ESTADISTICA ESPAÑOLA
Vol. 30, Núm. 119, 1989, págs. 295 a 378

Distancias Estadísticas

por
CARLES M. CUADRAS
Departament d'Estadística
Universitat de Barcelona

RESUMEN

Este artículo trata de la aplicación de las funciones de distancia a la estadística y al análisis de datos. Se exponen y discuten expresiones sobre distancias y coeficientes de similaridad entre individuos y poblaciones. Se incluyen también algunas aplicaciones a la biología, genética, psicología, arqueología, lingüística, análisis de la varianza, regresión y asociación estocástica.

Palabras clave: Distancia de Mahalanobis, distancia de Rao, distancia ultramétrica, coeficientes de similaridad, medidas de divergencia.

AMS 1980: 62H25; 62H30; 62P99.

1. INTRODUCCION

Las medidas de distancia entre poblaciones y dentro de poblaciones, han sido ampliamente utilizadas en numerosos campos científicos: antropología, agricultura, biología, genética, economía, lingüística, psicología, sociología, etc.

La noción de distancia estadística junto con sus propiedades constituyen una importante herramienta, tanto en la estadística matemática como en el análisis de datos. En el primer caso porque mediante una distancia se

pueden construir contrastes de hipótesis, estudiar propiedades asintóticas de estimadores, comparar parámetros, etc. En el segundo caso, porque la distancia es un concepto muy intuitivo, que permite obtener representaciones geométricas, fáciles de entender, ofreciendo al investigador una importante ayuda para interpretar la estructura de los datos.

En líneas generales consideramos dos clases de distancias estadísticas entre individuos y poblaciones:

a) Los n individuos de una población Ω quedan descritos por una matriz de datos $X(n \times p)$, donde p es el número de variables estadísticas (cuantitativas, cualitativas, binarias o categóricas). El número n suele ser el tamaño de una muestra de la población (ejemplo: $n = 75$ estudiantes universitarios), pero puede darse el caso de que Ω sea una población finita de n elementos (ejemplo: las $n = 50$ provincias españolas). Una distancia $\delta_{ij} = \delta(i,j)$ entre dos individuos o elementos i,j de Ω es una medida simétrica no negativa que cuantifica la diferencia entre ambos en relación con las variables. δ se puede sumarizar a través de la matriz de distancias

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} \ldots \delta_{1n} \\ \delta_{21} & \delta_{22} \ldots \delta_{2n} \\ \ldots\ldots\ldots\ldots\ldots \\ \delta_{n1} & \delta_{n2} \ldots \delta_{nn} \end{pmatrix} \qquad (1)$$

siendo $\delta_{ii} = 0$, $\delta_{ij} = \delta_{ji}$.

TABLA 1

Matriz de distancias genéticas entre 6 poblaciones de Drosophila subobscura: Heriot (H), Dalkeith (D), Groningen (G), Viena (V), Zurich (Z), Drobak (Dr)

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

b) Los individuos de cada población están caracterizados por un vector aleatorio $X = (X_1, \ldots, X_p)$, que sigue una distribución de probabilidad $f(x_1, \ldots, x_p; \theta)$. La distancia entre dos individuos i,j, caracterizados por los puntos x_i,

x_j de R^p, es una medida simétrica no negativa $\delta(x_i, x_j)$ que dependerá de θ. Análogamente la distancia entre dos poblaciones será una medida de divergencia $\delta(\theta_1, \theta_2)$ entre los parámetros que las caracterizan. También puede ser conveniente introducir una distancia $\delta(x_i, \theta)$ entre un individuo i y los parámetros θ.

Se pueden definir también distancias no paramétricas que miden la divergencia funcional entre funciones de densidad. En algunos casos están relacionadas con medidas de entropia.

Tanto en el caso *a)* como en el *b)*, en muchas aplicaciones interesa representar el conjunto Ω con la distancia δ, es decir, (Ω, δ), mediante un espacio geométrico modelo (V,*d*), donde V es un conjunto geométrico (espacio euclídeo, variedad de Riemann, grafo, curva, etc.) y *d* es una distancia sobre V. Según la técnica de representación utilizada (análisis de componentes principales, análisis de coordenadas principales, análisis de proximidades, análisis de correspondencias, análisis de cluster, etc.), la distancia *d* puede ser euclídea, ultramétrica, aditiva, no euclídea, riemaniana, etc.

La tabla 1 contiene un ejemplo de distancia genética entre un conjunto Ω de poblaciones europeas de *D. subobscura*. Aunque la distancia δ no es ultramétrica ni aditiva, puede representarse aproximadamente mediante un espacio ultramétrico (figura 3) o un espacio aditivo (figura 8).

1.1. Propiedades generales

Una distancia δ sobre un conjunto Ω es una aplicación de $\Omega \times \Omega$ en R, tal que a cada par (i,j) hace corresponder un número real $\delta(i,j) = \delta_{ij}$, cumpliendo algunas de las siguientes propiedades:

P. 1 $\quad \delta_{ij} \geq 0$

P. 2 $\quad \delta_{ii} = 0$

P. 3 $\quad \delta_{ij} = \delta_{ji}$

P. 4 $\quad \delta_{ij} \leq \delta_{ik} + \delta_{jk}$

P. 5 $\quad \delta_{ij} = 0$ si y sólo si $i = j$

P. 6 $\quad \delta_{ij} \leq max\{\delta_{ik}, \delta_{jk}\}$ (desigualdad ultramétrica)

P. 7 $\quad \delta_{ij} + \delta_{kl} \leq max\{\delta_{ik} + \delta_{jl}, \delta_{il} + \delta_{jk}\}$ (desigualdad aditiva)

P. 8 $\quad \delta_{ij}$ es euclídea

P. 9 $\quad \delta_{ij}$ es riemanniana

P.10 $\quad \delta_{ij}$ es una divergencia

Observaciones:

1) Una distancia debe cumplir por lo menos P.1, P.2, P.3. Cuando sólo cumple tales propiedades recibe el nombre de *disimilaridad.*

2) P.8 significa que existen dos puntos $x_i = (x_{i1}, \ldots, x_{im})'$, $x_j = (x_{j1}, \ldots, x_{jm})'$ de R^m tales que

$$\delta_{ij}^2 = (x_i - x_j)' \ (x_i - x_j) \tag{2}$$

es decir, δ_{ij} es la distancia euclídea entre los puntos x_i, x_j. Entonces (Ω, δ) puede representarse mediante el espacio euclídeo (R^m, d).

3) P.9 significa que (Ω, δ) puede ser representado mediante una variedad de Riemann (M, d_M).

4) P.6 $\Longrightarrow$ P.8 $\Longrightarrow$ P.4.

5) P.6 $\Longrightarrow$ P.7 $\Longrightarrow$ P.4.

6) Supongamos que hemos definido una medida de probabilidad μ sobre Ω. Entonces P.10 significa que δ es una expresión funcional sobre μ.

Algunas distancias poseen un calificativo propio según las propiedades que cumplen (cuadro 1). Todas estas propiedades las hemos referido a (Ω, δ). En algunos casos, como la distancia de Mahalanobis, δ verifica directamente las propiedades P.1 a P.4 y P.8 a P.10. Sin embargo, en general δ cumple solo aproximadamente algunas de las propiedades expuestas. Se trata entonces de representar (Ω, δ) a través de un modelo (V,d), aproximando δ a d, donde δ cumple con las suficientes propiedades requeridas. Por ejemplo, si podemos aproximar δ a d, siendo d una distancia ultramétrica, entonces (V,d) es un espacio ultramétrico y (Ω, δ) puede ser representado a través de un dendograma.

CUADRO 1

Calificación de una distancia según sus propiedades.

Disimilaridad: P.1, P.2, P.3
Distancia métrica: P.1, P.2, P.3, P.4, P.5
Distancia ultramétrica: P.1, P.2, P.3, P.6
Distancia euclídea: P.1, P.2, P.3, P.4, P.8
Distancia aditiva: P.1, P.2, P.3, P.7
Divergencia: P.1, P.2, P.10

2. DISTANCIAS SOBRE MATRICES DE DATOS

2.1. Similaridades

Una **similaridad** s sobre un conjunto Ω con n individuos, es una aplicación de $\Omega \times \Omega$ en R verificando las siguientes propiedades:

1) $0 \leq s_{ij} \leq 1$

2) $s_{ii} = 1$

3) $s_{ij} = s_{ji}$

La cantidad $s_{ij} = s(i,j)$ es una medida del grado de semejanza entre dos elementos i,j, en el sentido de que si ambos son muy parecidos entonces s_{ij} se aproxima 1. El concepto de similaridad es especialmente utilizado cuando sobre Ω se han introducido p características cualitativas, que se asocian a otras tantas variables binarias, que toman el valor 0 si la característica está ausente y el valor 1 si está presente. La matriz de incidencia individuos x características es una matriz $X = (x_{ik})$, cuyos elementos son ceros y unos. La similaridad entre dos individuos i,j queda bien descrita a través de a, b, c, d, siendo

$$a = \sum_{k=1}^{n} x_{ik}\, x_{jk}$$

$$b = \sum_{k=1}^{p} (1 - x_{ik})\, x_{jk}$$

$$c = \sum_{k=1}^{p} x_{ik}\, (1 - x_{jk})$$

$$d = \sum_{k=1}^{p} (1 - x_{ik})\, (1 - x_{jk})$$

es decir, a es el número de caracteres presentes comunes, b es el número de caracteres ausentes en i pero presentes en j, etc. Una similaridad s_{ij} es entonces una función de a, b, c.

$$s_{ij} = f\,(a, b, c)$$

tal que es creciente en a, decreciente y simétrica en b y c, vale $s_{ij} = 0$ si $b + c = p$, y $s_{ij} = 1$ si $a + d = p$.

Numerosos autores (Jaccard, 1900; Kulcynski, 1928; Russell y Rao, 1940; Sorensen, 1948; Sokal y Michener, 1958) han propuesto coeficientes de similaridad verificando tales propiedades:

$$s_{ij} = \frac{a}{a + b + c} \qquad \text{(Jaccard)}$$

$$s_{ij} = \frac{a}{p} \qquad \text{(Russell y Rao)}$$

$$s_{ij} = \frac{a + d}{p} \qquad \text{(Sokal y Michener)}$$

Sin embargo, otros coeficientes como

$$s_{ij} = \frac{a}{b + c} \qquad \text{(Kulczynski)}$$

cuyo rango es (0, ∞), no las cumplen.

La asociación entre los n elementos de Ω se expresa a través de una matriz de similaridades

$$S = \begin{pmatrix} s_{11} & s_{12} \ldots s_{1n} \\ s_{21} & s_{22} \ldots s_{2n} \\ \ldots\ldots\ldots\ldots\ldots \\ s_{n1} & s_{n2} \ldots s_{nn} \end{pmatrix} \qquad (3)$$

A menudo S se puede expresar operando en forma elemental la matriz X. Por ejemplo, para los coeficientes de Russell-Rao y Sokal-Michener, se tiene, respectivamente:

$$S_1 = (XX') \,/\, p$$

$$S_2 = [XX' + (J - X)\,(J - X)'] \,/\, p$$

En otros casos (Gower, 1971), la expresión es más compleja. Por ejemplo, para el coeficiente de Jaccard, se tiene:

$$S_3 = \frac{XX'}{p} + \frac{XX'}{p} * \left[\sum_{k=1}^{\infty} (J - X)\,(J - X)' * (J - X)\,(J - X)' \,/\, p^k \right]$$

indicando A * B la matriz cuyos elementos son $a_{ij} \times b_{ij}$ (producto matricial de Hadamard).

