**Qüestiió** - V. 9, n.º 4 (desembre 1985) pp. 219-250

# MÉTODOS GEOMÉTRICOS DE LA ESTADÍSTICA

C. M. CUADRAS, J. M. OLLER, A. ARCAS, M. RÍOS
UNIVERSITAT DE BARCELONA

*Se exponen diversos métodos geométricos, insistiendo en que propiedades se fundamentan, en orden a probar que diferentes espacios geométricos (euclídeo, no euclídeo, ultramétrico, - aditivo, riemanniano) juegan un importante papel en el análisis estadístico de datos.*

Keywords: MULTIVARIATE DATA ANALYSIS, ULTRAMETRIC DISTANCE,ADDITIVE INEQUALITY, GEODESIC DISTANCE, MULTIDIMENSIONAL SCALING.

## 1. INTRODUCCION.

La estadística tiene como soporte teórico diversas ramas de la matemática: el análisis, la probabilidad, el álgebra, la geometría , etc.

En este trabajo se demuestra como la estadística, especialmente las técnicas de análisis de datos multidimensionales, constituye un conjunto de fundamentos teóricos, criterios, propiedades y reglas de decisión que se basan en la geometría, y más exactamente, en los espacios métricos, es decir, los espacios topológicos, cuya topología viene inducida a través de una distancia.

Numerosos autores han contribuido a lo que K. Pearson llamó "Geometría de la Estadística". (Galton, Graun, Petty, Halley, Plaifair, Pearson, Fisher , Kendall, Gower, Déparcieux, Laplace, Quetelet, Benzecri, Lebart, Carroll, Gnanadesikan, Goodman, Dempster, Kruskal, Roy, Rao, Mahalanobis, Burbea, Matusita, Bose, Lerman, Shepard, Kruskal, Carroll, Sokal, Rohlf, etc.).

Así, existe el precedente de K. Pearson, quien entre Noviembre de 1891 y Enero de 1892 dictó un ciclo de conferencias (véase /54/), destacando que los métodos geométricos de representación de datos constituyen un aspecto fundamental en la investigación estadística.

Se reconoce también que las grandes contribuciones de R.A. Fisher a la estadística se deben, en buena parte, a que aplicó los criterios geométricos al análisis estadístico.

R.C. Bose introduce la llamada geometría parcial, en la que relaciona 4 axiomas sobre puntos y rectas con ciertos diseños parcialmente balanceados, identificando los vértices - geométricos con tratamientcs (véase /37/).

Quizás la mejor ilustración geométrica de un método estadístico lo constituye el modelo - lineal /57/

$$Y = X\beta + e$$

en el cual :

1) El vector Y se interpreta como un punto del espacio euclídeo $R^n$.

2) Las columnas de la matriz x generan un subespacio $F \subset R^n$, llamado espacio estimación.

3) La estimación por mínimos cuadrados de $\beta$ es aquel $\hat{\beta}$ tal que $X\hat{\beta}$ es la proyección de Y sobre F. Entonces se estima la varianza $\sigma^2$ del modelo a partir de la distancia -- euclídea

- C.M. Cuadras, J.M. Oller, A. Arcas y M. Ríos - Universitat de Barcelona - Facultat de Biologia Dep. de Bioestadística - Av. Diagonal, 637 - Barcelona 08028

- Article rebut el juliol de 1985.

$$|| Y - X\hat{\beta} ||$$

donde $(Y - X\hat{\beta})$ pertenece a $F^{\perp}$ , llamado espacio error.

4) Las hipótesis lineales y las funciones paramétricas estimables se identifican como subespacios $G \subset F$

5) La formulación geométrica de una hipótesis lineal es

$$H_o: E(Y) \in G \qquad H_1: E(Y) \in F$$

La hipótesis $H_o$ se rechaza si la distancia de Y a G es significativamente mayor que la distancia a F.

## 2. ESPACIOS GEOMETRICOS MODELOS.

En un problema de análisis estadístico de datos multidimensionales, dispondremos, en general, de un conjunto de individuos.

$$I = \{i_1, i_2, \ldots, i_n\}$$

y de un conjunto de variables $Y_1, Y_2, \ldots, Y_p$, que usualmente serán variables aleatorias.

Consideraremos entonces los siguientes espacios:

a) El espacio métrico $(I,\delta)$ donde $\delta$ es una distancia, es decir, una aplicación $\delta: I \times I \longrightarrow R$, que verifica las siguientes propiedades:

$$\begin{aligned} &\text{i)} \quad \delta(i,j) \geq 0 \\ &\text{ii)} \quad \delta(i,j) = \delta(j,i) \\ &\text{iii)} \quad \delta(i,i) = 0 \end{aligned} \qquad (1)$$

La distancia $\delta$ es una medida de la diferencia entre los individuos y constituye la información original del estadístico. $\delta$ puede calcularse sobre variables cualitativas --- (distancia de Battacharyya, ji-cuadrado, basada en índices de similaridad, etc.), cuantitativas (distancia de Mahalanobis, K.Pearson, euclídea, Minkowiski, etc.), sobre distribuciones de probabilidad (distancia de Fréchet, Levy, Hellinger, Matusita, Rao, -- etc.). Pero $\delta$ también puede obtenerse, en aplicaciones a la Psicología por ejemplo, preguntando a un grupo de sujetos el grado de similaridad entre cada par de individuos.

b) El espacio vectorial E generado por las variables $Y_i$

$$E = \{Y \mid Y = \sum_i a_i Y_i\}$$

Usualmente la métrica en E viene dada por la matriz de covarianzas $\Sigma$ . Entonces la norma de un vector se identifica con la varianza, y el coseno del ángulo entre dos vectores se identifica con la correlación.

c) El espacio $E^*$ dual del espacio E.

$$E^* = \{\alpha \mid \alpha \text{ es una forma lineal sobre } E\}$$

Usualmente, la métrica en $E^*$ viene dada por $\Sigma^-$ , donde $\Sigma$ es una g- inversa de $\Sigma$ . Se trata entonces de la métrica en $E^*$ inducida -- por la métrica en E.

d) El espacio modelo (V, d), cuya estructura es bien conocida, y que ha de servirnos de modelo de referencia para representar $(I,\delta)$. Diremos que (V, d) es el espacio modelo.

(V, d) es el resultado final de una representación de $(I,\delta)$. Según las diferentes propiedades de la distancia d, podemos hablar de diferentes formas geométricas de representación. En este trabajo expondremos diferentes geometrías sobre (V, d), a saber:

- euclídea
- no euclídea
- ultramétrica
- aditiva
- riemanniana

Una realización de I sobre V es una aplicación inyectiva e isométrica.

$$f: I \longrightarrow V$$

$$\delta(i,j) = d(f(i), f(j)) \qquad (2)$$

Una realización monótona de I sobre V es una aplicación f tal que

$$g(\delta(i,j)) = d(f(i), f(j)) \qquad (3)$$

donde g es alguna función monótona creciente.

En algunos planteamientos teóricos, la relación (3) es exacta. Pero en las aplicaciones, convendría que para una cierta función de ajuste $\| . \|$, la cantidad

$$\| g(\delta) - d \| \qquad (4)$$

sea mínima.

En (2) describimos exactamente $(I,\delta)$ a través del espacio modelo $(V,d)$. En (3), siendo (4) mínimo, describimos aproximadamente $(I,\delta)$ a través de $(V, d)$.

Un primer ejemplo de realización se obtiene definiendo

$$\begin{aligned} f:\ & I \longrightarrow E^* \\ & i \longrightarrow i^* \end{aligned} \qquad (5)$$

siendo $i^*$ tal que

$$i^*(Y) = Y(i) \qquad \forall Y \in E$$

Es decir, para toda variable Y, el valor $y = Y(i)$ que la variable toma sobre el individuo i, define una forma lineal $i^*$ que es la imagen de i. Entonces, la estructura geométrica de $E^*$, inducida por ejemplo por la matriz $\Sigma^-$, puede servirnos para representar I. Dempster /27/ realiza una excelente exposición de análisis multivariante basándose en la realización (5). Véase también /9/, /12/, /18/.

También tienen interés realizaciones del espacio de las variables E en un espacio modelo $(V,d)$ (generalmente un espacio euclídeo), que se suele combinar con una realización de I, dando lugar a las llamadas representaciones simultáneas (análisis de datos centrados, método biplot, análisis de correspondencias). Entonces la realización es

$$\begin{aligned} f_1:\ & I \\ & \searrow \\ & \quad V \\ & \nearrow \\ f_2:\ & E \end{aligned} \qquad (6)$$

donde una determinada métrica en V permite representar a la vez individuos y variables.

Finalmente, debemos considerar también el caso de que I sea un espacio de objetos a representar, $S = \{s_1,\ldots,s_q\}$ un espacio de sujetos, y considerar q espacios métricos

$$(I, \delta_1), (I, \delta_2),\ldots,(I, \delta_q)$$

donde $\delta_k$ es la distancia que el sujeto k imprime a I.

Entonces existirán dos clases de realizaciones:

a) La de I, o espacio común de los objetos

$$f: I \longrightarrow V$$

b) La particular del sujeto $s_k$

$$\phi_k : I \longrightarrow V_k$$

donde $V_k$ es la imagen de I desde el punto de vista del sujeto $s_k$.

Describimos a continuación las diferentes geometrías que resultan según la estructura del espacio modelo $(V, d)$.

## 3. GEOMETRIA EUCLIDEA.

Sea $V = R^m$, d la distancia euclídea

$$d(x,y) = \sqrt{(x-y)'(x-y)} \qquad (7)$$

Entonces $(I,\delta)$ es isométrico a un subconjunto finito de $R^m$, y es representable a través de unas coordenadas euclídeas X, donde X es una matriz nxm, cuyas n filas dan las n coordenadas de las n individuos de I.

Un aspecto fundamental de la realización $(I,\delta)$ en $(V,d)$, es establecer en que condiciones la realización euclídea existe, es decir, $\delta$ puede ser identificada como una distancia euclídea.

Sean

$$\text{card}(I) = |I| = n$$

$$\Delta = (\delta_{ij}) \qquad \delta_{ij} = \delta(i,j) \qquad i,j \in I$$

donde $\Delta$ es la matriz simétrica nxn de las distancias. Existen varios teoremas que caracterizan $\delta_{ij}$ como distancia euclídea. Indiquemos

I matriz identidad nxn
$1 = (1,\ldots,1)'$ vector nx1 formado por unos
$A = (-\frac{1}{2}\delta_{ij}^2)$ matriz nxn

$H = I_n - \frac{1}{n} 11'$ matriz idempotente

Teorema 1. $\Delta$ es euclídea si y sólo si

$\forall\ v \epsilon R^n$ tal que $v'1=1$, $v'A \neq 0 \Rightarrow$

$B_v = (I-1v')A(I-v1') \geq 0$

($B_v \geq 0$ significa matriz semidefinida positiva).

La caracterización que tiene consecuencias más importantes está contenida en el siguiente

Teorema 2. $(I;\delta)$ es realizable en $(R^m,d)$ y por tanto $\Delta$ es euclídea, si y sólo si

$$B = HAH \geq 0 \qquad \text{ran } B = m \tag{8}$$

Las coordenadas euclídeas de la realización están contenidas en la matriz X (nxm) tal que

$$\begin{aligned} B &= XX' \\ X'X &= D_\lambda \end{aligned} \tag{9}$$

donde $D_\lambda$ es la matriz diagonal con los valores propios de B, y por tanto X contiene los vectores propios $\lambda$-normalizados. Obsérvese que el teorema 2 es una consecuencia del teorema 1, tomando

$$v = \left(\frac{1}{n},\ldots,\frac{1}{n}\right)'$$

Esta elección de v hace que el origen de coordenadas coincida con el baricentro de la configuración X. Otras elecciones de v tienen interesantes propiedades geométricas /32/. Es sorprendente que los teoremas 1 y 2 fundamentales en geometría euclídea, no fueran demostrados hasta 1935 /58/.

-

## 3.1 ANÁLISIS DE COMPONENTES PRINCIPALES.

Como es sabido, la descomposición espectral de la matriz de covarianzas $\Sigma$

$$\Sigma = TD_\lambda T' = A.A' \tag{10}$$

define p componentes principales, que se interpretan como variables unitarias y ortogonales. A través de la matriz A tenemos una realización del espacio de la variables

$$E \xrightarrow{\ A\ } R^n$$

Tomando las n primeras columnas de A, que dan las coordenadas de las variables originales en las m primeras componentes, tenemos una proyección

$$E \longrightarrow R^m \qquad (m<p) \tag{11}$$

tras la cual representamos las p variables en dimensión reducida /13/.

En análisis factorial /17/, la descomposición (10) es

$$\Sigma = A.A' + F^2 \tag{12}$$

donde F es una matriz diagonal, A es la matriz factorial. (12) define la misma proyección (11), pero tomando como base de $R^m$, los m factores comunes.

Supongamos ahora que tenemos una matriz de datos

| | | Variables | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | .... | p |
| Individuos | 1 | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1p}$ |
| | 2 | $y_{21}$ | $y_{22}$ | $\cdots$ | $y_{2p}$ |
| | ⋮ | ..... | | | |
| | i | $y_{i1}$ | $y_{y2}$ | $\cdots$ | $y_{ip}$ |
| | ⋮ | ..... | | | |
| | n | $y_{n1}$ | $y_{n2}$ | $\cdots$ | $y_{np}$ |

$= Y$ (13)

donde $y_{ij} = Y_j(i)$ es el valor del individuo y en la variable $Y_j$.

Podemos suponer que la matriz es centrada, es decir, que las sumas por columnas son 0. Para ello basta restar cada $y_{ij}$, $i \epsilon I$, la media $\bar{y}_j$ de la variable j. La matriz de covarianzas es entonces

$$S = \frac{1}{n} Y'Y \tag{14}$$

La descomposición espectral

$$S = VD_\lambda V'$$

define entonces una realización seguida de una proyección

$$I \xrightarrow{\ V\ } R^p \longrightarrow R^m \tag{16}$$

en la que las coordenadas de los individuos son

$$Z = YV \qquad (17)$$

Tomando las m primeras coordenadas representadas por las m primeras columnas de Z, tenemos una representación óptima en dimensión m de I. La distancia$^2$ euclídea original entre i, j se transforma en

$$\delta(i,j)^2 = \sum_{h=1}^{m} (y_{in}-y_{jn})^2 \rightarrow \sum_{h=1}^{m} (z_{in}-z_{jn})^2 = d_m^2(i,j)$$

y se demuestra /17/ que

$$\sum_{i<j} d_m^2(i,j) = n(\lambda_1 + \ldots + \lambda_m) \qquad (18)$$

donde $\lambda_1 \geq \ldots \geq \lambda_m$ son los primeros m valores propios de S.

