comm.Stat. A7 (13), 1233-1241.

- MARDIA, K.V., KENT, J.T. y BIBBY, J.M. (1979). Multivariate Analysis. Academic Press, London
- MARRIOTT, F.H.C. (1982). Optimization methods of cluster analysis. Biometrika, V.69,2,417-421.
- MATULA, D.W. (1977). In Classification and Clustering, J. Van Ryzin ed., Academic Press, New York, 95-129.
- MATUSITA, K. (1957). Decision rule based on the distance for the classification problem. Ann. Inst. Statist. Math. 8, 67-77.
- MOOD, A. y GRAYBILL, F. (1969). Introducción a la teoría estadística. Aguilar, Madrid.
- OLLER, J.M. (1983). Utilización de métricas riemannianas en análisis de datos multidimensionales y su aplicación a la biología. Publicaciones de bioestadística y biomatemática N°11, Barcelona.
- OLLER, J.M. y CUADRAS, C.M. (1985a). Rao's distance for negative multinomial distributions. Sankhya, V.47,A,75-83.
- OLLER, J.M. y CUADRAS, C.M. (1985b). Sobre ciertas condiciones que deben verificar las distancias entre espacios probabilísticos. Actas XV Reunión nacional de estadística, Invest. Oper. e Inform.
- OCAÑA, J. (1975). Sobre la distancia genética. Tesina, Fac. Biolog., Univ. Barc.
- OHSUMI, N. y NAKAMURA, T. (1981). Some properties of monotone hierarchical dendrogram in numerical classification. Proc. Inst. Statist. Mathem., 28(1), 117-133.
- PETIT PIERRE, E. y CUADRAS, C.M. (1977). The canonical analysis applied to the taxonomy and evolution of the genus *Timarcha* Latr. Mediterránea, 1, 13-28.
- PREVOSTI, A. (1974). La distancia genética entre poblaciones. Miscellanea Alcobé, Univ. Barcelona, 109-118.
- PREVOSTI, A.; OCAÑA, J. y ALONSO, G. (1975). Distances between population of *Drosophila Subobscura* based on chromosome arrangement frequencies. Theor. and Appl. Genetics 45, 231-241.
- PRUZANSKY, S.; TVERSKY, A. y CARROLL, J.D. (1982). Spatial versus tree representations of proximity data. Psychometrika, V.47,1,3-24.
- RAO, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcuta Math. Soc., 37, 81-91.
- RAO, C.R. (1948). The utilization of multiple measurements in problems of biological classification. J. Roy. Statistical. Soc., B.10, 159-193.

- RAO, C.R. (1973). Linear Statistical Inference and its Applications. John Wiley, New York.
- RAO, C.R. (1982). Diversity and dissimilarity coefficients: a unified approach. J. Theoretical Population Biology, 21,24-43.
- RIOS, M. (1985). Distancias entre modelos lineales y su aplicación a la clasificación y diagnóstico de enfermedades. Tesis doctoral, Fac. de Medicina, Univ. de Barcelona.
- ROHLF, F.J. (1970). Adaptive hierarchical clustering schemes. Syst.Zool., 19,58-82.
- ROHLF, F.J. (1981). Spatial representation of phylogenetic trees computed from dissimilarity matrices. Inter.Symp. Compt. Meth.Paleo., Barcelona, 303-311.
- ROBERTS, W. y VARBERG, D. (1973). Convex Functions. Academic Press, New York.
- ROMEDER, J.M. (1973). Méthodes et programmes d'analyse discriminante. Dunod, Paris.
- SAITO, T. (1978). An alternative procedure to the additive constant problem in metric multidimensional scaling. Psychometrika, 43,193-201.
- SALICRU, M. (1983). Consideraciones sobre desemejanzas y clasificaciones asociadas. Tesina, Fac. Matemáticas, Univ. Barcelona.
- SATTATH, S. y TVERSKY, A. (1977). Additive similarity trees. Psychometrika, 42(3), 319-345.
- SEARLE, S.R. (1971). Linear Models. John Wiley, New York.
- SCHEFFE, H. (1959). The analysis of variance. John Wiley, New York.
- SHEPARD, R.N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika, 27,219-246.
- SNEATH, P.H.A. y SOKAL, R.S. (1973). Numerical Taxonomy. W.H. Freeman and Co., San Francisco
- SNEATH, P.H.A. (1980). Some empirical tests for significance of clusters. Data analysis and informatics. E. Diday et al. (eds.). North Holland Publishing Company, 491-507.
- SOKAL, R.R. y MICHENER, C.D. (1958). A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull., 38,1409-1438.
- SOKAL, R.R. y SNEATH, P.H.A. (1963). Principles of numerical taxonomy. W.H. Freeman and Co., San Francisco.
- SPIVAK, M. (1979). A comprehensive introduction to differential geometry. Publish or Perish, Inc. Berkeley.
- TORGERSSON, W.S. (1952). Multidimensional Scaling: I. Theory and method. Psychometrika, 17, 401-419.

- TORGERSON, M.S. (1958). Theory and methods of scaling. John Wiley, New York.
- TUCKEY, J.W. (1977). Exploratory data analysis. Addison-Wesley, Reading.
- WAGENAAR, W. (1971). Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. British J. of Mathematical and Statistical Psychology, 24, 101-110 .
- WATERMAN, M.S.; SMITH, T.F.; SINGH, M. y BEYER, N.A. (1977). Additive evolutionary trees. J. Theor. Biol., 64, 199-213.
- WILKS, S. (1962) . Mathematical statistics. John Wiley, New York.
- WISH, M. (1970) . Comparations among multidimensional structures of nations based on different measures of subjective similarity. General Systems, 15, 55-65.
- WISH, M.; DEUTSCH, M.; BIENER, L. (1972). Differences in perceived similarity of nations. P.P. 289-313 in A.K. Romney, R.N. Shepard and S. Nerlove (eds). Multidimensional Scaling: Theory and Applications in the Behavioral - Sciences, V.2, New York: Seminar Press.
- WISH, M. y CARROLL, J.D. (1982). Multidimensional Scaling and its Applications. P.R. Krishnaiah and L.N. Kernel Eds. , Handbook of Statistics, V.2, 317-345.