Para pasar de una disimilaridad a una distancia basta utilizar la fórmula

$$\delta_{ij} = 1 - s_{ij} \qquad (4)$$

Sin embargo, es más aconsejable utilizar

$$\delta_{ij} = \sqrt{1 - s_{ij}} \qquad (5)$$

En efecto, (5) da lugar a una distancia métrica, incluso euclídea, para la mayor parte de similaridades utilizadas en las aplicaciones (véase el cuadro 2). En general, dada una similaridad cualquiera (no necesariamente comprendida entre 0 y 1), podemos definir la distancia (Gower, 1966)

$$\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2\, s_{ij}} \qquad (6)$$

Si S es una matriz (semi) definida positiva, entonces δ_{ij} es euclídea (ver sección 4.1), y por lo tanto podremos representar (Ω, δ) a través del espacio euclídeo (R^m, d). En cambio, como se desprende del cuadro 2, pocas veces la distancia (4) es métrica y en ninguno de los casos presentados es euclídea.

Sobre los criterios que deben seguirse para elegir el coeficiente de similaridad (que dependerá del tipo de datos y el peso que se desea dar a las frecuencias *a, b, c, d*) véase Legendre y Legendre (1979), Gower y Legendre (1986).

CUADRO 2

Propiedades de algunos coeficientes de similaridad para variables binarias.

SIMILARIDAD	AUTOR	RANGO	$S \geq 0$	METRICA	EUCLIDEA
$\frac{a}{b+c}$	*Kulczynsky*	0, ∞	Sí		
$\frac{a}{a+b+c+d}$	*Russell y Rao*	0,1	Sí	Sí (Sí)	Sí
$\frac{a}{a+b+c}$	*Jaccard*	0,1	Sí	Sí (Sí)	Sí
$\frac{a+d}{a+b+c+d}$	*Sokal y Michener*	0,1	Sí	Sí (Sí)	Sí
$\frac{a}{a+2(b+c)}$	*Anderberg*	0,1	Sí	Sí (Sí)	Sí
$\frac{a+d}{a+2(b+c)+d}$	*Rogers y Tanimoto*	0,1	Sí	Sí (Sí)	Sí
$\frac{a}{a+\frac{1}{2}(b+c)+d}$	*Sorensen*	0,1	Sí	Sí (No)	Sí

(Sigue)

(1) $S \geq 0$ significa que la matriz de similaridades es (semi) definida positiva.

(2) La propiedad métrica se refiere a la distancia $d_{ij} = \sqrt{s_{ii}+s_{jj}-2s_{ij}}$ y a la distancia $\bar{d}_{ij} = 1 - s_{ij}$ (entre paréntesis).

(3) Ninguna de las distancias $\bar{d}_{ij}$ es euclídea.

CUADRO 2

(Final)

Propiedades de algunos coeficientes de similaridad para variables binarias.

SIMILARIDAD	AUTOR	RANGO	$S \geq 0$	METRICA	EUCLIDEA
$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$	*Sneath y Sokal*	0,1	No	Sí (No)	No
$\frac{a-(b+c)+d}{a+b+c+d}$	*Harman*	-1,1	Sí	Sí (Sí)	Sí
$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$	*Kulczynski*	0,1	No	No (No)	No
$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{c+d}+\frac{d}{b+d})$	*Anderberg*	0,1	No	No (No)	No
$\frac{a}{\sqrt{(a+b)(a+c)}}$	*Ochiai*	0,1	Sí	Sí (No)	Sí
$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$		0,1	Sí	Sí (No)	Sí
$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	*Pearson*	-1,1	Sí	Sí (No)	Sí
$\frac{ad-bc}{ad+bc}$	*Yule*	-1,1	No	No (No)	No

(1) $S \geq 0$ significa que la matriz de similaridades es (semi) definida positiva.

(2) La propiedad métrica se refiere a la distancia $d_{ij} = \sqrt{s_{ii}+s_{jj}-2s_{ij}}$ y a la distancia $\bar{d}_{ij} = 1-s_{ij}$ (entre paréntesis).

(3) Ninguna de las distancias $\bar{d}_{ij}$ es euclídea.

2.2. Distancias sobre datos cuantitativos

Supongamos que los valores observados para p variables aleatorias sobre n individuos, son cuantitativos, formando una matriz de datos

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \qquad (7)$$

Entonces cada individuo i puede representarse como un punto $x_i \in R^p$. La distancia más familiar entre dos individuos i,j es la distancia euclídea (2), es decir,

$$d_2(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{8}$$

Tal distancia es un caso particular de las distancias de Minkowski

$$d_q(i,j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right)^{1/q} \qquad 1 < q < \infty \tag{9}$$

que verifican P.1, P.2, P.3 y P.4. No son distancias euclideas, salvo el caso $q = 2$. Para $q = 1$ se tiene

$$d_1(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

que se denomina distancia "ciudad". Una expresión límite de (9) es

$$d_\infty(i,j) = max \{ |x_{ik} - x_{jk}| \}$$

llamada distancia "dominante". Véase la figura 1.

Volviendo de nuevo a la distancia euclídea (8), vemos que tiene algunos inconvenientes: *a)* no está acotada; *b)* no es invariante por cambios de escala; *c)* considera la p variables estocásticamente independientes.

Se han propuesto diferentes modificaciones sobre $d_2(i,j)$ a fin de evitar tales inconvenientes. Una primera modificación consiste, simplemente, en dividir por el número de variables, es decir, introducir la distancia (al cuadrado)

$$d_2^2(i,j) = \frac{1}{p} d_2(i,j)$$

Orloci (1967) ha propuesto otras transformaciones, que permiten acotar la distancia, y que en general están basadas en métricas geodésicas sobre la hiperesfera de radio 1 (ver sección 5.4)

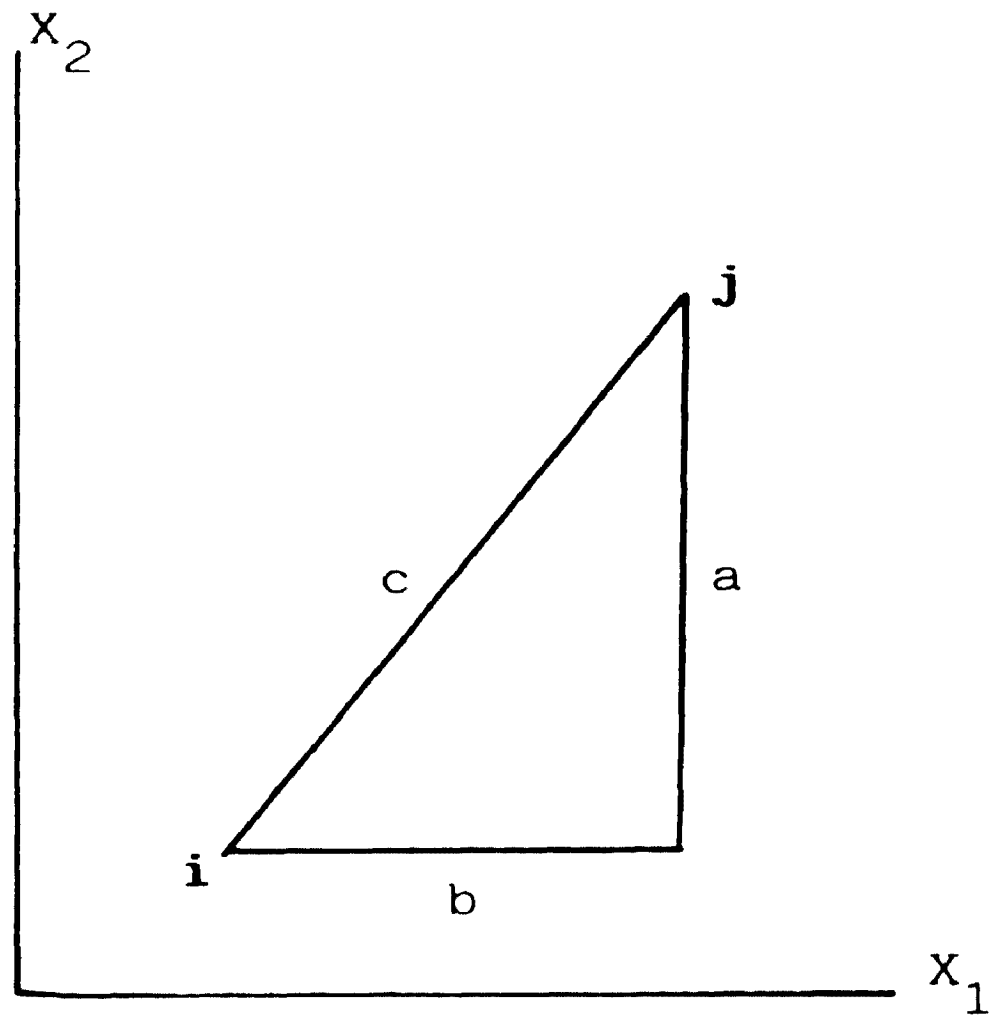


Fig.1. Distancias de Minkowski en dimensión p=2. Distancia euclídea d_2=c. Distancia "ciudad": $d_1 = a + b$. Distancia "dominante": d_∞ = a.

La invarianza por cambios de escala se resuelve dividiendo cada término $(x_{ik} - x_{jk})$ por la desviación típica de la variable k, lo que nos lleva a la distancia de K. Pearson (sección 3.1). El inconveniente *c)* puede resolverse introduciendo la distancia de Mahalanobis (sección 3.2) que tiene en cuenta las correlaciones entre las variables, y por tanto la redundancia existente entre las mismas.

Otras variantes de (9) son la métrica de Canberra

$$\sum_{K=1}^{p} \frac{| x_{ik} - x_{jk} |}{| x_{ik} | + | x_{jk} |}$$

que ha sido utilizada por Lance y Williams (1967), y el coeficiente de divergencia (Clark, 1952).

$$\sqrt{\frac{1}{p} \sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}} \right)^2}$$

Las propiedades métricas y euclídeas de estas y otras distancias para matrices de datos positivos, se consideran en el cuadro 3. Sobre los criterios que deben seguirse para la elección de la distancia, véase Legendre y Legendre (1979), Gower y Legendre (1986), y Legendre, Dallot y Legendre (1985). En este último trabajo, las distancias son clasificadas en tres tipos, según el peso que se quiera dar a las diferencias entre variables con diferentes rangos de variación (suponiendo variables dimensionalmente homogéneas y no negativas).

CUADRO 3

Distancias y disimilaridades para datos cuantitativos (no negativos).