## 3.2 ANÁLISIS DE COORDENADAS PRINCIPALES.

Supongamos que $(I,\delta)$ admite realización euclídea, de modo que B verifica (9). Propongamos aproximar B de rango m, por otra matriz $B^* \geq 0$ de rango $m' < m$ tal que (criterio de los mínimos cuadrados) poniendo $B^* = X_{(m)}X'_{(m)}$

$$f(X_{(m)}) = tra(B-B^*)^2 = \sum_{i,j} (b_{ij}-b^*_{ij})^2$$

sea mínimo. Derivando respecto $X_{(m)}$

$$\frac{\partial}{\partial X_{(m)}} f(X_{(m)} = -2(B-X_{(m)}X'_{(m)})X_{(m)} = 0$$

Si suponemos que los vectores columna de $X_{(m)}$ son ortogonales, entonces $X'_{(m)}X_{(m)} = D_m$ (matriz diagonal mxm), luego

$$BX_{(m)} = X_{(m)}D_m$$

y por lo tanto las columnas de $X_{(m)}$ constituyen los m primeros vectores propios de B. Pero éstas son las primeras m columnas de X de acuerdo con (9), que reciben el nombre de coordenadas principales en la representación euclídea de I (véase /33/, /65/). Obtenemos una realización de I seguida de una proyección análoga a la (16), verificando también (18), siendo la mejor representación euclídea de I en dimensión reducida.

El análisis canónico de poblaciones constituye una aplicación importante de esta técnica. Supongamos que I está formado por n poblaciones, y cada población se describe mediante $(\mu_i, \sum)$, donde $\mu_i$ es el vector de medias de las variables y $\sum$ la matriz de covarianzas común. Entonces se define entre cada par de poblaciones la distancia$^2$ de Mahalanobis /44/.

$$\delta^2_{ij} = (\mu_i-\mu_j)'\sum^{-}(\mu_i-\mu_j)$$

Calculando B y las coordenadas $X_{(m)}$, obtenemos una representación euclídea de las poblaciones.

Una generalización consiste en considerar un conjunto de funciones paramétricas

$$\{\psi_1, \psi_2, \ldots, \psi_q\}$$

siendo

$$\psi_i = a_{i1}\mu_1 + \ldots + a_{in}\mu_n$$

Definir entonces las distancias$^2$

$$\delta^2_{ij} = (\psi_i-\psi_j)'\sum^{-}(\psi_i-\psi_j)$$

y obtener análogamente las coordenadas principales. Entonces podemos representar funciones paramétricas estimables definidas sobre un modelo lineal multivariante (/12/, /15/, /18/), con aplicaciones al MANOVA --- (/16/, /51/).

## 3.3 REPRESENTACIONES SIMULTÁNEAS.

Las representaciones simultáneas en $R^m$ son aplicaciones

$$\begin{aligned} I &\longrightarrow R^m \\ E &\longrightarrow R^m \end{aligned} \qquad (19)$$

que permiten representar, bajo una métrica común, tanto los individuos como las variables.

Los diferentes métodos que realizan tal representación simultánea, pueden interpretarse como un caso particular del método BIPLOT /28/.

Toda matriz $Y_{nxp}$ (por ejemplo , la matriz de datos (13)), puede descomponerse en el producto

$$Y = GH' \qquad (20)$$

donde G es mxr, H es pxr, siendo r = ran(Y). Entonces, si $g_1',\ldots,g_n'$ son las filas de G, $h_1,\ldots,h_p$ son las columnas de H', (20) es equivalente al producto escalar

$$y_{ij} = g_i' h_j \qquad (21)$$

Luego podemos representar el individuo i por el vector $g_i$, la variable $y_j$ por el vector $h_j$, donde $g_i, h_j \in R^m$, para m=r.

La representación biplot toma la dimensión m=2. Se demuestra entonces que la mejor aproximación de la matriz Y de rango r por otra $Y_{(2)}$ de rango 2, tal que

$$tra(Y-Y_{(2)})^2$$

sea mínimo se obtiene a partir de la descomposición en valores singulares /31/ de Y

$$Y = U \Sigma V' \qquad (22)$$

donde $\Sigma$ es diagonal y contiene los valores singulares, U verifica $U'U = I_p$, V verifica $V'V = I_p$. La solución es

$$Y_{(2)} = U_{(2)} \Sigma_{(2)}^{\alpha} \Sigma_{(2)}^{1-\alpha} V_{(2)}' \qquad (23)$$

donde $U_{(2)}$, $V_{(2)}$ contienen las 2 primeras columnas de U, V, $\Sigma_{(2)}$ es diagonal 2x2 y contiene los 2 primeros valores singulares. Entonces tenemos dos tipos de soluciones:

a) $\alpha = 0 \quad G = U_{(2)} \quad H' = \Sigma_{(2)} V_{(2)}' \quad Y \simeq GH'$

llamada GH'-biplot.

b) $\alpha = 1 \quad J = U_{(2)} \Sigma_{(12)} \quad K = V_{(2)} \quad Y \simeq JK'$

llamada JK'-biplot.

El GH'-biplot da una buena representación de las variables, donde $\| h_j \|$ es proporcional a la desviación típica de la variable $Y_j$, $h_j' \cdot h_j$, es proporcional a la covarianza entre $Y_j$, $y_j$, . Por el contrario, el JK'-biplot da una buena representación de los individuos, en la que $\| j_i - j_{i'} \|$ aproxima la distancia de Mahalanobis entre i,i' (véase /28/).

También pueden estudiarse soluciones intermedias, por ejemplo tomando $\alpha = 1/2$.

Como ilustración del método biplot, representaremos la matriz Y de la tabla 1 (véase --/29/).

Los valores singulares son

$\sigma_1=108.01$, $\sigma_2=9.363$, $\sigma_3=0 \Rightarrow$ rang (Y) = 2.

Una descomposición en valores singulares de Y es

$$\begin{pmatrix} 0.292 & 0.580 & -0.015 \\ 0.450 & 0.008 & 0.877 \\ 0.450 & 0.175 & -0.288 \\ 0.563 & -0.717 & -0.222 \\ 0.438 & 0.344 & -0.311 \end{pmatrix}$$

$$\begin{pmatrix} 108.01 & 0 & 0 \\ 0 & 9.363 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0.181 & 0.961 & 0.205 \\ 0.840 & -0.260 & 0.477 \\ -0.512 & -0.085 & 0.854 \end{pmatrix}$$

$$Y = U \Sigma V'$$

De aquí se construye una descomposición $Y = AB'$, donde A es matriz 5x2, B' es matriz 2x3. Las filas de A representan las tribus, las filas de B representan las características demográficas, de modo que $y_{ij}=a_i' b_j$. La figura 1 da la representación simultánea. Obsérvese que $b_1$, $b_3$ aparecen muy relacionadas mientras $b_2$ presenta alta variabilidad. En cuanto a las tribus se observa que $a_4$ es la más diferenciada.

Es interesante conectar el biplot con el análisis de componentes principales. Supongamos que Y es una matriz centrada. Entonces la matriz de covarianzas es (14). Apliquemos el biplot a esta matriz (prescindiendo del factor escalar $n^{-1}$)

$$Y'.Y = VDV' \qquad (24)$$

siendo D matriz diagonal. Por (17), los individuos vienen representados por YV, pero según (22),

$$YV = U\Sigma V'V = U\Sigma \qquad (25)$$

Luego la representación de los individuos por análisis de componentes principales, coincide con el JK'-biplot sobre la matriz Y.

TABLA 1

CARACTERISTICAS DEMOGRAFICAS DE ALGUNAS TRIBUS DE INDIOS AMERICANOS

| Tribu | Mediana de los años de escolaridad | Porcentaje debajo límite pobreza | Indice Económico |
|---|---|---|---|
| Shoshones | 10.3 | 29.0 | 9.08 |
| Apaches | 8.9 | 46.8 | 10.02 |
| Sioux | 10.2 | 46.3 | 10.75 |
| Navajos | 5.4 | 60.2 | 9.26 |
| Hopis | 11.3 | 44.7 | 11.25 |

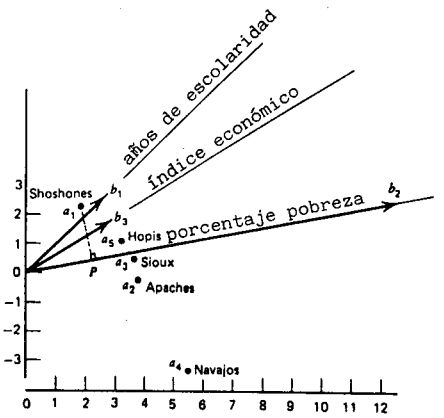


Fig. 1 : Biplot de la matriz de datos de la Tabla 1.

El análisis factorial de correspondencias /4/ se relaciona también con el biplot. Se trata de un método que, aplicado a matrices F de datos no negativos (usualmente frecuencias, porcentajes, medidas de abundancias), permite representar simultáneamente individuos y variables, donde los individuos aparecen como una media ponderada de las variables y recíprocamente. Supongamos que R y C son las matrices diagonales con los totales por filas y por columnas. Se utiliza entonces la llamada distancia ji-cuadrado, que es de hecho una distancia euclídea sobre la matriz

$$Y = R^{-1/2} F C^{-1/2} \qquad (26)$$

Consideremos la descomposición en valores singulares de Y

$$Y = \tilde{T} \sum T' \qquad (27)$$

Entonces la representación de los individuos o filas de F es (/17/, pág. 327)

$$A = XT = R^{-1/2}\tilde{T} \sum T'T = R^{-1/2}\tilde{T} \sum$$

Análogamente, la representación de las variables viene dada por

$$B = \tilde{X}'\tilde{T} = C^{-1/2}T \sum \tilde{T}'\tilde{T} = C^{-1/2}T \sum$$

Una clase de soluciones general es /35/.

$$A = R^{-1/2}\tilde{T} \sum^{\alpha} \qquad B = C^{-1/2}T \sum^{\alpha} \qquad (28)$$

Dando diferentes valores a $\alpha$, por ejemplo, $\alpha = 0$, $1/2$ ó $1$, se construyen otras tantas soluciones. Obsérvese que para $\alpha = 1/2$ se verifica

$$Y = AB' \qquad (29)$$

luego obtenemos un AB'-biplot de la matriz Y /30/.

Finalmente, /18/, /33/, /41/, en el llamado análisis de datos centrados se obtiene, de forma natural, una representación simultánea entre variables e individuos. Entonces, tanto la matriz Y'.Y como la matriz Y.Y' representan matrices de dispersión entre variables y entre individuos respectivamente, siendo fácil probar que dada la descomposición en valores singulares $Y = U\sum V'$, la representación de individuos es $A = YV$, la representación de variables es $B = Y'U$, estando A y B. relacionados por

$$A = YB\sum^{-} \qquad B = Y'A\sum^{-} \qquad (30)$$

Así, las coordenadas no triviales de B son

$$b_{ij} = \sigma_j^{-1}(y_{1i}\, a_{1j} + \ldots + y_{ni}\, a_{nj}) \qquad (31)$$

y por tanto (salvo el factor $\sigma_j^{-1}$) son media aritmética de las j coordenadas de los individuos, ponderadas por los valores que la variable $Y_i$ alcanza sobre los n individuos. En el caso del análisis de correspondencias se demuestra que (31), en la versión (28) con $\alpha = 1$, es la mejor representación $\beta$-baricéntrica, es decir, la mejor representación simultánea de filas y columnas de una matriz F, salvo un factor de proporcionalidad /40/.

## 3.4. ANALISIS INDIVIDUAL DE PROXIMIDADES.

Supongamos ahora que tenemos q sujetos y un conjunto I de objetos, de modo que cada sujeto define una métrica en I. Obtendremos entonces q espacios métricos.

$$(I,\delta_1),\ (I,\delta_2),\ \ldots,(I,\delta_q)$$

El análisis individual de proximidades ("individual multidimensional scaling") postula que existen dos representaciones euclídeas de I:

a) La general, formada por una configuración euclídea en $R^m$ representada por una matriz de coordenadas X.

b) La correspondiente a cada sujeto $s_j$, definida por una aplicación

$$\phi_j(X) = X_j \qquad j = 1,\ldots,q$$

que a los objetos I hace corresponder las coordenadas $X_j$.

Se interpreta que el sujeto $s_j$ deforma la configuración general X llegando a la configuración $X_j$. Los diferentes modelos que se han estudiado suponen la transformación lineal

$$\phi_j(X) = XT_j$$

donde $T_1,\ldots,T_q$ son matrices mxm. El nombre del modelo o método depende de las condiciones que verifican las $T_j$. Tenemos así:

1) $T_j=W_j$ (matriz diagonal). INDSCAL /10/.

2) $T_j=W_j$ S (S no restringida, $W_j$ diagonal) PARAFAC /36/.

3) $T_j$ matriz no restringida. IDIOSCAL /10/, PINDIS /69/.

El más utilizado es el INDSCAL, en el que se recoge en una matriz W los valores diagonales de las matrices $W_1,\ldots,W_q$. Entonces la representación de W informa sobre el peso que cada sujeto da a las diferentes dimensiones.

Existen dos teoremas que caracterizan la realización euclídea de los métodos INDSCAL e IDIOSCAL /25/. Indiquemos por $B_j=HA_jH$ la matriz asociada a $\delta_j$, como en el teorema 2.

Teorema 3. Los espacios $(I,\delta_j)$, $j=1,\ldots,q$, admiten una realización euclídea en $R^m$ por el método IDIOSCAL si y sólo si

a) $B_j \geq 0 \quad j=1,\ldots,q$

b) $\text{rang}\left(\sum_j B_j\right) = m$

Teorema 4. Los espacios $(I,\delta_j)$, $j=1,\ldots,q$, admiten una realización euclídea en $R^m$ por el método INDSCAL si y sólo si

a) $B_j \geq 0 \quad j=1,\ldots,q$

b) $\text{rang}\,(B_*) = m$, siendo $B_* = \sum_j B_j$

c) $B_i B_*^- B_j = B_j B_*^- B_i$ para todo $i,j=1,\ldots,q$.

## 4. GEOMETRIA EUCLIDEA.

Si la matriz B del teorema 2 tiene algún valor propio negativo, no existe una realización de I sobre $R^m$. Entonces diremos que la geometría sobre I no es euclídea, o que la distancia $\delta$ no es euclídea. Para representar $(I,\delta)$ en $R^m$ debemos recurrir entonces a una realización monótona, es decir, encontrar una función monótona no decreciente que transforme $\delta$ en una distancia euclídea. Podemos hablar de dos tipos de transformaciones: algebraicas y numéricas.

### 4.1. TRANSFORMACIONES ALGEBRAICAS.

La transformación más simple consiste en añadir una constante c a $\delta_{ij}$. Observemos que cualquiera que sea $\delta_{ij}$, si tomamos

$$\hat{c} = \max_{i,j,k} \{ \delta_{kj} - \delta_{ki} - \delta_{ij} \}$$

entonces $\delta_{ij} + \hat{c}$ verificará la desigualdad triangular. En general, el problema de hallar el mínimo $c^*$ tal que la distancia $\delta_{ij}(c)$

$$\delta_{ij}(c) = \delta_{ij} + c(1-\delta^{ij}) \qquad (32)$$

(donde $\delta^{ij}$ indica la delta de Kronecker) tenga realización euclídea, se conoce como "problema de la constante aditiva" /11/.

De (32) deducimos

$$-\tfrac{1}{2}\,\delta_{ij}{}^2(c) = -\tfrac{1}{2}\,\delta_{ij}{}^2 + c\delta_{ij} + c^2(1-\delta^{ij})$$

y la matriz B(c) asociada a $\delta_{ij}(c)$ (teorema 2) es

$$B(c) = B(o) + 2c\tilde{B} + \frac{c^2}{2}H$$

donde $\tilde{B}$ es la matriz asociada a la "distancia" $\sqrt{\delta_{ij}}$.

Se verifica entonces /25/.

Teorema 5. Existe una constante c tal que $\delta_{ij}(c)$ admite realización euclídea en dimensión (n-1).

Teorema 6. Supongamos que $\delta_{ij}$ no admite realización euclídea. Existe entonces una constante c tal que $\delta_{ij}(c)$ admite realización euclídea en dimensión (n-2).

El teorema 6 nos dice que si $(I,\delta)$ no admite realización euclídea, entonces admite realización monótona euclídea en $R^{n-2}$.

El problema de encontrar la mínima constante $c^*$ tal que $\delta_{ij}(c^*)$ sea euclídea, es decir, $\delta_{ij}(c)$ no es euclídea si $c<c^*$, mientras que $\delta_{ij}(c)$ es euclídea si $c\geq c^*$, ha sido resuelto recientemente /8/. Se demuestra que la constante $c^*$ es el mayor valor propio de la matriz

$$A = \begin{pmatrix} 0 & 2B(o) \\ -I & -4\tilde{B} \end{pmatrix}$$

Estudiemos ahora la llamada transformación aditiva en $\delta_{ij}^2$

$$\delta_{ij}^2(a) = \delta_{ij}^2 - 2a(1 - \delta^{ij}) \qquad (33)$$

La matriz asociada a $\delta_{ij}(a)$ es

$$B(a) = B - aH$$

Se demuestra que $B(a) \geq 0$, $rang(B(a)) = n-2$ para alguna a /43/.

La llamada solución de Lingoes consiste en tomar $a = \lambda_n$, siendo $\lambda_n$ el mínimo valor propio de B(0) (que es negativo). Entonces

$$\delta_{ij}(\lambda_n) = (\delta_{ij}^2 - 2\lambda_n(1-\delta^{ij}))^{1/2} \qquad (34)$$

representa la solución óptima para la clase de distancias transformadas (33).

Mardia /45/ observa que, aunque (34) proporciona una realización euclídea exacta en dimensión (n-2), $\delta_{ij}(\lambda_n)$ puede representar una distorsión considerable respecto a $\delta_{ij}$. Propone entonces tomar la constante

$$\hat{a} = \left(\sum_{i=m+1}^{n-1} \lambda_i\right)/(n-m-1) \qquad (35)$$

media aritmética de los n-m-1 menores valores propios de B(o), siendo m la dimensión de la representación euclídea. Para ciertas medidas de distorsión entre $\delta_{ij}$ y $\delta_{ij}(a)$, la solución (35) aplicada a (33) es óptima, aunque la representación euclídea no sea exacta.

Seguidamente estudiemos la transformación lineal

$$\delta_{ij}(a,b) = a\delta_{ij} + b(1-\delta^{ij}) \qquad (36)$$

que tiene la matriz asociada.

$$B(a,b) = a^2B(0) + 2ab\tilde{B} \quad \frac{b^2}{2} H$$

donde B(0) y $\tilde{B}$ son las matrices asociadas a $\delta_{ij}$ y $\sqrt{\delta_{ij}}$. La solución exacta para la clase de transformaciones lineales (36) es análoga a la del problema de la constante aditiva, así que es alcanzada también en dimensión (n-2). En /23/ se propone un algoritmo iterativo que permite obtener una solución aproximada para una dimensión euclídea m dada. El algoritmo puede generalizarse a una transformación polinómica en $\delta_{ij}$ de grado k.

$$\hat{\delta}_{ij} = P_k(\delta_{ij}) \qquad i \neq j.$$

pues el proceso descrito en /23/ es el mismo para el caso no lineal.

Como una consecuencia de los tipos de soluciones algebraicas comentadas, podemos enunciar

Teorema 7. Supongamos que $(I,\delta)$ no admite realización euclídea. Entonces $(I,\delta)$ admite una realización monótona en $(R^{n-2},d)$.

Es importante observar que no se ha podido encontrar hasta el presente una solución analítica exacta al problema de hallar f tal que $\hat{\delta}_{ij} = f(\delta_{ij})$ sea euclídea, mientras B se transforma en $\hat{B}$, con la propiedad de que $rang(\hat{B})$ = mínimo. En otras palabras: determinar la mínima dimensión euclídea m para la cual $(I,\delta)$ admite realización monótona en $(R^m,d)$.

## 4.2 TRANSFORMACIONES NUMERICAS.

Entendemos por transformación numérica a una aplicación f

$$\hat{d}_{ij} = f(\delta_{ij}) \qquad (37)$$

que transforma $\delta_{ij}$ en una distancia euclídea $\hat{d}_{ij}$, donde f puede tener cualquier expresión no necesariamente algebraica ni pertenecer a ninguna familia determinada de funciones. -- Shepard (/60/, /61/) propuso la idea de que las distancias euclídeas deben conservar la preordenación asociada a $(I,\delta)$, es decir

$$\delta_{ij} \leq \delta_{i'j'} \Rightarrow \hat{d}_{ij} \leq \hat{d}_{i'j'}, \qquad \forall\ i,j,i',j' \qquad (38)$$

La condición (38) significa que tanto $(I,\delta)$ como su imagen euclídea $(R^m,d)$ deben tener asociadas la misma preordenación /17/. En otras palabras: los elementos próximos (alejados) en I deben seguir estando próximos (alejados) en $R^m$.

Como en las aplicaciones debemos representar en dimensión reducida, deberemos hallar una realización monótona euclídea aproximada en

el sentido (4), de modo que la preordenación se conserve aproximadamente.

El algoritmo propuesto por Shepard /60/ consistía en ajustar directamente una distancia euclídea $d_{ij}$ a una transformación monótona $f(\delta_{ij})$. Sin embargo, se ha impuesto el método de Kruskal /67/, /68/, que ajusta una distancia euclídea $d_{ij}$ a la transformación $\hat{d}_{ij} = f(d_{ij})$, donde f se debe buscar entre la clase no paramétrica de todas las funciones monótonas, junto con una configuración euclídea X para una dimensión dada m, que define una distancia euclídea $d_{ij}$, tal que la medida de distorsión

$$\Lambda\alpha = \left( \frac{\sum_{i<j} |d_{ij} - \hat{d}_{ij}|^{1/\alpha}}{\sum_{i<j} d_{ij}^{1/\alpha}} \right)^{\alpha} \qquad (39)$$

sea mínima. Cuando $\alpha = 1/2$ esta medida se llama STRESS se indica por S y se suele expresar en forma de porcentaje. Para una m dada, una realización monótona de I en $R^m$ se califica de buena o excelente si $S \leq 5\%$. Si $S = 0$ tenemos una realización monótona exacta de $(I,\delta)$ en $(R^m,d)$. Obsérvese que minimizar S es un problema de regresión monótona: ajuste por mínimos cuadrados conservando la relación de monotomía entre las distancias.

Leeuw y Heiser /25/ exponen una relación sistemática de los algoritmos de MDS ("Multidimensional Scaling"), estudiando sus propiedades. En todos los casos se trata de minimizar una función por el criterio de los mínimos cuadrados, haciendo intervenir la configuración euclídea X, $\delta_{ij}$ y la distancia euclídea $d_{ij}(X)$.

a. Mínimos cuadrados sobre la matriz B (teorema 2).

a.1 : MDS métrico.
La función a minimizar es
$$L_1(X) = \text{tra}(B - X.X')^2 \qquad (40)$$
Si la distancia $\delta_{ij}$ es euclídea, es el mismo criterio que da lugar al análisis de coordenadas principales.

a.2 : MDS no métrico.
En el problema de la constante aditiva la función es
$$L_1(X,c) = \text{tra}(B(c) - X.X')^2 \qquad (41)$$

b. Mínimos cuadrados sobre las distancias al cuadrado.
La función de ajuste es
$$L_2(X) = \sum_{i<j} (\delta_{ij}^2 - d_{ij}^2(X))^2 \qquad (42)$$
y recibe el nombre de SSTRESS /64/.

c. Mínimos cuadrados sobre las distancias.
La función de ajuste es
$$L_3(X) = \sum_{i<j} w_{ij}(\delta_{ij} - d_{ij}(X))^2 \qquad (43)$$
donde $w_{ij}$ es un peso o carga. Si $w_{ij} \doteq 1$, minimizar (43) equivale a minimizar (39), para $\alpha = 1/2$, es decir, el STRESS de Kruskal. La introducción de los pesos $w_{ij}$ se hace para obtener distribuciones ji-cuadrado para (43), y para realizar comparaciones entre el STRESS y el SSTRESS.

Finalmente, digamos que el problema de encontrar una realización monótona de $(I,\delta)$ en $(R^m, d)$ está numéricamente resuelto para determinadas funciones de ajuste, aunque la realización exacta en dimensión mínima es un problema todavía no resuelto. Incluso, en algunos casos el problema no tiene solución. Lew /42/ comprueba dos contraejemplos:

1) $(I,\delta)$ no tiene realización monótona euclídea para $I = L_1^n$ con $n \geq 2$, es decir, los elementos de I son n-tuplas $(x_1,\ldots,x_n) = x$ con la métrica definida por la norma "ciudad"
$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

2) $(I,\delta)$ no tiene realización monótona euclídea para $I = L_\infty^n$, con $n \geq 2$, donde ahora la métrica viene definida por la norma "dominante"
$$\|x\|_\infty = \sup\{|x_1|,\ldots,|x_n|\}$$

Sin embargo, Lew /42/ demuestra que en ambos casos existe realización monótona sobre un espacio de Hilbert separable H con la norma cuadrática $\|x_2\|$, aunque también encuentra un contraejemplo de conjunto con una métrica que tampoco tiene realización monótona en H.

En el apartado 7 comentaremos realizaciones en variedades de Riemann.

## 5. GEOMETRIA ULTRAMETRICA.

Diremos que $(I,\delta)$ es un espacio ultramétrico si la distancia $\delta_{ij}$ verifica además la llamada desigualdad ultramétrica

$$\text{iv)} \quad \delta_{ij} \leq \max\{\delta_{ik}, \delta_{jk}\} \qquad \forall\, i,j,k \in I \qquad (44)$$

Es inmediato que se verifica entonces la desigualdad triangular. Una propiedad fundamental de una distancia ultramétrica es

Teorema 8. En un espacio ultramétrico todo triángulo $\{i,j,k\}$ es isósceles, siendo la base el lado menor

Es sabido /70/ que puede asociarse a I una jerarquía indexada (C, H), donde C es una colección de subconjuntos ("clusters") de I, y h es un índice sobre C, univocamente determinado por $\delta_{ij}$, con ciertas propiedades de monotonía . La idea principal es que en un espacio ultramétrico, la noción de proximidad $\delta_{ij} \leq x$ define una partición ("clustering") de I, y que aumentando x se obtienen particiones menos finas, que engloban a las anteriores, formando una estructura jerárquica. Se verifica /17/:

Teorema 9. Sea $(I,\delta)$ un espacio ultramétrico. Entonces:

a) I tiene asociado una jerarquía indexada (C, h). Reciprocamente, toda jerarquía indexada define una distancia ultramétrica sobre I.

b) La relación binaria en I

$$iR_x j \iff \delta_{ij} \leq x$$

es de equivalencia para todo real $x \geq 0$.

La representación geométrica de (C, h) se realiza a través de un grafo orientado llamado dendograma, que puede ser representado íntegramente en $R^2$. Las imágenes de los elementos de I son los extremos que no son raices del grafo. Entonces existe una realización

$$(I,\delta) \longrightarrow (G,u) \qquad (45)$$

siendo G un conjunto de grafos conexos sin ciclos, u la distancia que resulta de conectar cada par de extremos. Por ejemplo, en el dendograma de la figura 2, las distancias para el triángulo $\{1,2,5\}$ son

$$u_{12} = 1 < u_{15} = u_{25} = 4$$

Obsérvese que cualquier triángulo $\{i,j,k\}$ verifica (44).

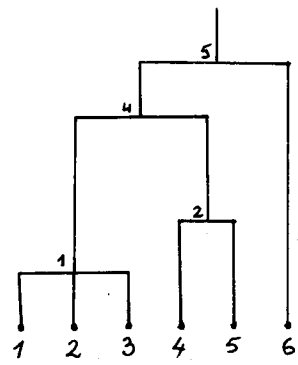


Figura 2: Ejemplo de dendograma.

Otra propiedad que tiene consecuencias interesantes es que $\delta_{ij}$ es distancia euclídea.

Teorema 10 Si $\delta$ es ultramétrica, $|I| = n$, entonces $(I,\delta)$ admite una realización euclídea en dimensión (n-1).

(Diferentes demostraciones en /9/, /20/, /34/, /39/).

Además de la realización (45), existe pues otra realización

$$(I,\delta) \longrightarrow (R^{n-1},d) \qquad (46)$$

que se describe a través de la descomposición $B = X.X'$ (teorema 2). Los valores propios de B y las coordenadas euclídeas X reflejan aspectos interesantes de la estructura jerárquica (C, h).

Agrupando elementos equidistantes de I tenemos

$$I = I_1 + I_2 + \ldots + I_k \qquad I_i \cap I_j = \emptyset \quad i \neq j \qquad (47)$$

siendo

$$I_1=[i_1,\ldots,i_{n_1}\,|\,\delta_{i_t i_{t'}}=h_1, h_1=\min\{\delta_{ij},\ i\neq j\in I\}]$$

$$I_2=[i_{n_1+n_2},\ldots,i_{n_1+n_2}\,|\,\delta_{i_t i_{t'}}=h_2, h_2=\min\{\delta_{ij}, i\neq j\in I-I_1\}]$$

... etcétera.

Suponemos además que son iguales los elementos de las filas de la matriz de distancias $(\delta_{ij})$ que resultan de eliminar los elementos $\delta_{ij}=h_q$, para $i,j\in Iq$, es decir,

$$\delta_{ij}=\delta_{i'j} \quad \text{si} \quad i,i'\in I_q \qquad j\notin I_q \qquad (48)$$

Resulta entonces que

$$|I_q| = n_q > 1$$

si (48) se cumple, y en caso contrario

$$|I_q| = 1$$

Expresaremos entonces la reunión (47) así

$$I = I_1+\ldots+I_r+I_{r+1}+\ldots+I_k$$

$$|I_j| = n_j>1 \quad 1\leq j\leq r \quad n_1\geq\ldots\geq n_r \qquad (49)$$

$$|I_j| = 1 \qquad j>r$$

De modo que (49) expresa la reunión de r -- clases o "clusters" con (k-r) elementos aislados de I. Por ejemplo, de la figura 2, tenemos

$$I = \{1,2,3\}+\{4,5\}+\{6\}$$

Ohsumi y Nakamura /46/ demuestran (con la ilustración de diversos ejemplos), que la formación de "clusters" sobre (C, h) se puede expresar como una combinación lineal de las distancias (al cuadrado) $h_1^2,\ldots,h_r^2$ que se relacionan directamente con algunos valores propios de B.