DISTANCIA	AUTOR	METRICA	EUCLIDEA
$\left(\sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \right)^{1/2}$	*Euclides*	SI	SI
$\left(\sum_{k=1}^{n} (x_{ik} - x_{jk})^q \right)^{1/q}$	*Minkowski*	SI	NO
$\left(\sum_{k=1}^{n} \left(\frac{x_{ik} - x_{jk}}{s_k} \right)^2 \right)^{1/2}$	*K. Pearson*	SI	SI
$\sum_{k=1}^{n} \frac{\lvert x_{ik} - x_{jk} \rvert}{\lvert x_{ik} \rvert + \lvert x_{jk} \rvert}$	*Canberra*	SI	NO
$\left(\frac{1}{n} \sum_{k=1}^{n} \frac{(x_{ik} - x_{jk})^2}{(x_{ik} + x_{jk})^2} \right)^{1/2}$	*Clark*	SI	SI

3. DISTANCIA DE MAHALANOBIS

3.1. Distancia euclídea normalizada

Dada una matriz de datos X en los términos de la sección 2.2, la distancia euclídea normalizada K (i,j) es la raíz cuadrada de

$$K^2 (i,j) = \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{\sigma_k^2} \qquad (10)$$

donde σ_k^2 es la varianza de la variable k. La distancia K (i,j) es invariante por cambios de escala y es una distancia entre individuos relacionada con el coeficiente de semejanza racial introducido por K. Pearson (1926), que ha sido utilizado en antropología para diferenciar cráneos. Dadas dos poblaciones representadas por (μ_1, Σ) y (μ_2, Σ), donde μ_1, μ_2 son los vectores de medias y Σ es la matriz de covarianzas (común) en relación a p variables aleatorias, el coeficiente de semejanza racial, también llamada distancia de K. Pearson, es proporcional a

$$K^2 = (\mu_1 - \mu_2)' \, [diag \, (\Sigma)]^{-1} \, (\mu_1 - \mu_2) \qquad (11)$$

K es también invariante por cambios de escala y puede considerarse un precedente de la distancia de Mahalanobis (14). Ambas distancias han sido comparadas por diversos autores. Véase Mardia (1977).

3.2. Definición y propiedades de la distancia de Mahalanobis

Supongamos que una población Ω está caracterizada por p variables aleatorias, siendo $\mu = (\mu_1, ..., \mu_p)'$ el vector de medias y Σ la matriz de covarianzas no singular. La distancia de Mahalanobis $M(i,j)$ entre dos individuos i,j, representados por los vectores x_i, x_j, se define como

$$M^2(i,j) = (x_i - x_j)' \Sigma^{-1} (x_i - x_j) \tag{12}$$

Análogamente, la distancia entre un individuo i y la población Ω es

$$M^2(i,\Omega) = (x_i - \mu)' \Sigma^{-1} (x_i - \mu) \tag{13}$$

La distancia entre dos poblaciones Ω_1, Ω_2 es

$$M^2(\Omega_1, \Omega_2) = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \tag{14}$$

Desde luego, estrictamente la distancia es $M(i,j)$, $M(i, \Omega)$ y $M(\Omega_1, \Omega_2)$ aunque es preferible manejarla elevada al cuadrado. La distancia (14) fue introducida por Mahalanobis (1936), alegando criterios heurísticos. Sin embargo, aparece de forma natural por diferentes caminos, como comentamos seguidamente.

a) Sea $E = < X_1, ..., X_p >$ el espacio vectorial generado por p variables aleatorias. Sea Σ la matriz de covarianzas no singular. Considerando E^*, espacio dual de E, a cada individuo i de Ω le podemos hacer corresponder la forma lineal i^* de E^* tal que $i^*(X) = X(i)$, donde $X \in E$. Luego, a través de las variables $X_1, ..., X_p$ podemos proyectar Ω en E^*. Entonces, como la métrica natural en E^* viene dada por la matriz inversa Σ^{-1} la distancia en E^*, es decir, la distancia entre individuos, es $M(i,j)$.

b) La función de densidad normal multivariante $N_p(\mu, \Sigma)$ es

$$f(x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)\right\}$$

es decir, es una función exponencial de la distancia de Mahalanobis entre x y μ. Obsérvese que la distribución de esta distancia es ji-cuadrado.

c) Consideremos g poblaciones $\Omega_1, ..., \Omega_g$, $g \geq 2$. Supongamos asociada a cada población Ω_j una distribución $N_p(\mu_j, \Sigma)$ y que sobre un individuo

ω tenemos la información multivariante x. En análisis discriminante es conocida la regla de la máxima verosimilitud para asignar ω a Ω_j, donde $j \in \{1,...,g\}$. Es fácil demostrar que esta regla es equivalente a asignar ω a la población Ω_j tal que la distancia de Mahalanobis $M(\omega, \Omega_j)$ es mínima (Mardia *et al*, 1979).

d) Consideremos el modelo estadístico $N_p(\mu, \Sigma)$ de las distribuciones normales, con Σ fijo, $\mu \in R^p$. Dotando al espacio paramétrico R^p de estructura de variedad riemanniana entonces la distancia de Mahalanobis $M(\mu_1, \mu_2)$ es una distancia geodésica entre dos puntos de la variedad (sección 5.4).

Por otra parte, la distancia de Mahalanobis goza de interesantes propiedades, que vamos a comentar para la versión (12).

1) $M(i,j) \geq 0$ y $M(i,j) = 0$ si y sólo si $x_i = x_j$.

2) $M(i,j) = M(j,i)$.

3) $M(i,j) \leq M(i,k) + M(j,k)$.

4) $M(i,j)$ es invariante por transformaciones lineales no singulares de las variables. En particular, es invariante por cambios de escala.

5) Introduciendo el cambio de variable $y = \Sigma^{-\frac{1}{2}} x$, es fácil ver que la distancia de Mahalanobis es euclídea.

6) Es una distancia normalizada, que puede expresarse en unidades de desviación típica. Además, tiene en cuenta las correlaciones entre las variables, es decir, la redundancia entre las variables.

7) Indiquemos por M_p la distancia basada en p variables y por M_{p+q} la distancia basada en $p+q$ variables, conteniendo estas $p+q$ a las p primeras. Entonces

$$M_p \leq M_{p+q}$$

8) Sean M_p, M_q las distancias tomando las variables $X = (X_1, ..., X_p)$ $Y = (Y_1, ..., Y_q)$. Supongamos que las variables X están incorrelacionadas con las variables Y. Entonces

$$M^2_{p+q} = M^2_p + M^2_q$$

Las propiedades **7)** y **8)** se ilustran en la figura 2, en la que puede apreciarse que la distancia de Mahalanobis es mayor para dos variables que para una sola, disminuyendo a medida que aumenta la correlación entre las variables.

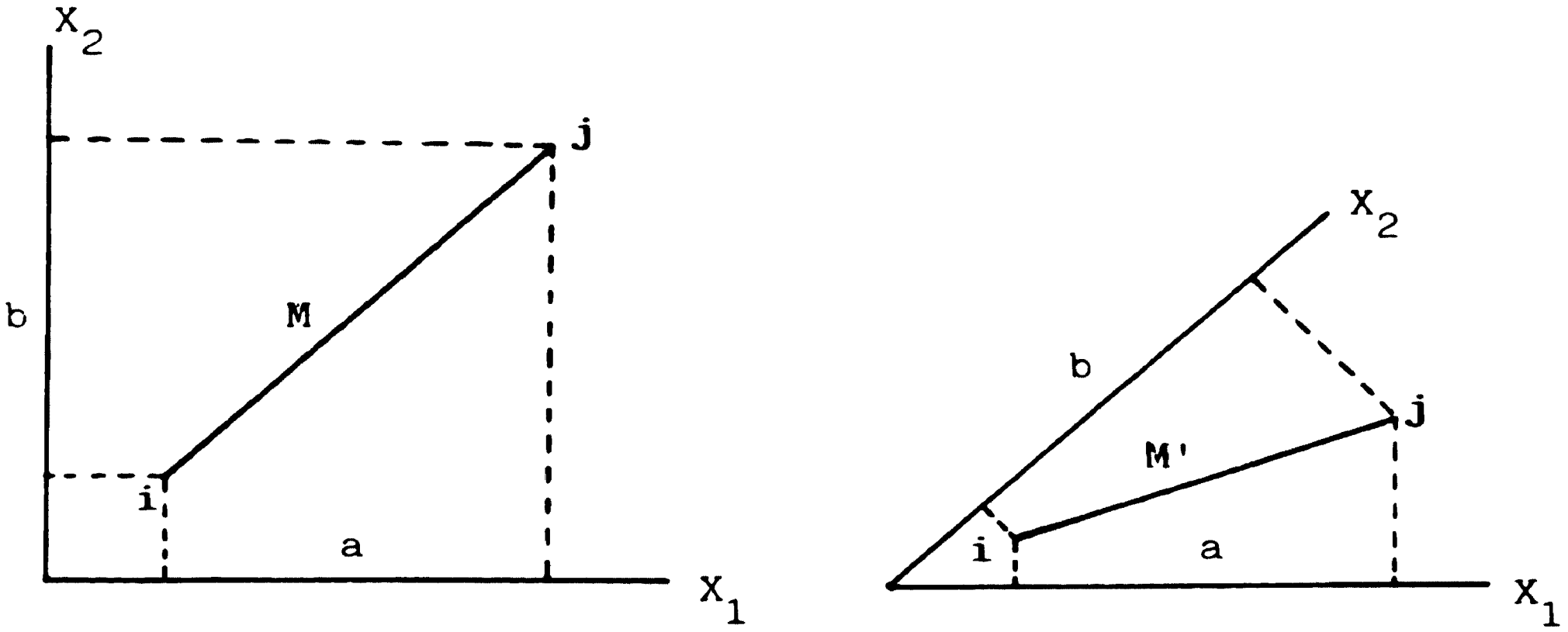


Fig. 2. Distancia de Mahalanobis en el caso de variables incorrelacionadas y de variables correlacionadas. Obsérvese que M > M' > *a* .

Es de esperar, por otra parte, que en las aplicaciones la distancia sea estable

$$\lim_{p \to \infty} M_p = \alpha < \infty$$

es decir, si el número de variables *p* es grande, la distancia de Mahalanobis no aumentará o al menos tenderá a un valor finito α, debido a que las variables añadidas serán redundantes respecto a las *p* anteriores.

3.3. Distancias singulares

Supongamos *ran* $\Sigma = r < p$ y que $\mu_1 - \mu_2$ es combinación lineal de las columnas de Σ. Se define la distancia de Mahalanobis singular entre las poblaciones de vectores de medias μ_1, μ_2 y matriz de covarianzas Σ como

$$M^2 = (\mu_1 - \mu_2)' \, \Sigma^- \, (\mu_1 - \mu_2)$$

donde Σ^- es una *g*-inversa de Σ. La distancia singular tiene aplicaciones a la genética (sección 7.2), al análisis factorial y a la comparación de curvas de crecimiento. Véase Rao (1954) y Mardia (1977).

4. DISTANCIAS EUCLIDEAS, ULTRAMETRICAS Y ADITIVAS

4.1. Caracterización de una distancia como distancia euclídea

Las distancias de K. Pearson, Mahalanobis y Minkowski para $q=2$, son euclídeas por propia definición. Sin embargo, en otros casos, la decisión sobre si la distancia δ definida en un conjunto Ω es euclídea o no, no puede tomarse tan directamente. Es entonces cuando adquiere importancia fundamental el teorema 1, puesto que nos permite representar (Ω, δ) a través de un espacio euclídeo (R^m, d).