Otro resultado es el siguiente /22/:

Teorema 11. Sea $\mu$ el mayor valor propio de B. Sea $h_j=\delta_{ii'}$, $i,i'\in I_j$, $1\leq j\leq r$ y consideremos

$$\lambda_j = \frac{1}{2}h_j^2 \qquad 1\leq j\leq r$$

Se verifica:

a) Cada $\lambda_j$ es valor propio de B de multiplicidad $(n_j-1)$. Además:

$$\mu\geq\lambda\geq\ldots\geq\lambda_1$$

siendo $\lambda_1 = \min\{\frac{1}{2}\delta_{ij}^2, i\neq j\in I\}$ el menor valor propio de B.

b) La igualdad $\mu=\lambda_j$ se cumple si y sólo si todos los elementos de I son equidistantes (I es isométrico a un simplex regular en dimensión n-1).

Sea ahora B = X.X'. Indíquenos

$$X = (X_o, X_r,\ldots,X_1)$$

donde $X_q$ son las coordenadas euclídeas que definen los $(n_q-1)$ vectores propios de valor propio $\lambda_q$. Por $X_o$ indicamos las restantes -- coordenadas principales euclídeas. La distancia euclídea $d_j$ definida a través de las --- coordenadas $X_j$, define una reducción de la dimensión

$$(I,\delta)\longrightarrow(R^{n-1},d)\longrightarrow(R^{n_j-1},d_j) \qquad (50)$$

siendo $n_j=|I_j|$, $1\leq j\leq r$, $n_o=n-1-\sum_{j=1}^{r}(n_j-1)=m-1$. Se tiene (/19/, /22/):

Teorema 12: Las distancias euclídeas que resultan de la proyección (50) verifican:

a) $d_o(i,j)=0$ si $i,j\in I_q$ $\quad 1\leq q\leq k$

$d_o(i,j)>0$ si $i\in Iq$ $j\in I_{q'}$ $\quad 1\leq q'\neq q''\leq k$

donde k es el número de clases construidas en (49).

b) $d_q(i,j) = hq$ si $i\neq j\in I_q$ $\quad 1\leq q\leq r$

$d_q(i,j) = 0$ si $i,j\notin I_q$ $\quad 1\leq q\leq k$

La interpretación del teorema 12 es clara. A través de las coordenadas $X_o$ representamos las clases $I_1,\ldots,I_q$, que quedan diferenciadas entre sí, y a través de cada matriz de coordenadas $X_q$ $(1\leq q\leq r)$, quedan representados exclusivamente los elementos de Iq (figura 3).

Holman /39/ observa que mientras una ultramétrica puede ser representada en un plano a través de un dendograma, la dimensión exacta en representación euclídea es n-1, luego no

se adapta bien a una reducción de la dimensión. Lo que ocurre en realidad es que deben realizarse varias representaciones euclídeas, que según el teorema 12 tienen diferentes interpretaciones y significados.

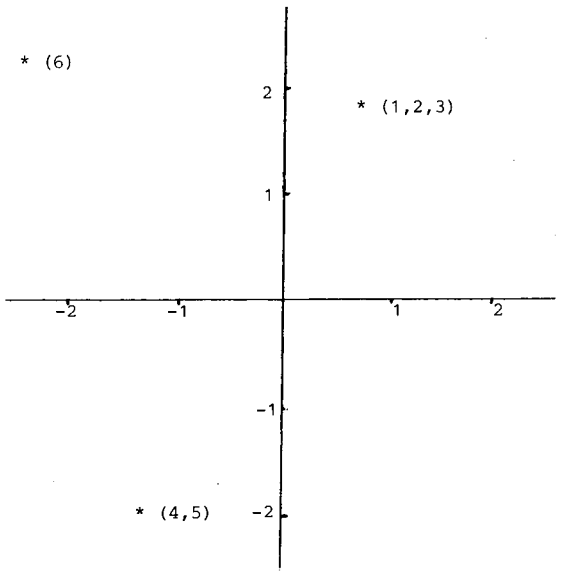


Figura 3: Pepresentación euclídea bidimensional del dendograma de la figura 2

## 6. GEOMETRIA ADITIVA.

Sea $(I,\delta)$ un espacio métrico y supongamos que $\delta$, además de las propiedades i), ii), iii) verifica

$$v)\ \delta_{ij}+\delta_{km} \leq \max\{\delta_{ik}+\delta_{jm}, \delta_{im}+\delta_{jk}\} \qquad (51)$$

para toda cuaterna (i,j,k,m). (51) se conoce como desigualdad aditiva o como axioma de los cuatro puntos. Diremos que $(I,\delta)$ es un espacio aditivo.

Supongamos que $\delta_{ij}+\delta_{km}$ es el menor de los sumandos en (51). Entonces es fácil verificar que

$$\delta_{ij}+\delta_{km} \leq \delta_{ik}+\delta_{jm} = \delta_{im}+\delta_{jk} \qquad (52)$$

Se puede dar a (52) una interpretación en la misma línea del teorema 8. Se verifica:

Teorema 13. En un espacio aditivo todo tetraedro (i,j,k,m) tiene dos pares de aristas opuestas cuyas sumas de longitudes coinciden, y además son ≥ que la suma del restante par.

Se comprueba también que la desigualdad ultramétrica es más restrictiva que la aditiva, la cual implica también la desigualdad triangular, es decir :

Desig. ultrametrica ⇒ desig. aditiva ⇒ desig.triangular

Así pues, un espacio ultramétrico es un caso particular de un espacio aditivo.

Una distancia aditiva viene a ser la distancia natural en un grafo simplemente conexo, con n extremos. La representación de un espacio aditivo la conseguiremos a través de una realización

$$(I,\delta) \longrightarrow (G,d)$$

donde G es un conjunto de grafos simplemente conexos con n extremos, llamados árboles aditivos, con una métrica en la que la distancia entre dos extremos (que están conectados por un único camino) es la longitud d del camino que los une. La imagen de cada uno de los individuos de I es un extremo del árbol aditivo. Por ejemplo, dada la matriz de distancias sobre I = {1,2,3,4,5}.

$$\Delta = \begin{pmatrix} 0 & 2.1 & 2.6 & 6 & 3.2 \\ & 0 & 0.7 & 6.1 & 3.3 \\ & & 0 & 6.6 & 3.8 \\ & & & 0 & 3.2 \\ & & & & 0 \end{pmatrix}$$

se comprueba fácilmente la desigualdad aditiva. Entonces, una posible representación de $(I,\delta)$ sería el árbol aditivo de la figura 4.

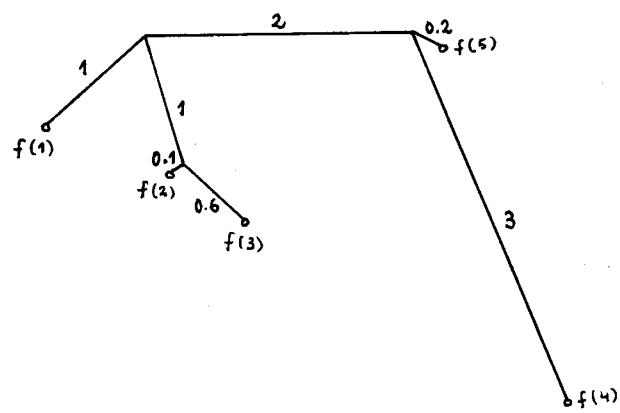


Figura 4: Ejemplo de árbol aditivo.

Un dendograma es un caso particular de árbol aditivo. Exactamente es un árbol aditivo con un nodo distinguido (llamado raíz) que es equidistante de todos los extremos. La geometría aditiva nace al considerar el axioma ultramétrico demasiado restrictivo, sustituirlo entonces por la desigualdad aditiva, más flexible, y que permite adaptarse mejor a una distancia experimental $\delta$. La representación, en lugar de un dendograma, se hace a través de un árbol aditivo, usualmente representado en forma paralela (ver figura 5). Las diferencias entre ambos tipos de representaciones son:

1) En el caso ultramétrico, las $n(n-1)/2$ interdistancias entre los individuos vienen determinadas por al menos $n-1$ valores independientes. En el caso aditivo son necesarios al menos $2n-3$ valores independientes.

2) Una ultramétrica define una jerarquía indexada $(C,h)$. La distancia entre individuos del mismo cluster (distancia intra-cluster) es siempre menor que la distancia entre individuos de distinto cluster (distancia inter-cluster).

3) Una distancia aditiva no ultramétrica no define ninguna jerarquía indexada. La distancia intra-cluster puede superar a la distancia inter-cluster.

4) Supongamos que mediante sendos algoritmos adecuados, ajustamos una distancia ultramétrica u y una aditiva d a una distancia experimental $\delta$. Entonces la distorsión (medida mediante correlación cofenética, stress, etc.) es menor para d que para u.

Los siguientes teoremas exponen tres resultados teóricos de notable interés e importancia.

Teorema 14. El espacio $(I,\delta)$ se puede representar a través de un árbol aditivo si y sólo si $\delta$ verifica la desigualdad aditiva /5/.

Teorema 15. La representación de $(I,\delta)$ es única /66/.

Teorema 16. Si $\delta$ es una distancia aditiva, existe entonces una distancia ultramétrica u y una función $\psi : I \longrightarrow R$ tal que

$$\delta_{ij} = u_{ij} + \psi(i) + \psi(j) \qquad (53)$$

(véase /71/).

La expresión (53) permite estudiar y clasificar las distancias aditivas. Destacan entonces tres tipos simples de árboles aditivos:

a) Ultramétricos: $\delta_{ij} = u_{ij}$

b) Singulares: $\delta_{ij} = \psi(i) + \psi(j)$

   En este caso, el árbol aditivo tiene un único modo interno.

c) Lineales: Si todos los puntos pueden representarse a lo largo de una línea recta.

Otras clases de árboles aditivos se pueden construir combinando estos tres tipos. Así se demuestra que si $\psi$ es positiva, el árbol aditivo es suma de un ultramétrico más un singular y esto ocurre si ninguna distancia entre nodos internos excede a las distancias entre extremos. Finalmente, la condición para que un árbol aditivo sea suma de uno singular más otro lineal es que a lo sumo hayan dos aristas internas partiendo de un nodo interno /71/.

Existen diversos algoritmos de construcción de árboles aditivos a partir de una distancia /24/, /26/, /71/.

Uno de los aspectos interesantes a plantear es la conveniencia de realizar una representación euclídea o una representación mediante un árbol aditivo. Diversos autores, /55/, /71/, señalan que cuando los individuos se pueden describir en términos de alguna estructura factorial, es preferible la representación espacial. Pero si los datos reflejan un esquema de clasificación o una estructura evolutiva, la representación debe hacerse mediante un dendograma o un árbol aditivo. Pruzansky et.al. /55/, como medidas objetivas, señala que el sesgo (momento central de tercer orden de la distribución de distancias) y la elongación (proporción de triángulos tales que la suma de las 2 caras extremas es menor que el doble de la intermedia)

pueden describir aproximadamente la conveniencia de uno u otro tipo de representación. Mientras la representación en árbol tiende a producir sesgo negativo y elongación alta, la representación espacial (particularmente en el plano), produce sesgo positivo y elongación baja (resultados que obtienen por simulación). De todos modos, es un problema abierto el estudio espacial de los árboles aditivos y de los criterios para decidir entre uno u otro tipo de representación, /1/, /2/.

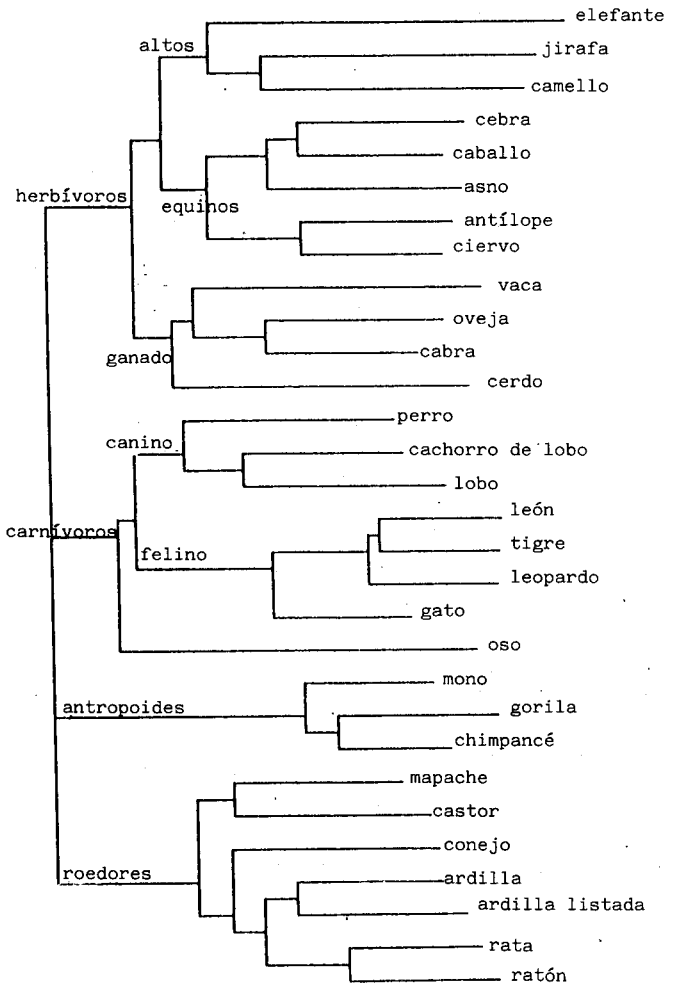


Figura 5: Representación de diversas especies de animales a través de un árbol aditivo en forma paralela (según /71/).

## 7. GEOMETRIA RIEMANNIANA.

Una realización de $(I,\delta)$ en un espacio euclídeo se puede generalizar a una variedad riemanniana

$$(I,\delta) \longrightarrow (M,d_M)$$

donde $d_M$ es una distancia definida a través de una métrica de Riemann. Si además la curvatura de Riemann es constante, podemos representar los elementos de I como puntos de un espacio hiperbólico, elíptico o esférico, según el signo de la curvatura. Si la curvatura es nula, entonces tenemos que I admite una realización euclídea. También ha sido estudiado e interpretado el caso de curvatura no constante /75/.

En las aplicaciones, $d_M$ se ajusta a $\delta$ mediante algún criterio de mínimos cuadrados /74/. Como $(M, d_M)$ tiene una estructura más amplia que la de un espacio euclídeo, cabe esperar una realización más aproximada de $(I,\delta)$ sobre una variedad riemanniana que sobre un espacio euclídeo.

Un caso importante en el cual la representación riemanniana es natural, se presenta --- cuando I puede ser identificado con una distribución de probabilidad paramétrica

$$p(x_1,\ldots,x_k;\ \theta^1,\ldots,\theta^n)$$

Consideramos entonces la variedad paramétrica

$$M = \{\theta \epsilon R^n \mid p(x,\theta) \text{ es f. densidad de } x \epsilon R^k\}$$

y a continuación tomamos como tensor métrico fundamental (Rao /73/) la matriz de información de Fisher

$$g_{ij} = E\left(\frac{1}{p^2}\ \frac{\partial p}{\partial \theta^i}\ \frac{\partial p}{\partial \theta^j}\right) \qquad i,j = 1,\ldots,m$$

La distancia entre dos distribuciones es, en general, la longitud de la geodésica que las une, interpretadas ambas como puntos de M.

En las secciones siguientes nos proponemos llegar a esta distancia partiendo de ciertos argumentos heurísticos y de cuatro condiciones deseables que deberían verificar una distancia, que a continuación concretaremos para algunas distribuciones clásicas.

## 7. CARACTERIZACION Y CONDICIONES GENERALES PARA UNA DISTANCIA.

Cuando se utiliza el concepto de distancia en Análisis de Datos es para expresar de - forma cuantitativa las analogías y diferencias entre los objetos que se distancian, de forma que objetos geométricamente próximos sea sinónimo de objetos semejantes y a la inversa, objetos geométricamente alejados debe corresponderse con la idea de objetos muy diferentes.

En otras palabras, la distancia entre dos objetos debe ser considerada como una medida de la información que poseemos sobre las analogías y diferencias de los mismos. Esta información, más que una propiedad de -- los objetos en sí que comparamos, es una -- propiedad del proceso de observación, y para su cuantificación, deberemos basarnos en las propiedades formales del mismo, atendiendo a la naturaleza probabilística de las variables que observamos y utilizamos para estudiar dichos objetos, antes que a la naturaleza física de los mismos.

El primer problema que hay que resolver es el problema de la caracterización de los objetos estudiados, a través de objetos formales adecuados, susceptibles de ser tratados en términos matemáticos. Con frecuencia efectuamos ciertas medidas sobre los objetos estudiados, medidas que pueden ser consideradas como valores de una muestra aleatoria de una población estadística, resultando entonces útil caracterizar a cada uno de los mismos por medio de la función de densidad de probabilidad conjunta de las variables aleatorias observadas sobre cada objeto. En muchos casos, según la naturaleza del problema a resolver, podremos suponer que dicha función de densidad pertenece a una determinada familia paramétrica: multinomial, normal, exponencial, etc. ...

Caracterizados los objetos a distanciar mediante funciones de densidad de probabilidad paramétrica $p(x,\theta)$, deberemos introducir una distancia entre éstas últimas, que introducirá de forma natural una métrica o pseudométrica en el conjunto de objetos estudiados.

Será conveniente exigir que la distancia definida entre las funciones de densidad paramétricas verifique las siguientes condiciones:

1.-La distancia entre dos funciones de densidad debe depender del concepto de información. A su vez, la información asociada a un resultado x, podemos definirla como $f(p(x,\theta))$, donde f es una función cuyo dominio son los reales positivos y que podemos suponer diferenciable las veces que haga falta.

2. La distancia entre dos funciones de densidad debe ser independiente de la parametrización de las mismas, es decir, debe de resultar invariante frente a transformaciones admisibles de los parametros, - entendiendo por tales, transformaciones - biyectivas definidas a través de funciones diferenciables con continuidad. Ello debe ser así debido a que los parámetros no son más que un "sistema de coordenadas"

en una variedad de funciones de densidad de probabilidad.

3. Por otra parte, la distancia entre dos funciones de densidad debe ser invariante también frente a transformaciones admisibles de las variables aleatorias, ya que esto no afecta para nada a los objetos que deseamos distanciar, sino que sólo es un cambio en la forma con que tomamos las medidas, por ejemplo, cambios de escala, etc.

4. Finalmente, la distancia entre los objetos comparados debe aumentar si agregamos variables aleatorias, estocásticamente independientes, a las anteriores, ya que aumentará la información sobre las diferencias de los mismos /72/ .

Una forma razonable de proceder consiste en exigir que la distancia al cuadrado entre dos funciones de densidad de probabilidad infinitamente próximas $p(x,\theta)$, $p(x,\theta+d\theta)$ sea la esperanza del cambio infinitesimal de información al cuadrado. Teniendo en cuenta que $\theta = (\theta^1,\ldots,\theta^n)$ resulta

$$ds^2 = E[(df(p(x,\theta))^2] =$$

$$= \sum_{\mu=1}^{n} \sum_{\nu=1}^{n} E[(f'(p))^2 \partial_{\theta\mu} p \partial_{\theta\nu} p] d\theta^\mu d\theta^\nu \qquad (54)$$

Véase Rao /73/, Atkinson y Mitchell /3/, Burbea y Rao /6/, /7/, Oller /47/, /48/, Oller y Cuadras /49/, /50/, /52/, /53/.

Nótese que

$$g_{\mu\nu}(\theta) = E[(f'(p(x,\theta)))^2 \partial_{\theta\mu} p(x,\theta) \partial_{\theta\nu} p(x,\theta)] \qquad (55)$$

son las componentes, en coordenadas $\{\theta^i\}$, de un tensor métrico (o pseudométrico) sobre la variedad M (definida anteriormente), es decir, (54) define una métrica sobre M con la cual tendremos una estructura de variedad riemanniana (o semiriemanniana).

Dadas dos distribuciones $p(x,\theta_A)$, $p(x,\theta_B)$, identificaremos la distancia entre ellas con la distancia definida por la métrica (54) entre $\theta_A$, $\theta_B$. Dicha distancia no resulta alterada al efectuar transformaciones admisibles de los parámetros. Véase Hicks /38/, Spivak /63/ .

Por otra parte, si X es un vector aleatorio con distribución obsolutamente contínua, $p(x,\theta)$, fijado $\theta$, es una densidad escalar. Por tanto, al transformar, de forma admisible las X en Y , la función de densidad se transforma según:

$$\bar{p}(y.\theta) = p(x.\theta).\left| \det\left(\frac{\partial x}{\partial y}\right) \right| \qquad (56)$$

Para que la forma cuadrática diferencial, y por tanto, la distancia riemanniana, resulte invariante frente a cualquier cambio admisible $X \xrightarrow{T} Y$, deberá verificarse:

$$E((f'(p))^2 \partial_{\theta\mu} p \partial_{\theta\nu} p) = E((f'(\bar{p}))^2 \partial_{\theta\mu} \bar{p} \partial_{\theta\nu} \bar{p} \qquad (57)$$

$$\mu,\nu = 1,\ldots,n$$

con $p(x,\theta)$ arbitrarias, y para toda transformación T. Si llamamos $A_x$ al soporte de $p(x,\theta)$ $A_y$ al soporte de $\bar{p}(y,\theta)$ y $J = \left| \det\left(\frac{\partial x}{\partial y}\right) \right|$ entonces podemos escribir:

$$\int_{A_y} (f'(\bar{p}))^2 (\partial_{\theta\mu} \bar{p}\ \partial_{\theta\nu} \bar{p})\ \bar{p}\ dy =$$

$$= \int_{A_x} (f'(pJ))^2 (\partial_{\theta\mu} pJ\ \partial_{\theta\nu} pJ)\ p\ dx \qquad (58)$$

Y para que se cumpla (57), es necesario, dada la arbitrariedad de T y la generalidad de p, y también suficiente que:

$$f'(p) = f'(p.J).J \qquad (59)$$

derivando parcialmente respecto p obtenemos:

$$f''(p) = f''(p.J).J^2 \qquad (60)$$

y derivando ahora respecto J, resulta:

$$0 = f'(p.J) + f''(p.J).J\ p \qquad (61)$$

Combinando (59), (60), (61), se obtiene:

$$pf''(p) + f'(p) = 0 \qquad (62)$$

Ecuación diferencial cuya integral general viene dada por:

$$f(p) = \alpha \ell n\ p + \beta \qquad (63)$$

donde $\alpha$ y $\beta$ son constantes de integración arbitrarias.

Por tanto, para satisfacer, en el caso absolutamente contínuo, la condición de invariancia del elemento de línea definido a través de la forma cuadrática diferencial (54), frente a transformaciones admisibles de las variables aleatorias, la información asociada a un resultado x, debe venir definido por:

$$I(x,\theta) = \alpha \ell n(p(x,\theta)) + \beta \qquad (64)$$

El valor de la constante β no afecta a la forma cuadrática (54), ya que desaparece al derivar. La forma cuadrática (54) queda definida salvo una constante de proporcionalidad α, que en la práctica resulta recomendable escoger α= -1, (y β=0 por ejemplo), reduciéndose (64) a la definición de información dada por Shannon /59/.

El caso discreto es menos restrictivo, ya que en este caso p(x,θ) es un invariante, en vez de una densidad escalar, frente a transformaciones admisibles de variables aleatorias, por tanto independientemente de la función f que escojamos para definir la información asociada a un resultado, se garantizará la invariancia de la forma cuadrática diferencial (54).

Por otra parte, ni x e y son variables aleatorias, vectoriales, estocásticamente independientes, al verificarse:

$$p(x,y,\theta) = p_x(x,\theta).p_y(y,\theta) \qquad (65)$$

entonces definiendo la información a través de (64), obtenemos:

$$g_{\mu\nu}(\theta) = E\left(\frac{1}{p^2}\partial_{\theta\mu}p\partial_{\theta\nu}p\right) = E\left(\frac{1}{p_x^2}\partial_{\theta\mu}p_x\partial_{\theta\mu}p_x\right) +$$

$$+ E\left(\frac{1}{p_y^2}\partial_{\theta\mu}p_y\partial_{\theta\nu}p_y\right) = g^{(x)}_{\mu\nu}(\theta) + g^{(y)}_{\mu\nu}(\theta) \qquad (66)$$

$$\mu,\nu = 1,\ldots,n$$

por tanto:

$$ds^2 = ds_x^2 + ds_y^2 \qquad (67)$$

es decir, la forma cuadrática diferencial que define al elemento de línea es suma de formas cuadráticas, cada una de ellas calculada a partir de (54) y (64), considerando las variables x e y por separado.

La descomposición (67) asegura que al añadir variables aleatorias independientes, la distancia entre los objetos estudiados aumenta.

Como conclusión podemos pues asegurar que tenemos un procedimiento para definir una distancia entre las p(x,θ), (tanto en el caso absolutamente contínuo como en el caso discreto), y por ende, entre los objetos estudiados, que satisface los requerimientos exigidos inicialmente 1 a 4. Es la distancia riemanniana inducida por la forma cuadrática diferencial (54), definiendo f a partir de (64) con α = -1 y β = 0.

El campo tensorial métrico puede escribirse como:

$$g_{\mu\nu}(\theta) = E\left(\frac{1}{p^2}\partial_{\theta\mu}p\partial_{\theta\nu}p\right) = E(\partial_{\theta\mu}\ell np\partial_{\theta\nu}\ell np) \qquad (68)$$

$$\mu,\nu = 1,\ldots,n$$

y bajo condiciones de regularidad, de forma alternativa:

$$g_{\mu\nu}(\theta) = -E(\partial^2_{\theta\mu\theta\nu}\ell np) \qquad (69)$$

$$\mu,\nu = 1,\ldots,n$$

Nótese que las componentes del tensor métrico (68) coinciden con los elementos de la matriz de información de Fisher. La métrica obtenida se conoce con el nombre de métrica informacional.

## 7.2. OTRAS PROPIEDADES.

A continuación veamos algunas características y propiedades generales de esta distancia, para después discutir algunos resultados concretos.

Consideremos la variedad funcional definida por:

$$S = \{f \mid f = 2\sqrt{p} \quad p\in M\} \qquad (70)$$

donde, para simplificar, supondremos que p(x,θ) es una función de densidad de probabilidad absolutamente contínua (el caso discreto podría desarrollarse de forma análoga).

Esta variedad está inducida, de forma natural, en un espacio de Hilbert H, formado por funciones de cuadrado integrable, sobre un soporte fijo A, y cuyo producto escalar vie-

ne definido por:

$$\langle f,g\rangle = \int_A f.\bar{g} \qquad (71)$$

Entonces se sigue el siguiente:

Teorema 17

El tensor métrico definido en M a partir de (68) es el tensor métrico inducido en S por la métrica de H.

El resultado es consecuencia inmediate de - que el elemento de línea inducido por el producto escalar de H, en S, viene dado por:

$$ds^2 = \int_A (2\sqrt{p(x,\theta)} - 2\sqrt{p(x,\theta+d\theta)})^2 =$$

$$= \int_A \left( \frac{dp}{\sqrt{p(x,\theta)}} \right)^2 dx =$$

$$= \int_A \frac{1}{p(x,\theta)} \left( \sum_{\mu=1}^{n} \partial_{\theta\mu} p(x,\theta) d\theta\mu \right)^2 dx =$$

$$= \sum_{\mu=1}^{n} \sum_{\nu=1}^{n} E(\partial_{\theta\mu} \ell np \partial_{\theta\nu} \ell np)\ d\theta^{\mu} d\theta^{\nu} \qquad (72)$$

Nótese además que S está incluida en una -- parte de la superficie de una esfera de radio 2 en H.

Otro resultado notable viene dado por:

Teorema 18

Definiendo $I(x,\theta) = - \ell np(x,\theta)$, entonces - bajo ciertas condiciones de regularidad, el único tensor métrico $g_{\mu\nu}(\theta)$ que satisface la ecuación:

$$g_{\mu\nu} = E(I,\mu\nu) \qquad (73)$$

$$\mu,\nu = 1,\ldots,n$$

es el tensor métrico definido en (68).

Es consecuencia inmediata de (69).

Este resultado nos sugiere que podría utilizarse (73) para definir al tensor métrico - de la variedad riemanniana M, como el tensor igual a la esperanza de la segunda derivada covariante del invariante información.

Otros resultados proporcionan condiciones suficientes de variedad riemanniana euclídea:

Teorema 19.

Si la función de densidad $p(x,\theta)$ puede expresarse como producto de funciones de densidad marginales uniparamétricas:

$$p(x,\theta) = p_1(x_1,\theta^1) \ldots p_n(x_n,\theta^n) \qquad (74)$$

entonces la variedad riemanniana M, es ecuclídea.

Este resultado es consecuencia inmediate de que, bajo estas hipótesis, se anula el tensor de Riemann-Christoffel.

Teorema 20.

Si la función de densidad conjunta viene dada por:

$$p(x,\theta) = A(x)\exp\left( - \frac{1}{2} \| C-\theta \|^2 \right) \qquad (75)$$

donde $A(x)$ y $X(x) = (C_1(x),\ldots,C_n(x))$ son funciones que no dependen de $\theta$ y $\| . \|$ es la norma euclídea ordinaria, entonces la variedad M es euclídea.