Sea Ω un conjunto formado por n elementos y sea $\delta_{ij} = \delta(i,j)$ una distancia sobre Ω. Diremos que la matriz de distancias $\Delta = (\delta_{ij})$ es euclídea m-dimensional si existen n puntos $x_1, x_2, \ldots, x_n$ se un espacio euclídeo R^m tales que

$$\delta_{ij}^2 = (x_i - x_j)' (x_i - x_j) \tag{15}$$

En otras palabras, Δ es euclídea si existe una matriz de datos X cuyas filas son $x_1', x_2', \ldots, x_n'$

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix} \tag{16}$$

Verificándose

$$\delta_{ij}^2 = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2$$

Consideremos ahora la matriz identidad I_n, la matriz J_n de orden $n \times n$ cuyos elementos son todos iguales a 1, la matriz

$$H = I_n - \frac{1}{n} J_n$$

y la matriz $A = (a_{ij})$, siendo $a_{ij} = -\frac{1}{2} \delta_{ij}^2$, para $i,j = 1,2,\ldots,n$. Sea también

$$B = H A H \tag{17}$$

Obsérvese que los elementos de la matriz $B = (b_{ij})$ verifican

$$b_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}$$

siendo $\bar{a}_{i\cdot}$, $\bar{a}_{\cdot j}$ las medias de la fila i y de la columna j respectivamente, y $\bar{a}$ la media de los n^2 elementos de A.

Teorema 1

La matriz de distancias Δ es euclídea m-dimensional si y sólo si B es semidefinida positiva de rango m.

Demostración: Véase Seber (1984).

Si B es semidefinida positiva y de rango m, entonces existe una matriz $X(n \times m)$ tal que

$$B = X X' \tag{18}$$

Cualquier matriz X cumpliendo (18), puede tomarse como matriz de datos, es decir, tal que sus filas contengan las coordenadas de los puntos de R^m cuyas intradistancias reproduzcan Δ.

Este importante resultado fue primeramente obtenido por Schoenberg (1935), fecha en verdad tardia tratándose de una propiedad fundamental de la geometría euclídea. Cuando X proviene de la descomposición espectral de B, se obtiene la solución del análisis de proximidades ("metric multidimensional scaling"), utilizada por Torgerson (1958) en psicología, y replanteada con mayor claridad por Gower (1966), con el nombre de análisis de coordenadas principales. Desde entonces ha sido ampliamente utilizada también en ecología, botánica, etc. Gower (1982) propone y discute ciertas generalizaciones del Teorema 1.

Sea ahora S una matriz de similaridad (semi) definida positiva. Luego $S = X X'$ para alguna matriz $X(n \times m)$ y por tanto $s_{ij} = x_i' . x_j$, donde $x_1', \ldots, x_n'$ representan las filas de X. Entonces, tomando la distancia δ_{ij} definida por S, como

$$\delta_{ij}^2 = s_{ii} + s_{jj} - 2\, s_{ij} = (x_i - x_j)'\ (x_i - x_j) \tag{19}$$

resulta que δ_{ij} es una distancia euclídea. En consecuencia, para conseguir una representación euclídea de Ω de modo que la proximidad entre puntos sea el equivalente geométrico de la similaridad entre individuos, es preferible utilizar (5) ó (6) en vez de (4).

Supongamos ahora que la matriz de distancias Δ no es euclídea. Entonces B tiene valores propios negativos, no existe ninguna matriz X verificando (18) y por lo tanto Ω no puede representarse en un espacio euclídeo. Este problema se presenta con frecuencia en psicología. Para solventarlo, se debe aproximar δ a una distancia euclídea d, a fin de que (Ω, δ) admita

una representación euclídea aproximada. Después de los trabajos de Shepard (1962 *a,b*) y Kruskal (1964 *a,b*), quedó establecido que δ debía transformarse en $d = f(\delta)$, donde f es una función monótona no decreciente, a fin de que se conservara la preordenación de las distancias originales, es decir,

$$\delta_{ij} \leq \delta_{i'j'} \iff d_{ij} \leq d_{i'j'} \tag{20}$$

Las transformaciones pueden ser algebraicas o numéricas. Entre las primeras, las más sencillas son (suponiendo $i \neq j$) :

$d_{ij} = \delta_{ij} + c$ (Cooper, 1972; Cailliez, 1983) ,

$d_{ij}^2 = \delta_{ij}^2 - 2a$ (Lingoes, 1971; Mardia, 1978) ,

$d_{ij} = a\,\delta_{ij} + b$ (Cuadras y Ruiz-Rivas, 1980) .

Se demuestra que si δ_{ij} no es euclídea, existen constantes adecuadas *a, b, c* tales que d_{ij} es una distancia euclídea. Respecto a las segundas, existe abundante literatura sobre el tema de transformar, por procedimientos numéricos, una distancia no euclídea en euclídea. Véase De Leeuw y Heiser (1982), Cuadras *et al.* (1985).

4.2. Distancias ultramétricas

En esta sección estudiamos la propiedad

P.6 $\quad \delta_{ij} \leq max\{\delta_{ik}, \delta_{jk}\} \quad \forall\ i,j,k$

asi como sus principales consecuencias. Cuando se cumple P.6 se dice que (Ω,δ) es un espacio ultramétrico. Sin embargo, dificilmente una distancia calculada a partir de unos datos estadísticos, cumplirá una propiedad tan restrictiva como la desigualdad ultramétrica. En realidad, se trata de aproximar δ a una ultramétrica δ_u, y representar aproximadamente (Ω,δ) a través de un espacio ultramétrico, en el sentido de la sección 1.

El axioma ultramétrico para una distancia fue introducido por primera vez por M. Krassner en 1930, en ciertas investigaciones de la teoría de números y series formales, demostrando que ciertas distancias, como la distancia *p-ádica* entre números enteros y la distancia entre los rangos de dos series, cumplían la propiedad ultramétrica. Posteriormente Benzecri (1965), Jardine, Jardine y Sibson (1967) y Johnson (1967), establecieron la relación entre distancia ultramétrica y jerarquia indexada. Desde entonces, las

distancias con la propiedad P.6 juegan un papel fundamental en análisis de datos.

Supongamos que (Ω,δ) es un espacio ultramétrico. Es bien conocido (Jonhson, 1967; Benzecri, 1976), que puede asociarse a Ω una jerarquia indexada (C,α) donde C es una colección de subconjuntos ("clusters") de Ω, α es un índice sobre C, unívocamente determinado por la distancia ultramétrica δ, con ciertas propiedades de monotonía. La idea principal es que en un espacio ultramétrico, la noción de proximidad entre dos individuos, en el sentido de que su distancia es inferior a un valor $x > 0$ dado, define una relación de equivalencia en Ω y por tanto una partición de Ω. Es decir, la relación $i \sim j$ si y sólo si $\delta_{ij} \leq x$, es de equivalencia y define una partición ("clustering") de Ω para cada nivel x. Aumentando x se obtiene particiones progresivamente menos finas, que engloban a las anteriores, formando una estructura jerárquica. La representación geométrica de (Ω,δ) se lleva a cabo a través de un dendograma, que refleja métrica (a través de α) y jerarquicamente la estructura de Ω (figura 3).

La aplicación de estos conceptos para resolver problemas de clasificación, taxonomía y sistemática, es bien conocida (Jardine y Sibson, 1971; Sneath y Sokal, 1973; Benzecri, 1976; Cuadras, 1981). Por otra parte, es bastante frecuente en las aplicaciones, la doble representación de (Ω,δ), bien a través de un espacio euclídeo, bien a través de un espacio ultramétrico (véase, por ejemplo, Escarré, 1972; Canton y Sancho, 1976, D'Andrade *et al.*, 1972; Rapoport y Fillenbaum, 1972; Del Castillo, 1986) hasta el punto de que la relación entre ambas representaciones ha interesado a los estadísticos.

La pregunta básica es la siguiente: ¿cuál es la conexión que existe entre la representación a lo largo de unos ejes de coordenadas y a través de un dendograma?. Un primer paso sobre este tema fue dado por Holman (1972), que demostró el siguiente:

Teorema 2

Si (Ω,δ) es un espacio ultramétrico, Ω tiene n elementos y $\delta_{ij} \neq 0$ para todo $i \neq j$, entonces la matriz de distancias $\Delta = (\delta_{ij})$ es euclídea $(n-1)$-dimensional.

De este resultado, que habría sido conjeturado por Gower (1971), se han dado diferentes demostraciones: Gower y Bandfield (1975), Cailliez y Pagés (1976), Cuadras y Carmona (1983). Otros autores han relacionado ambas clases de representaciones. Ohsumi y Nakamura (1981) estudian la relación entre la formación de "clusters" y los valores propios de la matriz asociada a Δ (teorema 1). Carroll (1976) introduce estructuras de árbol

como modelos intermedios entre los modelos espaciales y el esquema de representación jerárquico, describiendo un algoritmo para ajustar una estructura de árbol a (Ω,δ). Por otra parte, Pruzansky *et al.* (1982) estudian y comparan representaciones en el plano euclídeo y a través de un árbol aditivo (ver sección siguiente), proponiendo índices y criterios de ajuste para decidir el modelo más apropiado.

Por lo demás, no parece fácil interpretar el teorema 2. Holman (1972) observa que mientras un conjunto finito en el que hay definida una distancia ultramétrica, puede representarse integramente en el plano mediante un dendograma, la dimensión exacta en una representación euclídea vale exactamente $(n\text{-}1)$, luego parece estar reñida con una reducción de la dimensión, es decir, con una representación tomando (por ejemplo), los 2 primeros ejes principales. Sin embargo, Critchley (1985) demuestra que una distancia ultramétrica puede llegar a estar arbitrariamente próxima a una distancia euclídea m-dimensional, incluyendo $m=1$.

Cuadras (1983), Cuadras y Oller (1987) discuten este problema analizando las coordenadas euclídeas $X(n \times (n\text{-}1))$ que verifican (18), obtenidas por descomposición espectral de B, es decir, las coordenadas principales asociadas a una matriz de distancias ultramétricas Δ. La estructura de los vectores propios de B está relacionada con la formación de "clusters" y algunos valores propios pueden obtenerse explicitamente.

Existe una partición

$$\Omega = \Omega_1 + \ldots + \Omega_r + \Omega_{r+1} + \ldots + \Omega_k \tag{21}$$

tal que cada Ω_i, $1 \leq i \leq r$, contiene $n_i > 1$ elementos, mientras que cada Ω_i, $r < i \leq k$, contiene $n_i=1$ elementos, siendo

$$n = \sum_{i=1}^{k} n_i$$

el número de elementos de Ω. La partición (21) está formada por r "clusters" maximales de elementos equidistantes y $(n\text{-}r)$ elementos aislados (pudiendo verificarse $r \geq 1$ y $r=k$). Si $h_1 \leq \ldots \leq h_r$ son las distancias (comunes) en $\Omega_1, \ldots, \Omega_r$ respectivamente, entonces

$$\lambda_1 = \frac{1}{2} h_1^2 \leq \ldots \leq \lambda_r = \frac{1}{2} h_r^2$$

son valores propios de B y cada λ_i tiene multiplicidad $(n_i\text{-}1)$. Además λ_1 es el menor valor propio de B. Por otra parte, la matriz de coordenadas principales puede ser arreglada en la forma

$$X = (X_0, X_1, \ldots, X_r)$$

donde X_i ($n \times (n_i - 1)$) contiene las coordenadas principales asociadas a λ_i que discriminan solamente los objetos contenidos en Ω_i, siendo $1 \leq i \leq r$. Las demás coordenadas principales ubicadas en X_0 permiten representar, a lo largo de $(k-1)$ dimensiones, los "clusters" que constituyen la partición (21). Luego, la dimensión $(n-1)$ necesaria según el teorema de Holman, puede quedar reducida a $(k-1)$. Otra consecuencia es que la utilización de los dos primeros ejes principales (asociados a los dos primeros valores propios de B), puede ser inadecuada puesto que, en ciertos casos, podrían no discriminar adecuadamente los diferentes "clusters" de Ω .