Una de las características de una variedad euclídea es la existencia de un sistema de coordenadas tal que, bajo el mismo, el tensor métrico es un tensor constante (en particular igual a la identidad). Entonces, en dicho sistema de referencia es posible calcular la distancia entre puntos de M, a partir de

$$d(p(x,\gamma_A), p(x,\gamma_B)) = \sqrt{\sum_{i=1}^{n} (\gamma_A^i - \gamma_B^i)^2} \qquad (76)$$

La transformación de coordenadas $\theta \to \gamma$, viene definida como es bien sabido, /62/, - por el sistema de ecuaciones diferenciales:

$$\frac{\partial^2 \gamma^m}{\partial\theta^i \partial\theta^j} - \sum_{\alpha=1}^{n} \Gamma_{ij}^{\alpha} \frac{\partial\gamma^m}{\partial\theta\theta^{\alpha}} = 0 \qquad (77)$$

donde los $\Gamma_{ij}^{\alpha}$ son los símbolos de Christoffel de segunda especie.

## 7.3 DISTANCIAS RIEMANNIANAS PARA ALGUNAS DISTRIBUCIONES.

Se ha calculado explícitamente (Atkinson y Mitchel /3/, Burbea y Rao /6/, /7/, Oller /47/, /48/, Oller y Cuadras /49/, /50/, /53/) las distancias entre distribuciones de algunas familias paramétricas importantes, así como la curvatura riemanniana asociada a la variedad paramétrica. Citemos algunos de los resultados más remarcables.

### 7.3.1. DISTRIBUCIONES UNIPARAMETRICAS.

La variedad M es unidimensional, y por tanto euclídea, (curvatura cero). Veamos la distancia entre algunas distribuciones concretas:

Exponencial.

La expresión de la distancia riemanniana entre dos distribuciones exponenciales de parámetros $\lambda$ y $\mu$ viene dada por:

$$d = \left| \ell n \frac{\lambda}{\mu} \right| \qquad (78)$$

Poisson.

Dadas dos distribuciones de Poisson de parámetros $\lambda$ y $\mu$ , la distancia riemanniana vendrá dada por:

$$d = 2 \left| \sqrt{\lambda} - \sqrt{\mu} \right| \qquad (79)$$

Bernoulli.

La distancia riemanniana entre dos distribuciones de Bernoulli de parámetros p y q vendrá dada por:

$$d = 2 \left| \text{arc sen } \sqrt{p} - \text{arc sen } \sqrt{q} \right| \qquad (80)$$

### 7.3.2. DISTRIBUCIONES MULTIPARAMETRICAS

En estos casos la variedad funcional M no es en general euclídea, por lo que la expresión de la distancia será compleja de hallar. Habrá que resolver en general las ecuaciones diferenciales de las curvas geodésicas dadas por:

$$\frac{d^2\theta^r}{ds^2} + \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \Gamma^r_{\alpha\beta} \frac{d\theta^\alpha}{ds} \frac{d\theta^\beta}{ds} = 0 \qquad (81)$$

$$r = 1 \ldots n$$

y posteriormente determinar las constantes de integración para que la geodésica una dos puntos dados de la variedad. Veamos algunos casos concretos:

NORMAL UNIVARIANTE.

La curvatura gaussiana asociada a la variedad M, es igual a:

$$K = -\frac{1}{2} \qquad (82)$$

y la distancia riemanniana entre dos distribuciones normales univariantes de parametros $(\mu_A, \sigma_A)$ y $(\mu_B, \sigma_B)$ , medias y desviaciones típicas respectivamente), viene dada por:

$$d = \sqrt{2}\, \ell n \left( \frac{1 + \delta_{AB}}{1 - \delta_{AB}} \right) \qquad (83)$$

siendo:

$$\delta_{AB} = \left( \frac{(\mu_A - \mu_B)^2 + 2(\sigma_A - \sigma_B)^2}{(\mu_A - \mu_B)^2 + 2(\sigma_A + \sigma_B)^2} \right)^{1/2} \qquad (84)$$

NORMAL MULTIVARIANTE ( $\Sigma$ constante).

Este es un caso particular del Teorema 4. La variedad paramétrica M es euclídea,

$$K = 0 \qquad (85)$$

y la distancia riemanniana entre dos distribuciones normales multivariantes de vector de medias $\mu_A$ y $\mu_B$ (covarianza fija $\Sigma$) viene dada por:

$$d = \sqrt{(\mu_A - \mu_B)' \Sigma^{-1} (\mu_A - \mu_B)} \qquad (86)$$

que coincide con la distancia de Mahalanobis, /44/.

MULTINOMIAL.

Sean las funciones de densidad paramétricas:

$$p(x,\theta) = \frac{(x_1 + \ldots + x_n)!}{x_1! \ldots x_n!} (\theta^1)^{x_1} \ldots (\theta^n)^{x_n} \qquad (87)$$

con $x_1 + \ldots + x_n = N$, fijo. Entonces la variedad M tiene por curvatura:

$$K = \frac{1}{4N}$$

y la distancia riemanniana entre dos distribuciones multinomiales de parámetros

$(p^1,\ldots,p^n)$ y $(q^1,\ldots,q^n)$ viene dada por

$$d = 2\sqrt{N}\cos^{-1}\left(\sum_{j=1}^{n}\sqrt{p^j q^j}\right) \qquad (88)$$

que salvo constantes coincide con la distancia de Bhattacharyya.

MULTINOMIAL NEGATIVA.

Consideremos las funciones de densidad paramétricas:

$$p(x,\theta) = \frac{(x_1+\ldots+x_n+r-1)!}{x_1!\ldots x_n!(r-1)!}(\theta^1)^{x_1}\ldots$$

$$\ldots(\theta^n)^{x_n}(1-\theta^1-\ldots-\theta^n)^r \qquad (89)$$

para una constante positiva r. Entonces la variedad paramétrica M tiene curvatura igual a:

$$K = -\frac{1}{4r} \qquad (90)$$

y la distancia riemanniana entre dos distribuciones multinomiales negativas de parámetros $(p^1,\ldots,p^n)$ y $(q^1,\ldots,q^n)$, si llamanos $p^{n+1} = 1-p^1-\ldots-p^n$ y $q^{n+1}=1-q^1-\ldots-q^n$, resulta:

$$d = 2\sqrt{r}\cosh^{-1}\left(\frac{1-\sum_{j=1}^{n}\sqrt{p^j q^j}}{\sqrt{p^{n+1}q^{n+1}}}\right) \qquad (91)$$

## 7.4. PRUEBA DE SIGNIFICACION PARA UNA DISTANCIA.

En la práctica los parámetros resultan desconocidos, por tanto para calcular la distancia entre dos poblaciones estadísticas caracterizadas por $p(x_1\theta_A)$ y $p(x_1\theta_B)$ habrá que estimar previamente, por máxima verosimilitud, los vectores paramétricos $\theta_A$ y $\theta_B$. Asi, la distancia estimada D, será una variable aleatoria que para tamaños muestrales grandes tomará valores cercanos a la distancia real $\Delta$, en el sentido de converger en probabilidad hacia la misma. Bajo condiciones de regularidad (existencia de geodésicas en el interior de la variedad, definidas por funciones analíticas, etc.) es posible obtener la distribución asintótica de D, en el siguiente

Teorema 21.

La función de densidad de probabilidad asintótica del estadístico D (distancia estimada), dada la distancia real , viene dada por:

$$f(D,\Delta)=\left(\frac{-i}{\Delta}\right)^{\frac{n-2}{2}}\frac{N_A N_B}{N_A+N_B}D^{\frac{n}{2}}.\exp\left\{-\frac{1}{2}\frac{N_A N_B}{N_A+N_B}(D^2+\Delta^2)\right\}:$$

$$\frac{J_{n-2}}{2}\left(i\frac{N_A N_B}{N_A+N_B}D\Delta\right) \qquad D\geq 0$$

donde n es la dimensión de la variedad M, $N_A$ y $N_B$ los tamaños muestrales en las poblaciones A y B respectivamente, $i = \sqrt{-1}$ y $J_{\frac{n-2}{2}}$ la función de Bessel de argumento imaginario puro e índice $\frac{n-2}{2}$.

Como consecuencia, se tiene, para $\Delta = 0$, el siguiente resultado:

Teorema 22.

La función de densidad asintótica del estadístico $U = \frac{N_A N_B}{N_A+N_B}D^2$ es una distribución ji-cuadrado con n-grados de libertad.

Ambos teoremas pueden utilizarse para contrastar $H_o$: $\Delta=0$ frente $H_1$: $\Delta>0$, mediante la prueba ji-cuadrado.

## 8. UN EJEMPLO DE APLICACION.

El siguiente ejemplo nos va a servir para ilustrar algunos de los métodos geométricos de representación de datos, pertenecientes a geometría euclídea, ultramétrica y diferencial. En el primer caso utilizaremos una representación euclídea por análisis de coordenadas principales, en el segundo caso un dendograma y en el tercero una distancia geodésica entre modelos lineales.

El estudio se realizó sobre una poblcción de niños sospechosos de padecer Diabetes Melitus a los cuales se les practicó (en el H.S.J. de Dios de Barcelona) un TTOG, que como es sabido consiste en medir la Glucosa y la Insulemia del niño a los 0 (basal), 30, 60, 90, 120, 150 y 180 minutos de haber ingerido 1.75 grs.

de glucosa por Kg. de peso, hasta un máximo de 75 grs.

Los datos que proporcionaban estas observaciones nos permitían conocer el funcionamiento del páncreas endocrino, así como diagnosticar o clasificar a un individuo de diabético o no siguiendo los criterios establecidos por el National Diabeth Data Group.

Los criterios de diagnóstico generalmente usados en la práctica clínica, aun aceptando que son objetivos, creemos que no aprovechan en su totalidad la información que sobre la curva de glucemia/insulina nos proporciona el TTOG, ya que se refiere únicamente a observaciones puntuales de las curvas y no hacen una valoración global de las mismas.

Ahora bien, si podemos definir una distancia haciendo uso de toda la información de las curvas, nos será posible mediante los métodos geométricos de análisis de datos, llegar a una clasificación objetiva y sistemática de la población de niños estudiada.

## 8.1. METODOLOGIA.

Los datos corresponden a las N=33 curvas de glucemia/insulina de N=33 niños después de practicar un TTOG.

Dado que el error de las determinaciones de glucosa y de insulina no eran constantes, es decir, las variables observadas no tenían igual varianza para distintos valores de t (tiempo), se aplicó la transformación logarítmica. La transformación elegida se justificó por el hecho de que el cociente entre las medias muestrales de las determinaciones y sus derivaciones típicas eran constantes, exactamente 5 para la glucosa y 2.5 para la insulina.

## 8.2. MODELO EMPLEADO.

Se supuso que las 2 curvas se ajustaban a un polinomio de tercer grado. El modelo es pues

$$\begin{aligned} y_i &= \alpha_0+\alpha_1 t_i+\alpha_2 t_i^2+\alpha_3 t_i^3+e_i \\ w_i &= \beta_0+\beta_1 t_i+\beta_2 t_i^2+\beta_3 t_i^3+e_i^* \end{aligned} \qquad (92)$$

siendo $y_i$, $w_i$ los logaritmos de las determinaciones de glucosa e insulina en el tiempo $t_i$, y $e_i, e_i^*$ las variables aleatorias de los errores de las medidas de glucosa e insulina, que, tras la transformación efectuada, se ajustan a variables normales de media 0 y varianzas $\sigma_1^2, \sigma_2^2$, siendo $\sigma_2^2 = 4\sigma_1^2$. Suponemos además que $y_i$, $w_i$ son independientes.

Por otra parte, como de la variable insulina se tomaban 2 réplicas por cada $t_i$, se adoptó como variable dependiente la media muestral de las dos réplicas.

$$z_i = (w_i(1) + w_i(2))/2$$

Por consiguiente, el modelo (92) queda

$$\begin{aligned} y_i &= \alpha_0+\alpha_1 t_i+\alpha_2 t_i^2+\alpha_3 t_i^3+e_i \\ z_i &= \beta_0+\beta_1 t_i+\beta_2 t_i^2+\beta_3 t_i^3+e_i^* \end{aligned} \qquad (93)$$

donde $e_i$ es $N(0, \sigma_1^2)$, $e_i^*$ es $N(0, 2\sigma_1^2)$.

La expresión matricial de (93) es

$$y = X\alpha + e \qquad z = X\beta + e^* \qquad (94)$$

siendo:

$$y = (y_0, y_1, \ldots, y_n) \qquad z = (z_0, z_1, \ldots, z_n)$$

$X = (x_{ij})$ es la matriz de diseño, con

$$x_{ij} = t_{i-1}^{j-1} \qquad \begin{matrix} 1 \le j \le 4 \\ 1 \le i \le n+1 \end{matrix}$$

$$e = (e_0, \ldots, e_n)' \qquad e^* = (e_0, \ldots, e_n)'$$

$\alpha$ y $\beta$ son los vectores de parámetros

$$\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)' \qquad \beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$$

Expresando conjuntamente el modelo (94) tenemos

# TABLA 2

VALORES MEDIOS POR GRUPO

| | | | |
|---|---|---|---|
| GRUPO I | GLUCOSA | T1 | 96.00 |
| INDIVIDUO NUM. 26 | GLUCOSA | T2 | 202.0 |
| | GLUCOSA | T3 | 253.0 |
| | GLUCOSA | T4 | 177.0 |
| | GLUCOSA | T5 | 131.0 |
| | GLUCOSA | T6 | 67.00 |
| | GLUCOSA | T7 | 60.00 |
| | INSULINA | T1 | 1180. |
| | INSULINA | T2 | 2814. |
| | INSULINA | T3 | 4000. |
| | INSULINA | T4 | 3808. |
| | INSULINA | T5 | 3428. |
| | INSULINA | T6 | 2060. |
| | INSULINA | T7 | 1233. |
| GRUPO IIA | GLUCOSA | T1 | 231.0 |
| INDIVIDUO NUMS. | GLUCOSA | T2 | 364.0 |
| 21-22 | GLUCOSA | T3 | 425.0 |
| | GLUCOSA | T4 | 459.5 |
| | GLUCOSA | T5 | 455.0 |
| | GLUCOSA | T6 | 390.0 |
| | GLUCOSA | T7 | 354.0 |
| | INSULINA | T1 | 17.65 |
| | INSULINA | T2 | 21.70 |
| | INSULINA | T3 | 19.60 |
| | INSULINA | T4 | 25.60 |
| | INSULINA | T5 | 21.10 |
| | INSULINA | T6 | 15.55 |
| | INSULINA | T7 | 16.00 |
| GRUPO IIB | GLUCOSA | T1 | 301.0 |
| INDIVIDUO NUM. 14 | GLUCOSA | T2 | 485.0 |
| | GLUCOSA | T3 | 517.0 |
| | GLUCOSA | T4 | 519.0 |
| | GLUCOSA | T5 | 558.0 |
| | GLUCOSA | T6 | 534.0 |
| | GLUCOSA | T7 | 522.0 |
| | INSULINA | T1 | 5.700 |
| | INSULINA | T2 | 5.000 |
| | INSULINA | T3 | 5.000 |
| | INSULINA | T4 | 5.000 |
| | INSULINA | T5 | 5.000 |
| | INSULINA | T6 | 5.000 |
| | INSULINA | T7 | 5.000 |
| GRUPO III | GLUCOSA | T1 | 139.0 |
| INDIVIDUO NUM. 20 | GLUCOSA | T2 | 252.0 |
| | GLUCOSA | T3 | 331.0 |
| | GLUCOSA | T4 | 365.0 |
| | GLUCOSA | T5 | 392.0 |
| | GLUCOSA | T6 | 385.0 |
| | GLUCOSA | T7 | 321.0 |
| | INSULINA | T1 | 107.0 |
| | INSULINA | T2 | 112.0 |
| | INSULINA | T3 | 108.0 |
| | INSULINA | T4 | 104.0 |
| | INSULINA | T5 | 109.0 |
| | INSULINA | T6 | 112.0 |
| | INSULINA | T7 | 115.0 |
| GRUPO IVA | GLUCOSA | T1 | 92.84 |
| INDIVIDUOS NUMS. | GLUCOSA | T2 | 143.1 |
| 4-11-13-19-25 | GLUCOSA | T3 | 139.9 |
| 27-29-30 | GLUCOSA | T4 | 123.4 |
| | GLUCOSA | T5 | 120.7 |
| | GLUCOSA | T6 | 109.2 |
| | GLUCOSA | T7 | 102.0 |
| | INSULINA | T1 | 33.51 |
| | INSULINA | T2 | 185.8 |
| | INSULINA | T3 | 201.6 |
| | INSULINA | T4 | 146.7 |
| | INSULINA | T5 | 144.6 |
| | INSULINA | T6 | 121.7 |
| | INSULINA | T7 | 95.59 |
| GRUPO IVB | GLUCOSA | T1 | 82.00 |
| INDIVIDUO NUM.9 | GLUCOSA | T2 | 131.0 |
| | GLUCOSA | T3 | 197.0 |
| | GLUCOSA | T4 | 249.0 |
| | GLUCOSA | T5 | 205.0 |
| | GLUCOSA | T6 | 152.0 |
| | GLUCOSA | T7 | 111.0 |
| | INSULINA | T1 | 16.20 |
| | INSULINA | T2 | 21.20 |
| | INSULINA | T3 | 22.40 |
| | INSULINA | T4 | 22.80 |
| | INSULINA | T5 | 86.00 |
| | INSULINA | T6 | 63.00 |
| | INSULINA | T7 | 26.20 |
| GRUPO IVC | GLUCOSA | T1 | 88.55 |
| INDIVIDUOS NUMS. | GLUCOSA | T2 | 132.9 |
| 1-2-3-5-6-7-8-10 | GLUCOSA | T3 | 118.3 |
| 12-15-16-17-18-23 | GLUCOSA | T4 | 104.6 |
| 24-28-31-32-33 | GLUCOSA | T5 | 99.71 |
| | GLUCOSA | T6 | 96.02 |
| | GLUCOSA | T7 | 88.09 |
| | INSULINA | T1 | 19.22 |
| | INSULINA | T2 | 89.02 |
| | INSULINA | T3 | 72.40 |
| | INSULINA | T4 | 57.59 |
| | INSULINA | T5 | 48.84 |
| | INSULINA | T6 | 42.84 |
| | INSULINA | T7 | 31.48 |

DENDROGRAM

| CYCLE | ITEMS GROUPED I | J | COEFF |
|---|---|---|---|
| 1 | 18 | 32 | 0.134 |
| 2 | 28 | 33 | 0.369 |
| 3 | 15 | 28 | 0.456 |
| 4 | 17 | 31 | 0.477 |
| 5 | 5 | 17 | 0.577 |
| 6 | 21 | 22 | 0.603 |
| 7 | 4 | 30 | 0.651 |
| 8 | 5 | 6 | 0.654 |
| 9 | 3 | 15 | 0.672 |
| 10 | 1 | 2 | 0.712 |
| 11 | 4 | 19 | 0.719 |
| 12 | 5 | 24 | 0.821 |
| 13 | 5 | 7 | 0.891 |
| 14 | 1 | 23 | 0.901 |
| 15 | 4 | 11 | 0.916 |
| 16 | 1 | 5 | 1.046 |
| 17 | 4 | 13 | 1.089 |
| 18 | 1 | 12 | 1.093 |
| 19 | 8 | 18 | 1.116 |
| 20 | 4 | 29 | 1.184 |
| 21 | 3 | 10 | 1.249 |
| 22 | 1 | 3 | 1.399 |
| 23 | 1 | 8 | 1.722 |
| 24 | 4 | 27 | 1.873 |
| 25 | 4 | 25 | 1.903 |
| 26 | 1 | 16 | 2.035 |
| 27 | 1 | 9 | 2.588 |
| 28 | 1 | 4 | 3.091 |
| 29 | 14 | 21 | 3.608 |
| 30 | 1 | 20 | 3.795 |
| 31 | 1 | 14 | 5.799 |
| 32 | 1 | 26 | 10.241 |

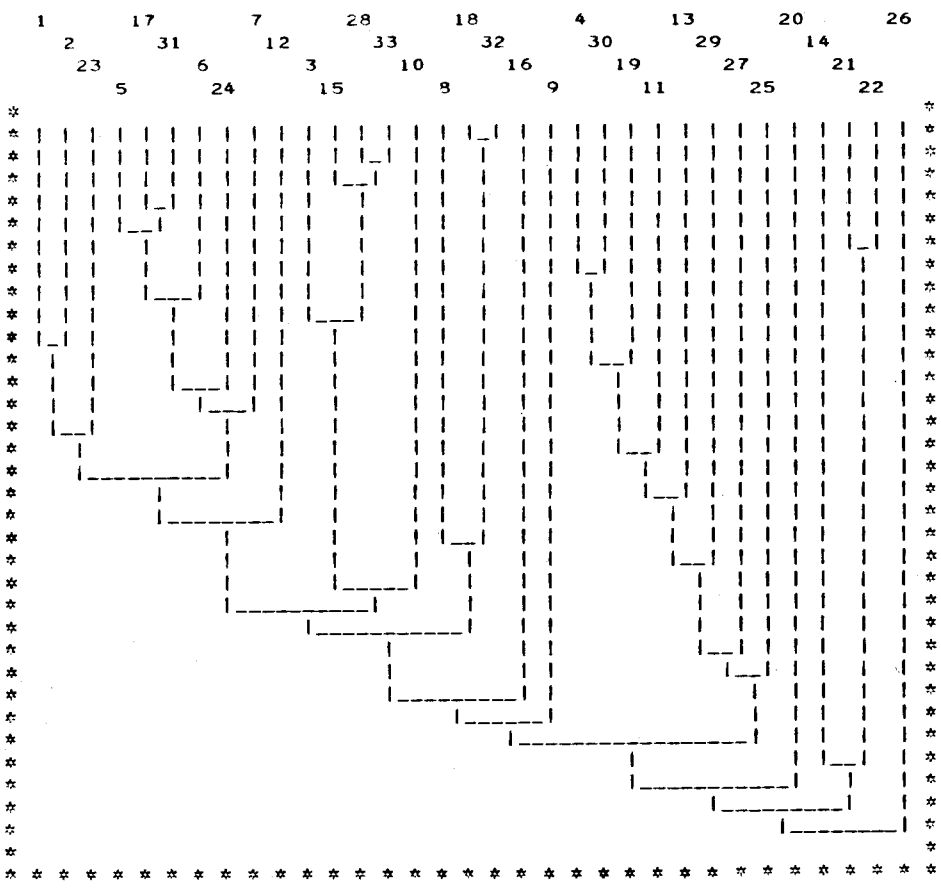


Figura 6 : Dendograma construido a partir de las distancias entre curvas de respuesta, respecto a glucosa e insulemia, sobre 33 niños.

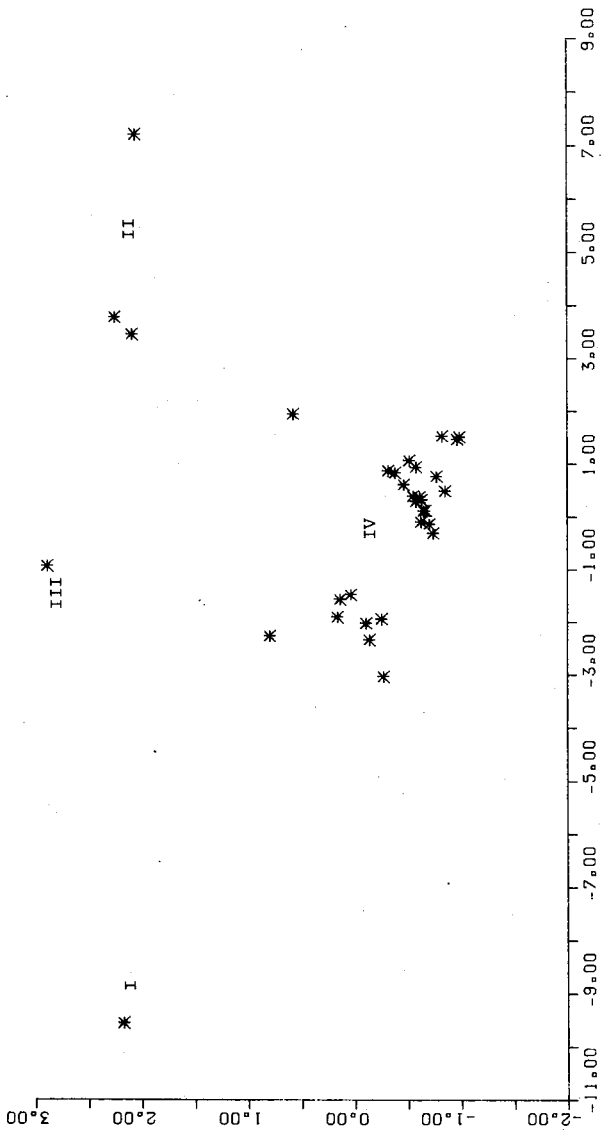


Figura 7: Representación por coordenadas principales de las distancias entre curvas de respuesta sobre 33 niños.

$$Y = XB + E \qquad (95)$$

siendo

$$Y = \begin{pmatrix} y_0 & z_0 \\ y_1 & z_1 \\ \vdots & \vdots \\ y_n & z_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & t_0 & t_0^2 & t_0^3 \\ 1 & t_1 & t_1^2 & t_1^3 \\ & \cdots & \cdots & \\ & \cdots & \cdots & \\ 1 & t_n & t_n^2 & t_n^3 \end{pmatrix}$$

$$B = \begin{pmatrix} \alpha_0 & \beta_0 \\ \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \\ \alpha_3 & \beta_3 \end{pmatrix} \qquad E = \begin{pmatrix} e_0 & e_0^* \\ e_1 & e_1^* \\ \vdots & \vdots \\ e_n & e_n^* \end{pmatrix}$$

En nuestro caso es $n = 6$, $t_i = 30i$ $(0 \le i \le 6)$.

Ambos diseños son de rango máximo

$$\text{rang } X = 4$$

## 8.3. DISTANCIA ENTRE DOS MODELOS.

Supongamos que tenemos 2 modelos distintos e independientes

$$Y_a = XB_a + E_a \qquad Y_b = XB_b + E_b$$

Aplicando las técnicas descritas en el apartado 7, una estimación de la distancia$^2$ entre ambos modelos es /56/,

$$\hat{D}^2 = \text{traza}\ ((Y_a - Y_b)'X(X'X)^{-1}X'(Y_a - Y_b)S^{-1}) \qquad (96)$$

donde S es la estimación insesgada de la matriz de covarianzas (supuesta común) obtenida por el procedimiento usual en modelos lineales multivariantes /17/. $\hat{D}^2$ permite comparar los 2 modelos, representando una generalización multivariante al problema de comparar curvas experimentales /14/.

Pero, al suponer las variables y, z independientes, tenemos en nuestro caso

$$\hat{D}^2 = \hat{L}_1^2 + \hat{L}_2^2 \qquad (97)$$

donde $\hat{L}_1^2$ en la estimación de la distancia$^2$ entre los modelos correspondientes a la variable y (glucosa),

$$y_a = X\alpha_a + e_a \qquad y_b = X\alpha_b + e_b \qquad (98)$$

cuya expresión es

$$\hat{L}_1^2 = \frac{1}{\sigma_1^2}\ ((y_a - y_b)'X(X'X)^{-1}X'(y_a - y_b)) \qquad (99)$$

Análogamente, $L_2^2$ es la distancia$^2$ para z (insulina), cuya expresión es análogamente

$$\hat{L}_2^2 = \frac{1}{2\sigma_1^2}\ ((z_a - z_b)'X(X'X)^{-1}X'(z_a - z_b)) \qquad (100)$$

Por otra parte, se sabe como dato de laboratorio, que el cociente entre media m y desviación típica $\sigma(y)$ de la determinación de glucosa, es

$$\frac{m}{\sigma(y)} = 5$$

Se obtiene entonces /56/ que

$$\sigma_1^2 = \frac{1}{25} \qquad \sigma_2^2 = \frac{2}{25}$$

Luego, en (99) y (100) las varianzas son conocidas.

## 8.4. RESULTADOS OBTENIDOS Y DISCUSION.

Una vez obtenida la matriz de distancias, se construyó un dendograma, aplicando el método UPGMA. La correlación cofenética resultó ser 0.92, lo que indica que la jerarquía construida reproduce bien la distancia inicial.

La fig. 6 representa el dendograma de la población de niños a los cuales se les realizó el TTOG. De este dendograma, cortando a distancia 4, se obtienen cuatro grupos que denotaremos con los números I, II, III, IV, y que pasamos seguidamente a describir.

El grupo I está formado por el niño número 26. Este niño, como se ve en la tabla 2, tiene una gran insulino-resistencia, pues tiene valores altos de glucosa mientras dura la prueba y sin embargo, los valores de insulina son extremadamente altos.

El grupo II está formado por tres niños, con los números 14, 21, 22 (ver tabla 2). Estos niños son diabéticos insulinopénicos que ya en ayunas tienen valores de glucemia superiores a los 200 mg/100 c.c. Al cortar a distancia 2.5 en este grupo se distinguen dos sub-

grupos: el IIa formado por los niños con los números 21 y 22, cuyas curvas de glucemia-insulina corresponden a individuos diabéticos insulinopénicos, y el grupo IIb formado por un solo individuo con el número 14. Como puede observarse, no hay prácticamente respuesta insulínica.

El grupo III está formado por un solo individuo con el número 20 (ver tabla 2). Este es un caso típico de diabetes diagnosticada a través del TTOG. Su glucemia en ayunas era de 139 mg/100 c.c. y tras la estimulación se alcanzan en todos los tiempos observados valores de glucemia superiores a los 200 mg/c.c.

Finalmente el grupo IV, al que pertenecen la mayor parte de los niños, son niños con una curva de glucemia normal salvo en un caso que ahora analizaremos.

Al cortar a nivel de distancia 2.5 en el grupo IV se distinguen tres subgrupos. El subgrupo IVa está formado por 8 individuos que presentan una curva de glucemia normal, sin embargo, presentan valores de insulina sensiblemente más elevados que los normales (ver tabla 2). El grupo IVb está formado por un solo individuo con el número 9 de la tabla 2, cuyos valores de glucosa en los tiempos 90 m. y 120 m. son mayores que 200 mg/100 c.c. por lo que, según el National Dialetes Data Group se diagnostica como diabético.