Ejemplo: Consideremos el conjunto $\Omega = \{ 1, \ldots, 7 \}$ y la matriz de distancias ultramétricas sobre Ω .

$$\Delta = \begin{pmatrix} 0 & 1 & 2 & 2 & 4 & 4 & 5 \\ & 0 & 2 & 2 & 4 & 4 & 5 \\ & & 0 & 1 & 4 & 4 & 5 \\ & & & 0 & 4 & 4 & 5 \\ & & & & 0 & 4 & 5 \\ & & & & & 0 & 5 \\ & & & & & & 0 \end{pmatrix}$$

Los valores propios de la matriz B asociada a Δ son

$$\lambda_1 = 21 \quad \lambda_2 = 16{,}43 \quad \lambda_3 = 3{,}5 \quad \lambda_4 = \lambda_5 = \lambda_6 = \frac{1}{2}$$

La figura 4 contiene la representación de la distancia a través de un dendograma. La partición (21) es en este caso

$$\Omega = \{ 1,2 \} + \{ 3,4 \} + \{ 5,6 \} + \{ 7 \}$$

Las figuras 5,6,7, contienen las representaciones euclídeas a lo largo de los diferentes ejes principales. Obsérvese que la representación tradicional tomando los dos primeros ejes principales (figura 5), coloca los elementos 1 a 4 en el mismo punto. Sin embargo, tomando los ejes primero y tercero, conseguimos discriminar el "cluster" $\{ 1,2 \}$ del $\{ 3,4 \}$, es decir, la figura 6 refleja la partición anterior. Por otra parte, los elementos dentro de los "clusters" $\{ 1,2 \}$, $\{ 3,4 \}$, $\{ 5,6 \}$, quedan diferenciados a lo largo de los ejes cuarto, quinto y sexto (figura 7).

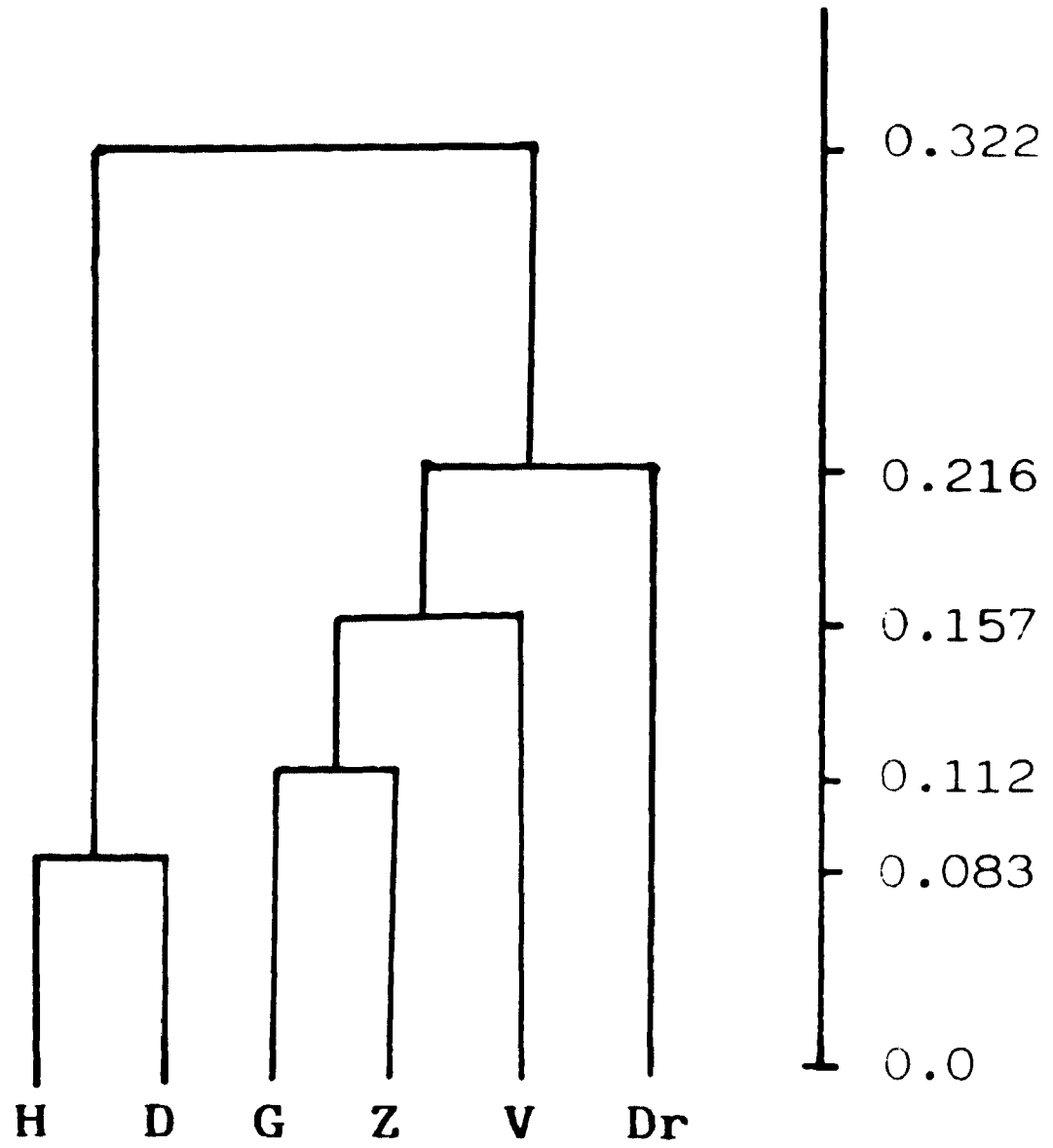


Fig. 3.– Representación mediante un dendograma (distancia ultramétrica) de las poblaciones cuya matriz de distancias genéticas viene dada en la tabla 1.

Figura 4

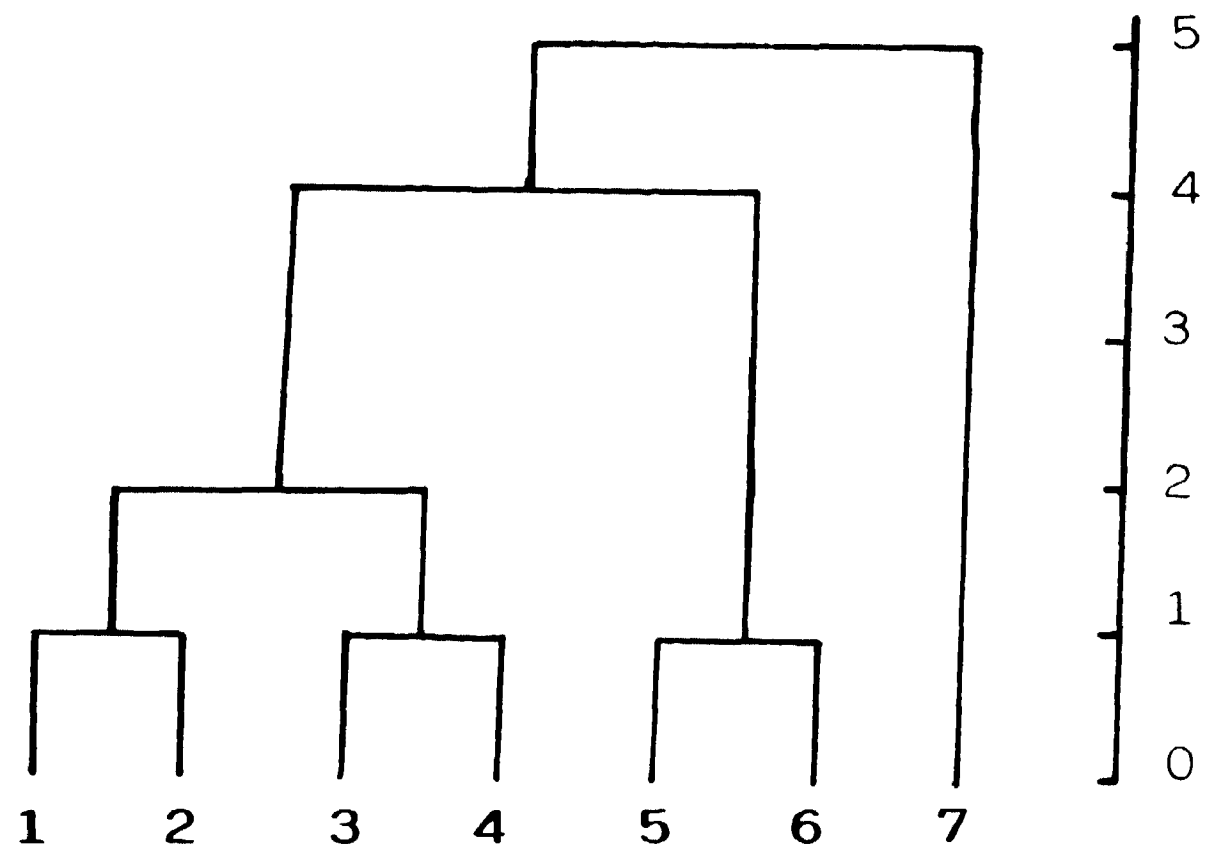


Figura 5

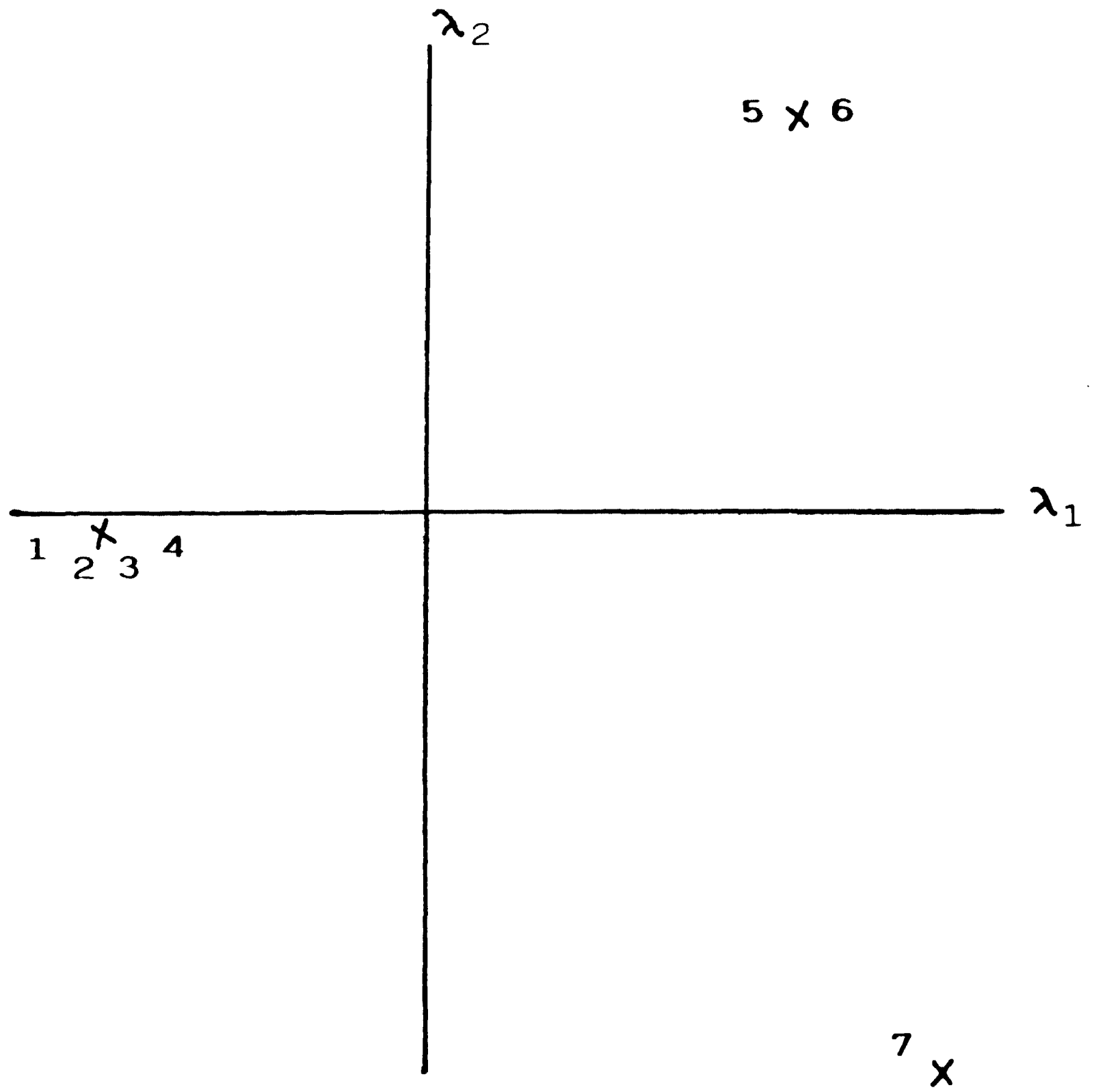


Figura 6

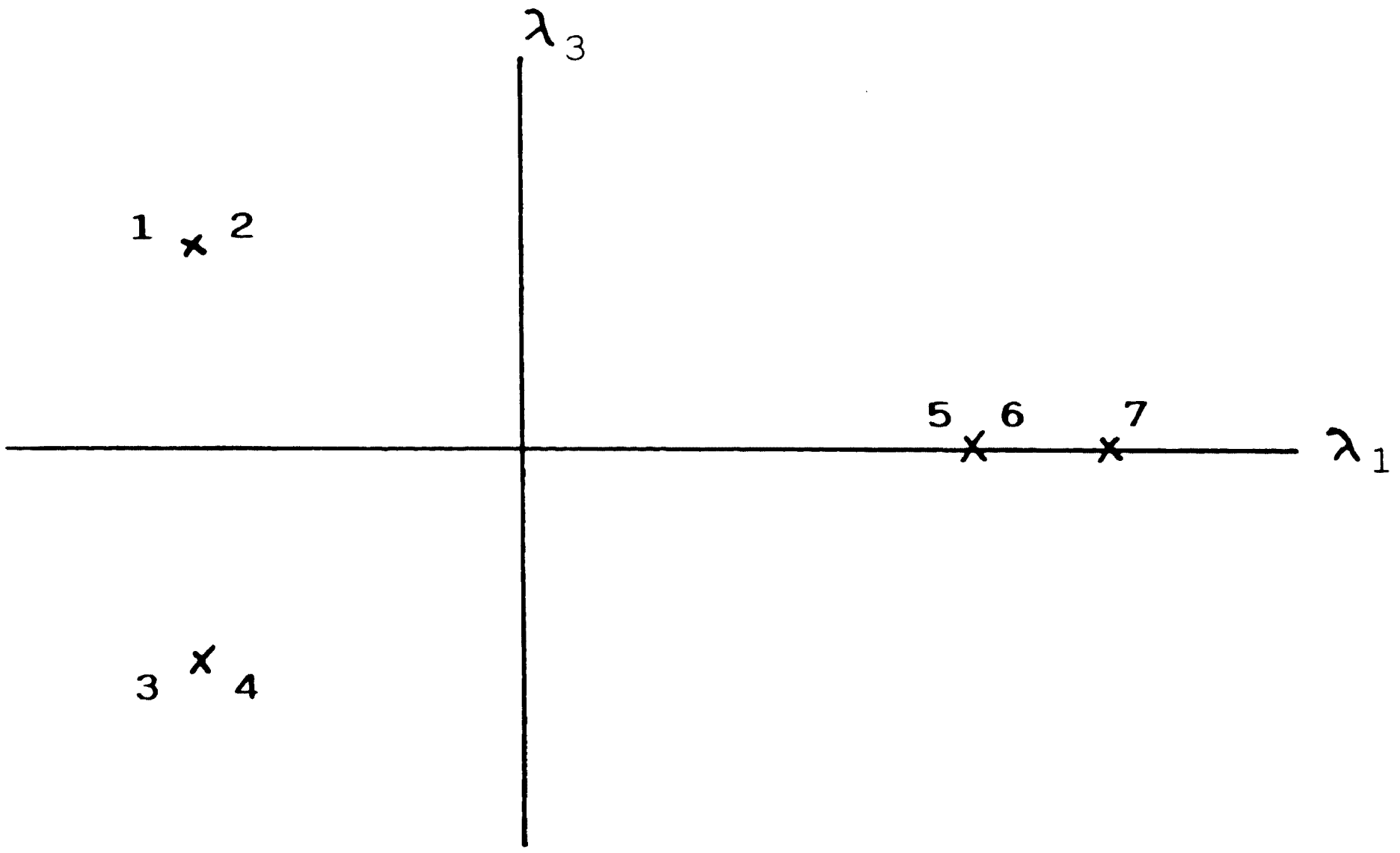
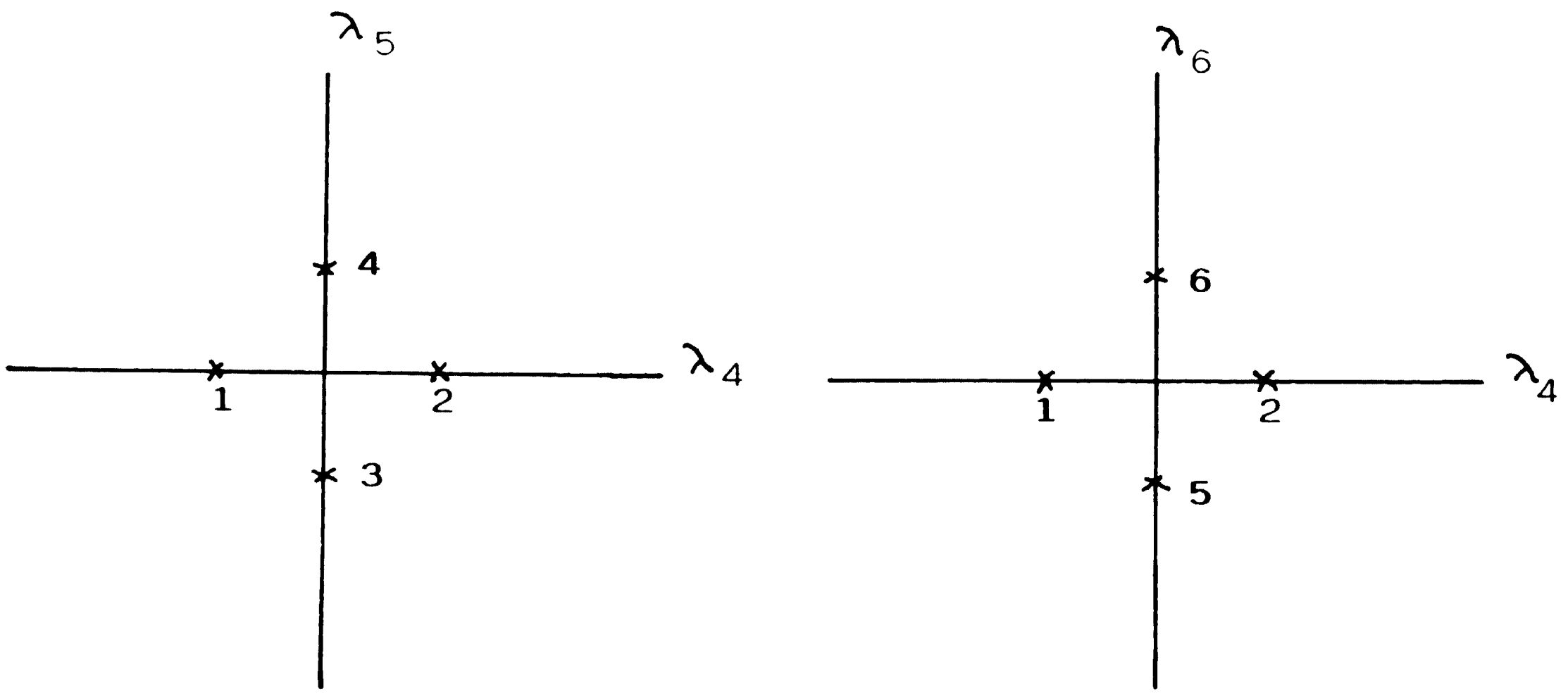


Figura 7



4.3. Distancias aditivas

Una distancia δ_{ij} es aditiva si verifica la desigualdad aditiva (también llamada axioma de los cuatro puntos)

P.7 $$\delta_{ij} + \delta_{kl} \leq max \{ \delta_{ik} + \delta_{jl} , \delta_{il} + \delta_{jk} \} \quad \forall\, i,j,k,l$$

Cuando se cumple P.7 diremos que (Ω,δ) es un espacio aditivo. De hecho, una distancia estadística no cumplirá, en general, la desigualdad aditiva, sino que se trata de aproximar δ a una distancia aditiva δ_a, a fin de poder aproximar (Ω,δ) a través de un espacio aditivo, en el sentido de la sección1.

El interés por la desigualdad aditiva surge al considerar la desigualdad ultramétrica como demasiado restrictiva para ajustarle una distancia estadística. Se puede probar la siguiente implicación entre las desigualdades P.4, P.6, P.7:

$$\text{ultramétrica} \Longrightarrow \text{aditiva} \Longrightarrow \text{triangular}$$

Por lo tanto, un espacio ultramétrico es un caso particular de un espacio aditivo. Si Ω es un conjunto finito y δ es una distancia métrica no se conoce ninguna forma de representación de (Ω,δ) a través de una estructura geométrica conocida. Pero si δ es ultramétrica entonces la representación puede hacerse en forma de dendograma. Y si δ es aditiva, Ω puede representarse a través de los extremos de un grafo simplemente conexo, tomando como distancia la longitud del camino que los une. Esta es la llamada representación a través de un árbol aditivo (figura 8).