El subgrupo IVc está formado por 19 individuos con curva de glucemia e insulina normales.

La clasificación se ha completado con un análisis de coordenadas principales, partiendo de la matriz de distancias entre los N=33 niños, calculada a partir de (97), (99), (100). La variabilidad explicada por los 2 primeros ejes fue del 92% . La fig. 7 contiene los resultados de este análisis. Vemos reflejados los cuatro grupos obtenidos por taxonomía numérica.

## 9. CONCLUSIONES.

Los diferentes métodos geométricos para representar un conjunto I provisto de una distancia d, a través de un espacio geométrico modelo $(V,\delta)$, pueden ser clasificados en función de las propiedades de la distancia $\delta$ . Tales propiedades pueden ser:

1) $\delta(i,j)\geq 0$
2) $\delta(i,j)=\delta(j,i)$
3) $\delta(i,i)=0$
4) $\delta(i,j)\leq\delta(i,k)+\delta(j,k)$
5) $\delta(i,j)$ es distancia euclídea
6) $\delta(i,j)\leq\max\{\delta(i,k);\delta(j,k)\}$
7) $\delta(i,j)+\delta(k,m)\leq\max\{\delta(i,k)+\delta(j,m);\delta(i,m)+\delta(j,k)\}$
8) $\delta(i,j)$ se obtiene a partir de una métrica de Riemann.

Cualquier distancia verifica 1), 2), 3). Existen diversas implicaciones entre las demás propiedades. Por ejemplo:

$6) \Rightarrow 5) \Rightarrow 4)$ $\quad$ $6) \Rightarrow 7) \Rightarrow 4)$

La representación de (I,d) a través de $(V,\delta)$ puede ser exacta o aproximada. La estructura de una representación vendrá condicionada a la estructura de $(V,\delta)$. Así hablaremos de representación euclídea si $\delta$ verifica 5), de representación ultramétrica si $\delta$ verifica 6), de representación aditiva si $\delta$ verifica 7) y de representación riemanniana si $\delta$ verifica 8).

Finalmente, el caso no euclídeo se resuelve mediante realizaciones monótonas en espacios euclídeos o de Riemann, conservando la preordenación asociada a la distancia d.

## 10 REFERENCIAS.

/1/ ARCAS, A.: "Relaciones entre arboles aditivos y ultramétricos". En: Homenatge a F. de A. Sales, Cont. Cient., F. Matem., Univ. Barcelona, 8-15.1985.

/2/ ARCAS, A.: "Contribuciones a la representación de datos multidimensionales mediante árboles aditivos". Tesis doctoral (inédita). - 198 .

/3/ ATKINSON, C. & MITCHEL, A.F.S.: "Rao's distance measure". Sankhya, 43A, 345-365, 1981.

/ 4/ BENZECRI, J.P.: "L'analyse des donnees I. La taxinomie. L'analyse des donnees II. L'analyse des correspondances" - Dunod, Paris 1976.

/ 5/ BUNEMAN, P.: "The recovery of trees from measures of dissimilarity". En: F.R. Hodson, D.G. Kendall & P. Tautu (Eds.), Mathematics in the Archeological and Historical Sciences. Edinburgh, Edinburgh University Press. 1971.

/ 6/ BURBEA, J. & RAO, C.R.: "Entropy differential metric, distance and divergence measures in probability spaces: a unified approach". J. of Multivariate analysis, 12, 575-596. 1982.

/ 7/ BURBEA, J. & RAO, C.R.: "Differential metrics in probability spaces". Prob. Math. Stat., 3, 241-258. 1984.

/ 8/ CAILLIEZ, F.: "The analytical solution of the additive constant problem". Psychometrika. Vol. 48, No. 2. 305-308. 1983.

/ 9/ CAILLIEZ, F. & PAGES, J.P.: "Introduction a l'analyse des donnees". SMASH, Paris. 1976.

/10/ CARROLL, J.D. & CHANG, J.J.: "Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young descomposition". Psychometrika 35, 283-319. 1970.

/11/ COOPER, C.H.: "A new solution to the additive constant problem in metric multidimensional scaling". Psychometrika, 37, 311-322. 1972.

/12/ CUADRAS, C.M.: "Analisis discriminante de funciones parametricas estimables". Trab. Estad. Inv. Oper., 25(3), 3-31. 1974.

/13/ CUADRAS, C.M.: "Sobre la reducción de la dimensión en análisis estadístico multivariante". Trab. Estad. I.O. 28, 63-76. 1977.

/14/ CUADRAS, C.M.: "Sobre la comparació estadística de corbes experimentals. QUESTIIO, 3(1), 1-10. 1979.

/15/ CUADRAS, C.M.: "Mètodes de representació de dades i la seva aplicació en Biologia". Col. Soc. Catalana Biologia, 13, 95-133. 1980.

/16/ CUADRAS, C.M.: "Análisis y representación multidimensional de la variabilidad". Inter. Sym. Concept.Meth. Paleo., Barcelona, 287-297. 1981.

/17/ CUADRAS, C.M.: "Métodos de análisis multivariante". Eunibar, 642pp., Barcelona 1981.

/18/ CUADRAS, C.M.: "Conceptos metodológicos probabilísticos y geométricos en Bioestadística". Pub. Bioestadística y Biomatemática, 4, Ed. Pub. Univ. Barna. 288 pp. 1983.

/19/ CUADRAS, C.M.: "Análisis algebraico sobre distancias ultramétricas". Actas 44 Per. de sesiones del Ist. Intern. de Estadística, Madrid. Cont. Libres, Vol. II, 554-557. 1983

/20/ CUADRAS, C.M. & CARMONA, F.: "Dimensionalitat euclidiana en distancies ultrametriques". QUESTIIO, 7(1), 353-358. 1983.

/21/ CUADRAS, C.M. & OLLER, J.M.: "Geometria finita aplicada a la estadística". Actas XIV Cong. Nac. Est. Inv. Op. Inf., Granada 287-296. 1984.

/22/ CUADRAS, C.M. & OLLER, J.M.: "Eigenanalysis and metric multidimensional scaling on hierarchical structures". Sometido a Psychometrika, 1985.

/23/ CUADRAS, C.M. & RUIZ-RIVAS, C.: "Una contribución al análisis de proximidades". Pub. Secc. Matem. Univ. Au. Barcelona, 22, 103-106. 1980.

/24/ CUNNINGHAM, J.P.: "Free trees and bidirectional trees as representations of psychological distance". J. of mathematical psychology, 17, 165-188. 1978.

/25/ DE LEEUW, J. & HEISER, W.: "Theory of multidimensional sccaling". En: Handbook of statistics. Vol. 2., P.R. Krishnaiah, L.N. Kanal (Eds.), North-Holland Pub. Co., Amsterdam, 1982.

/26/ DE SOETE, G.: "A least square algorithm for fitting additive trees to proximity data". Psychometrika, 48(4), 621-626. 1983.

/27/ DEMPSTER, A.P.: "Elements of continuous multivariate analysis". Addison-Wesley, Reading, Mass. 1969.

/28/ GABRIEL, K.R.: "The biplot graphic display of matrices with application to principal component analysis". Biometrika, 58, 453-467. 1971.

/29/ GABRIEL, K.R.: "Biplot". En: Enzyclopedia of statistical sciences (S.Kotz, N.L. Johnson, Eds.) Wiley, New York. 1981.

/30/ GALINDO, P.: "Contribuciones a la representación simultánea de datos multidimensionales". Tesis Doctoral (inédita). U. de Salamanca, 1985.

/31/ GOLUB, G.H. & REINSCH, C.: "Singular value decomposition and least squares solutions". Numer. Math., 14, 403-420, 1970.

/32/ GOWER, J.C.: "Euclidean distance geometry". Math. Scientist, 7, 1-14. 1982.

/33/ GOWER, J.C.: "Some distance properties of latent root and vector methods in multivariate analysis". Biometrika, 53, 315-328. 1966.

/34/ GOWER, J.C. & BANFIELD, C.F.: "Goodness-of-fit criteria for hierarchical classification and their empirical distributions". En : Proceedings of the 8th Inter. Biometric Conference, 347-361. 1975.

/35/ GOWER, J.C. & DIGBY, P.G.N.: "Expressing complex relationships in two dimensions". En: Interpreting Multivariate Data (V. Barnett, Ed.) 83-118, J. Wiley, New York, 1981.

/36/ HARSMAN, R.A.: "Parafac 2: Mathematical and technical notes". Ucla, Working papers in phonetics, No. 22. 1972.

/37/ HERR, D.G.: "Geometry in statistics". En: Encyclopedia of Statistical Sciences. (S. Kotz, N.L. Johson, Eds.), Wiley, New York. 1981.

/38/ HICKS, N.J.: "Notes on differential geometry". Van Nostran, Princeton.1965.

/39/ HOLMAN, E.W.: "The relation between hierarchical and euclidean models for psychological distances". Psychometrika Vol.No.4,417-423. 1972.

/40/ LEBART, L. & MORINEAU, A. & FENELON,J.P. "Tratamiento de datos". Marcombo, Boixareu Editores, Barcelona, 1985.

/41/ LEFEBVRE, J.: "Introducción aux analyses statistiques multidimensionnelles". Masson, Paris, 2 Ed., 1980. 1976.

/42/ LEW, J.S.: "Some counterexamples in multidimensional scaling". J. of Math.Psych. 17, 247-254. 1978.

/43/ LINGOES, J.C.: "Some boundary conditions for a monotone analysis of symmetric matrices". Psychometrika, 36, 195-203. -- 1971.

/44/ MAHALANOBIS, P.C.: "On the generalized distance in statistics". Proc. Nat.Inst. Sci. India, 2(1), 49-55. 1936.

/45/ MARDIA, K.V.: "Some properties of classical multidimensional scaling". Comm. Stat., A7 (13), 1233-1241.

/46/ OHSUMI, N. & NAKAMURA, T.: "Some properties of monotone hierarchical dendrogram in numerical classification". Proc.Inst. Statist.Mathem.(Tokio), 28(1), 117-133. 1981.

/47/ OLLER, J.M.: "Utilización de métricas riemanianas en análisis de datos multidimensionales y su aplicación a la biología". Publicaciones de Bioestadística y Biomatemática N.11. 288pp. 1983.

/48/ OLLER, J.M.: "Sobre la curvatura riemanniana asociada a distribuciones de series de potencias generalizadas multivariantes". Homen F. de A. Sales, F. Matem., Univ. Barcelona, 118-120. 1985.

/49/ OLLER, J.M. & CUADRAS, C.M.: "Defined distances for some probability distributions". Proceed. II Word Conf. Math. Serv. Man. (A. Ballester, D. Cardus, E. Trillas, Eds.) Un. Pol. de las Palmas, 563-565. 1982.

/50/ OLLER, J.M. & CUADRAS, C.M.: "On a distance defined for multivariate negative multinomial distribution". Abstracts of contr. papers. XI Intern. Biometric. Confer., Toulouse,P.69. 1982

/51/ OLLER, J.M. & CUADRAS, C.M.: "Representación canónica en manova: aplicación a una clase de diseño anidado". QUESTIIO 6(3), 221-229. 1982.

/52/ OLLER, J.M. & CUADRAS, C.M.: "Sobre una distancia definida para la distribución normal multivariante". XIII Jorn. de Estad. I.Op. e Inform., D. de Estadística U. de Valladolid, Actas Vol.II, Sec.III 32-36. 1983.

/53/ OLLER, J.M. & CUADRAS,C.M.: "Rao's distance for negative multinomial distributions". Sankkya series A, 47(1), 75-83, 1985.

/54/ PEARSON, E.S.: "Pearson, creador de la estadística aplicada". Espasa-Calpe Argentina, Buenos Aires-México. 1948.

/55/ PRUZANSKY, S., TVERSKY, A. & CARROL,J.D: : "Spatial versus tree representations of proximity data". Psychometrika, Vol. 47, 1, 3-24. 1982.

/56/ RIOS, M.: "Distancias entre modelos lineales y su aplicación a la clasificación y diagnóstico de enfermedades". Tesis doctoral ( inédita) . 1985.

/57/ SCHEFFE, H.: "The analysis of variance". J. Wiley, New York. 1959.

/58/ SCHOENBERG, I.J.: "Remarks to Maurice Frechet's article sur la definition axiomatique d'une classe d'espaces vectorieelles distanciés applicables vectoriellement sur l'espace de Hilbert". Ann. Math. 36, 724-732. 1935.

/59/ SHANNON, C.E.: "A mathematical theory of communications". Bell System Tech. J.27, 379-423. 1948.

/60/ SHEPARD, R.N.: "The analysis of proximities. Multidimensional scaling with an unknown distance function. I". Psychometrika, 27, 125-140. 1962.

/61/ SHEPARD, R.N.: "The analysis of proximities. Multidimensional scaling with an unknown distance function. II". Psychometrika, 27, 219-246. 1962.

/62/ SOKOLNIKOFF, I.S.: "Analisis tensorial". Ed. Index. 1971.

/63/ SPIVAK, M.: "A comprehensive introduction to differential geometry". Publish or perish, inc.Berkeley. 1979.

/64/ TAKANE, Y., YOUNG, F.W. & DE LEEUW,J.: "Nonmetric individual differences multidimensional scaling alternating least square method with optimal scaling features". Psychometrika, 42, 7-67. 1977.

/65/ TORGERSON, W.S.: "Theory and methods of scaling". J. Wiley, New York. 1958.

/66/ WATERMAN, M.S., SMITHY, T.F., SINGH, M. & BEYER, W.A.: "Additive evolutionary trees". J. Theor. Biol., 64, 199-213. 1977.

/67/ KRUSKAL, J.B.: "Multidimensional scaling by optimising goodness of fit to a nonmetric hiphotesis". Psychometrika, 29, 1-27. 1964.

/68/ KRUSKAL, J.B.: "Nonmetric multidimensional scaling: a numerical method". Psychometrika, 29, 115-129. 1964.

/69/ LINGOES, J.C. & BORG, I.: "A direct approach differences scaling using increasingly complex transformations. Psychometrika, 43(4), 491-519, 1978.

/70/ JOHNSON, S.C.: "Hierarchical clustering schemes". Psychometrika, 32, 241-254. 1967.

/71/ SATTATH, S.& TVERSKY, A.: "Additive similarity trees". Psychometrika, 42 (3), 319-345. 1977.

/72/ OLLER, J.M. & CUADRAS, C.M.: "Sobre ciertas condiciones que deben verificar las distancias entre espacios probabilísticos". Actas XV Reunión SEIO, Gijón, 1985.

/73/ RAO, C.R.: "Information and accuracy attainable in the estimation of statistical parameters". Bull. Calcuta Math. Soc. 37, 81-91. 1945.

/74/ LINDMAN, H. & CAELLI, T.: "Constant curvature riemmannian scaling". J. of Mathem. Psychol., 17, 89-109, 1978.

/75/ PIESZKO, H.: "Multidimensional scaling in riemannian space". J. of Mathem. Psychol., 12, 449-477, 1975.