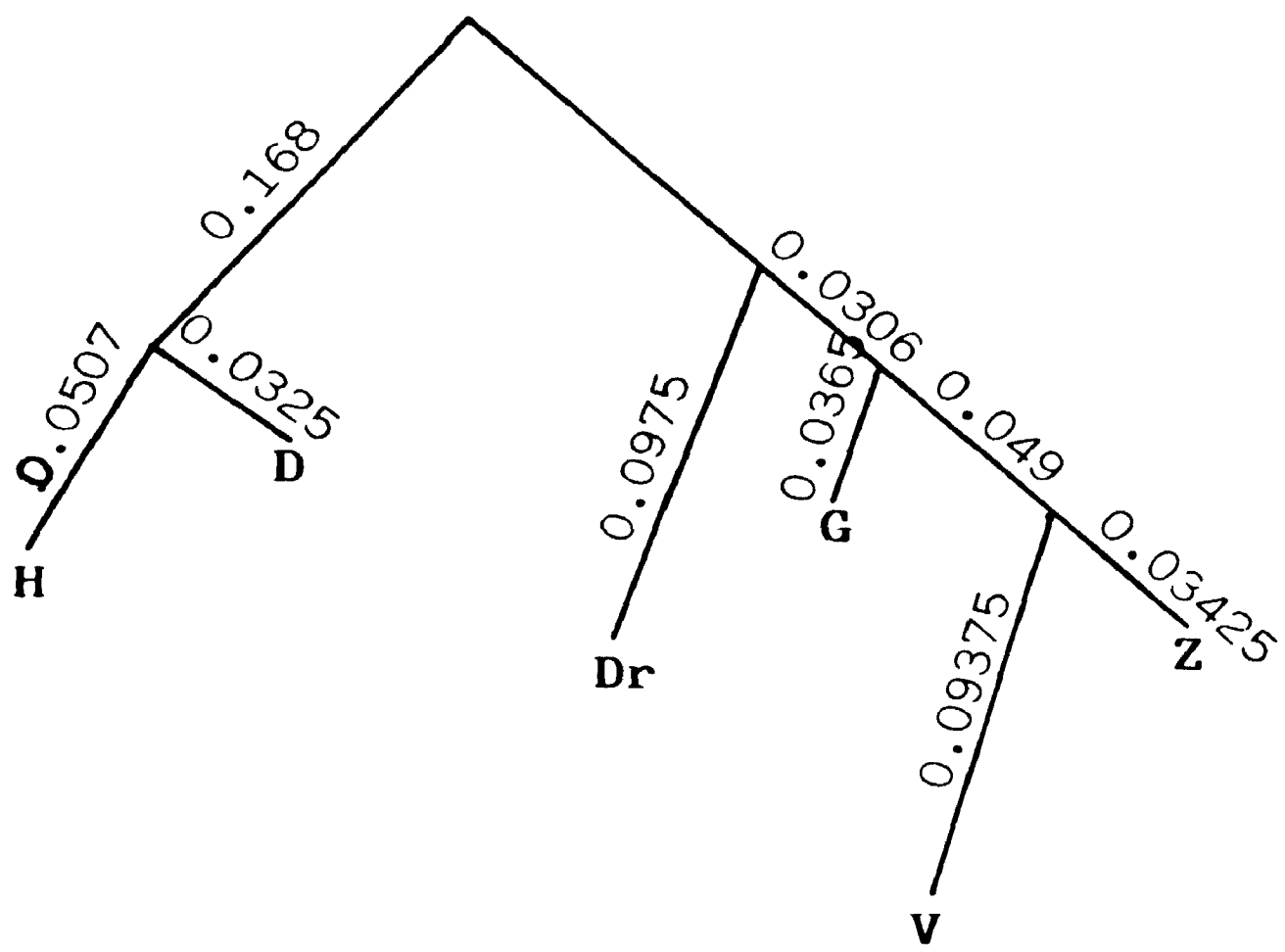


Fig. 8.– Representación mediante un árbol aditivo (distancia aditiva) de las poblaciones cuya matriz de distancias genéticas viene dada en la tabla 1.

Buneman (1971) demuestra que un espacio (Ω,δ) puede representarse a través de un espacio aditivo si y sólo si δ verifica la desigualdad aditiva o axioma de los cuatro puntos. Además, como Waterman *et al.* (1977) demuestran esta representación es única. Otra formalización de la representación aditiva ha sido desarrollada por Arcas (1987).

Un dendograma es un caso particular de un árbol aditivo. En efecto, es un árbol aditivo con un nodo distinguido (llamado raíz) que es equidistante de todos los extremos. Como la desigualdad aditiva es más flexible que la ultramétrica, resulta más fácil ajustar una distancia aditiva a una distancia estadística. En otras palabras, en lugar de un dendograma, resulta más aproximada la representación a través de un árbol aditivo, que usualmente se realiza en forma paralela.

Las diferencias entre ambos tipos de representación son:

a) En el caso ultramétrico, las $n(n-1)/2$ interdistancias entre los individuos vienen determinados por al menos $(n-1)$ valores intermedios, mientras que en el caso aditivo este número se eleva a $(2n-3)$.

b) Una distancia ultramétrica define una jerarquía indexada (C,α), que es la forma más perfecta de clasificación. La distancia entre individuos del mismo "cluster" (distancia intracluster) es siempre menor que la distancia entre individuos de distinto "cluster" (distancia intercluster).

c) Una distancia aditiva no ultramétrica no define ninguna jerarquía indexada. La distancia intracluster puede superar a la distancia intercluster.

d) Toda distancia ultramétrica es euclídea (teorema 2). Una distancia aditiva puede ser no euclídea.

e) En un árbol aditivo no existe un nodo equidistante de los extremos. El problema de fijar una raíz (similar a la elección del origen de coordenadas en una representación espacial) depende del algoritmo de clasificación. Diferentes raices inducen diferentes jerarquías de particiones o "clusters".

f) Si mediante algoritmos adecuados ajustamos una distancia ultramétrica y una distancia aditiva a una misma distancia estadística, la distorsión (que puede medirse utilizando la correlación cofenética) es menor en el caso aditivo.

Un importante resultado es el teorema 3, que permite estudiar y clasificar las distancias aditivas.

TEOREMA 3

Si (Ω,δ) es un espacio aditivo, existe entonces una distancia ultramétrica y una función $\Psi : \Omega \longrightarrow$ R tal que

$$\delta_{ij} = u_{ij} + \Psi(i) + \Psi(j) \qquad (22)$$

Demostración: Ver Buneman (1971).

De la descomposición (22) podemos obtener tres clases de distancias aditivas, dando lugar a 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 nodo interno.

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

Otros tipos de árboles pueden construirse combinando los tres anteriores. Véase Sattah y Tversky (1977), Barthelemy y Guénoche (1988).

Por otra parte, existen diversos algoritmos para ajustar una distancia aditiva, que presentan diversas ventajas e inconvenientes (Arcas y Cuadras, 1987). El más conocido es el ADDTREE, elaborado por A. Tversky e implementado por Corter (1982).

Un algoritmo sencillo puede deducirse partiendo de (22). Mediante alguno de los algoritmos de clasificación jerárquica, ajustemos una distancia estadística d_{ij} a una ultramétrica u_{ij} . Sea

$$y_{ij} = d_{ij} - u_{ij}$$

Ajustemos seguidamente y_{ij} a un modelo lineal de la forma $\alpha_i + \alpha_j$ por el criterio de los mínimos cuadrados. Entonces es fácil ver que

$$\hat{\alpha}_i = [\ (2n-2) \sum_{k=1}^{n} (d_{ik}-u_{ij}) - \sum_{k \neq j} (d_{jk}-u_{jk})\ ] \ / \ 2(n-2)\ (n-1)$$

De este modo, $\hat{\delta}_{ij} = u_{ij} + \hat{\alpha}_i + \hat{\alpha}_j$ es una distancia aditiva que se ajusta a d_{ij} .

5. DISTANCIA DE RAO

En esta sección iniciamos el estudio de las distancias estadísticas definidas sobre distribuciones de probabilidad, en el sentido del apartado *b)* de la sección 1. A causa de sus interesantes propiedades y su conexión con otras distancias, empezaremos comentando una distancia introducida por Rao (1945) y estudiada por Atkinson y Mitchell (1981), Burbea y Rao (1982a,b), Oller y Cuadras (1982a, 1985), Oller (1987), Amari (1985), Cuadras *et al.* (1985), Burbea (1986), Mitchell (1988).

5.1. Definición y propiedades generales

Sea $S = \{ p(X,\theta) \}$ un modelo estadístico, donde X es un vector aleatorio, $\theta = (\theta_1, \ldots, \theta_n)$ es un parámetro n-dimensional, $p(X,\theta)$ es una función de densidad de probabilidad de X parametrizada por θ. Podemos considerar que θ pertenece a una variedad diferenciable Θ y tomar la matriz de información de Fisher

$$G = E \{ [\frac{\partial}{\partial \theta} log\, p(X;\theta)] [\frac{\partial}{\partial \theta} log\, p(X;\theta)]' \} \tag{24}$$

como tensor métrico fundamental sobre Θ. Indicando $G = [g_{ij}(\theta)]$, el elemento del arco (al cuadrado) es

$$d s^2 (\theta) = \sum_{i,j=1}^{n} g_{ij}(\theta)\, d\theta_i\, d\theta_j \tag{25}$$

Debido a que G se comporta como un tensor covariante simétrico de segundo orden para todo θ, resulta que (25) es invariante por transformaciones admisibles de los parámetros. Fijando entonces dos puntos θ_A, θ_B de Θ, y una curva paramétrica $\theta = \theta(t)$, $t_A \leq t \leq t_B$, con $\theta(t_A) = \theta_A$, $\theta(t_B) = \theta_B$, la distancia entre ambos a lo largo de la curva es

$$\int_{t_A}^{t_B} \frac{d s(\theta)}{d t}\, dt = \int_{t_A}^{t_B} [\sum_{i,j=1}^{n} g_{ij}(\theta)\, \dot{\theta}_i\, \dot{\theta}_j]^{1/2}\, dt \tag{26}$$

donde $\dot{\theta}_i$ significa la derivada respecto t. La distancia geodésica entre θ_A y θ_B es la distancia a lo largo de una curva geodésica, es decir, a lo largo de una curva tal que (26) sea mínima. La curva geodésica se obtiene resolviendo la ecuación de Euger-Lagrange

$$\sum_{i=1}^{n} g_{ih}\, \dot{\theta}_i + \sum_{i,j=1}^{n} \Gamma_{ijh}\, \dot{\theta}_i\, \dot{\theta}_j = 0 \qquad h = 1, \ldots, n$$

con las condiciones de contorno $\theta(t_A) = \theta_A$, $\theta(t_B) = \theta_B$, siendo

$$\Gamma_{ijh} = \frac{1}{2} \left[\partial_i g_{jh} + \partial_j g_{hi} - \partial_h g_{ij} \right]$$

los símbolos de Christoffel de primera clase.

La distancia geodésica $R(\theta_A, \theta_B)$ sobre Θ, basada en la matriz de información de Fisher, recibe el nombre de distancia geodésica informacional o *distancia de Rao*. Utilizada como distancia entre dos densidades de probabilidad de S goza de las siguientes propiedades:

1) Es invariante por transformaciones admisibles, tanto de las variables como de los parámetros.

2) Si las variables aleatorias contenidas en X son estocasticamente independientes y con distribuciones uniparamétricas, entonces la distancia geodésica es euclídea.

3) Se puede relacionar con el contraste de hipótesis y está conectada con las propiedades asintóticas de ciertos estimadores de parámetros.

La distancia de Rao puede también introducirse relacionándola con ciertas medidas de divergencia (ver sección siguiente), o por via axiomática, exigiendo ciertas condiciones generales a una distancia entre distribuciones de probabilidad (Cuadras et al., 1985). Otro camino para introducir esta distancia consiste en considerar el espacio tangente T_θ definido en cada punto de la variedad. T_θ es un espacio vectorial local, que está generado por los n vectores $\partial / \partial \theta_i$, $i = 1, \ldots, n$. Todo vector tangente puede ser representado como una combinación lineal de la base natural $\partial_i = \partial / \partial \theta_i$

$$V = \sum_{i=1}^{n} V^i \partial_i \qquad (27)$$

Consideremos ahora el modelo estadístico $S = \{ p(X,\theta) \}$ y las n variables aleatorias

$$Z_i = \frac{\partial}{\partial \theta_i} log\ p\ (X,\theta) \qquad i = 1, \ldots, n \qquad (28)$$

y supongamos que son linealmente independientes en X para cada valor de θ. Definimos entonces el espacio vectorial

$$T_\theta^{(1)} = < Z_1, \ldots, Z_n >$$

de las variables aleatorias que son combinación lineal de Z_i .

Existe un isomorfismo natural entre T_θ y $T_\theta^{(1)}$. En efecto, basta establecer la correspondencia

$$\partial_i \iff Z_i = \frac{\partial}{\partial\, \theta_i} \log p\,(X,\theta) \tag{29}$$

según la cual la imagen de (27) será la variable aleatoria

$$V_x = \sum_{i=1}^{n} V^i Z_i$$

Luego podemos identificar T_θ con $T_\theta^{(1)}$, y referirnos a $T_\theta^{(1)}$ como la representación en términos de variables aleatorias de T_θ.

Observando que, bajo ciertas condiciones de regularidad, se verifica

$$E\,(Z_i) = 0 \qquad i = 1, \ldots, n$$

el producto escalar natural en $T_\theta^{(1)}$ es

$$< U_x, V_x > = cov\,(U_x, V_x) = E\,(U_x . V_x)$$

luego los productos escalares de la base $\{Z_i\}$ son

$$g_{ij}\,(\theta) = E\,(Z_i . Z_j) \qquad i,j = 1, \ldots, n$$

que constituyen la matriz de información de Fisher (24). Puesto que hemos definido un producto escalar en cada espacio tangente $T_\theta^{(1)}$, resulta entonces que hemos dotado a S de una estructura de espacio de Riemann. La distancia geodésica entre dos puntos de la variedad se obtiene minimizando (26).

Con el propósito de estructurar las propiedades intrínsecas de un modelo estadístico, Amari (1985) considera la α-conexión

$$\Gamma_{ijk}^{(\alpha)} = \Gamma_{ijk} + \frac{1 - \alpha}{2} T_{ijk}$$

siendo T_{ijk} el tensor simétrico

$$T_{ijk} = E(Z_i . Z_j . Z_k)$$

Las geodésicas asociadas a un α-conexión son aquellas curvas cuyo vector tangente se desplaza paralelamente a lo largo de si misma y pueden considerarse como rectas. Tomando como tensor métrico la matriz de

información de Fisher, resulta que la 0-conexión coincide con la conexión de Levi-Civita, que es la única que hace compatible la noción afín de paralelismo con la noción métrica de distancia. Las demás conexiones son menos naturales desde el punto de vista métrico, pero tienen interesantes interpretaciones estadísticas.

La 1-conexión fue introducida por Efron (1975), y tiene una interpretación natural si consideramos la familia exponencial. Para esta familia se verifica

$$\Gamma_{ijk}^{(\alpha)} = \frac{1 - \alpha}{2} T_{ijk}$$

que es identicamente 0 para $\alpha = 1$. La familia exponencial constituye un espacio sin curvatura respecto a la 1-conexión y las geodésicas asociadas pueden interpretarse como rectas.

La -1-conexión fue introducida por Dawid (1975). Si consideramos una mixtura de distribuciones

$$p(x, \theta) = (1-\theta)\, p_1(x) + \theta\, p_2(x) \qquad 0 \leq \theta \leq 1$$

entonces

$$\Gamma_{ijk}^{(\alpha)} = \frac{1 + \alpha}{2} T_{ijk}$$

que es idénticamente 0 para $\alpha = -1$. Obtenemos una familia de distribuciones que constituyen un espacio sin curvatura respecto a la -1-conexión. La familia puede considerarse como una línea recta conectando dos distribuciones.

La teoría de las α-conexiones puede aplicarse para estudiar la familia de distribuciones exponenciales, así como la familia exponencial curvada de Efron. Véase Amari (1985), Burbea (1986).

La distancia de Rao, es decir, la distancia geodésica en S basada sobre la métrica asociada a (24), ha sido calculada para la mayoría de distribuciones univariantes y algunas distribuciones multivariantes. Si en ciertos casos el cálculo es sencillo, en otros es bastante complejo o no ha sido resuelto todavía. Por ejemplo, para el sistema uni-paramétrico bivariante propuesto por Cuadras y Augé (1981), en la expresión de la distancia intervienen diversos desarrollos en serie (Ruiz-Rivas y Cuadras, 1988). La distancia entre dos normales multivariantes con distinta matriz de covarianzas, todavía no ha sido resuelta, aunque se han intentado algunas aproximaciones al problema (Oller y Cuadras, 1983, Calvo, 1988).

5.2. Distancias entre distribuciones univariantes

Supongamos que $p(x \mid \theta)$ es una función de densidad univariante y uniparamétrica. Es fácil ver que la distancia de Rao entre $a, b \in \Theta$ viene dada por

$$R(a, b) = \left| \int_a^b \sqrt{g(\theta)}\, d\theta \right| \tag{30}$$

A continuación damos las distancias para algunas de estas distribuciones.

a. Binomial:

$$p(x \mid \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x} \qquad x = 0,1,\ldots,N \qquad 0 < \theta < 1$$

$$R(a,b) = 2\sqrt{N}\ \arccos\left\{\sqrt{ab} + \sqrt{(1-a)(1-b)}\right\}$$

b. Poisson:

$$p(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!} \qquad x = 0,1,2,\ldots\ \theta > 0$$

$$R(a,b) = 2\left|\sqrt{a} - \sqrt{b}\right|$$

c. Binomial negativa (r fijo):

$$p(x \mid \theta) = \frac{\Gamma(x+r)}{x!\,\Gamma(r)} \theta^x (1-\theta)^r \qquad x = 0,1,2,\ldots \qquad 0 < \theta < 1$$

$$R(a,b) = 2\sqrt{r}\ \cos h^{-1}\left(\frac{1 - \sqrt{ab}}{\sqrt{(1-a)(1-b)}}\right)$$

d. Gamma (r fijo)

$$p(x \mid \theta) = \frac{1}{\Gamma(r)} x^{r-1} e^{-x\theta} \theta^r \qquad x > 0,\ \theta > 0$$

$$R(a,b) = \sqrt{r}\,\left|\log(a/b)\right|$$

e. Weibull (α fijo)

$$p(x \mid \theta) = \alpha x^{\alpha-1} \theta\, e^{-x^{\alpha}\theta} \qquad x > 0,\ \theta > 0,\ \alpha > 0$$

$$R(a,b) = \left|\log(a/b)\right|$$

Nótese que, como la distribución de Weibull es la que sigue $X^{1/\alpha}$ donde X es Gamma ($r=1$), se obtiene un caso particular de la distancia anterior, pues la distancia de Rao es invariante.

f. Pareto (r fijo)

$$p(x \mid \theta) = \theta\, r^{\theta}\, x^{-(\theta+1)} \qquad x \geq r,\ \theta \geq o$$

$$R(a,b) = \mid log\ (a/b) \mid$$

g. Normal N (μ, θ^2), (μ fijo)

$$R\ (a,b) = \frac{1}{\sqrt{2}} \mid log\ (a/b) \mid$$

h. Normal N (μ, σ^2) , (σ^2 fijo)

$$R\ (a,b) = \frac{\mid a - b \mid}{\sigma}$$

5.3. Distancia entre distribuciones univariantes biparamétricas

Supongamos que ahora Θ tiene dimensión $n=2$. Entonces es necesario calcular la matriz de información de Fisher y proceder como hemos explicado al principio de esta sección, es decir, resolviendo las correspondientes ecuaciones geodésicas. Las distancias que exponemos seguidamente han sido obtenidas por Atkinson y Mitchell (1981), Burbea y Rao (1982 *a,b*), Oller (1987).

a. Distribución normal

La distancia de Rao entre $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, es decir, entre los puntos (μ_1,σ_1^2), (μ_2,σ_2^2), es

$$R(1,2) = \frac{1}{\sqrt{2}}\ log\ \frac{1 + \delta\ (1\ ,\ 2)}{1 - \delta\ (1\ ,\ 2)}$$

siendo

$$\delta(1,2) = \left[\ \frac{(\mu_1 - \mu_2)^2 + 2\ (\sigma_1 - \sigma_2)^2}{(\mu_1 - \mu_2)^2 + 2\ (\sigma_1 + \sigma_2)^2}\ \right]^{1/2}$$

b. Distribución de valores extremos de Gumbel

$$p(x \mid \alpha,\theta) = \frac{1}{\theta}\ exp\ (-exp\ (x-\alpha)\ /\ \alpha)\ )\ exp\ (-\ (x-\alpha)\ /\ \theta)$$

$$\theta > 0,\ x > 0,\ \alpha \in R$$

Sea γ la constante de Euler. Indicando $a=1-\gamma$, $b=\pi\ /\ \sqrt{6}$, la distancia entre los puntos (α_1,θ_1) y (α_2,θ_2) es

$$R\ (1,2) = b\ log\ \frac{1 + \delta\ (1,2)}{1\ -\ \delta\ (1,2)}$$

donde

$$\delta(1,2) = \left\{ \frac{[\ (\alpha_2-\alpha_1)\ -\ a(\theta_2-\theta_1)\ ]^2 + b^2\ (\theta_2-\theta_1)^2}{[\ (\alpha_2-\alpha_1)\ -\ a(\theta_2-\theta_1)\ ]^2 + b^2\ (\theta_2+\theta_1)^2} \right\}^{1/2}$$

c. Distribución de valores extremos de Cauchy-Frechet

$$p(x|\beta,\lambda) = exp\ (-\ (x/\beta)^{-\lambda}\ )\ (x/\beta)^{-(\lambda+1)}\ \lambda/\beta \qquad \beta,\lambda\ ,\ x > 0$$

La distancia entre los puntos (β_1,λ_1) y (β_2,λ_2) es

$$R(1,2) = b\ .\ log\ \frac{1 + \delta(1,2)}{1\ -\ \delta(1,2)}$$

donde

$$\delta(1,2) = \left\{ \frac{[\ log\ (\beta_2/\beta_1)+a(\lambda_2-\lambda_1)\ /\ \lambda_1\lambda_2]^2 + b^2(\lambda_2-\lambda_1)^2\ \lambda_1^2\ \lambda_2^2}{[\ log\ (\beta_2/\beta_1)+a(\lambda_2-\lambda_1)\ /\ \lambda_1\lambda_2]^2 + b^2(\lambda_2+\lambda_1)^2\ \lambda_1^2\ \lambda_2^2} \right\}^{1/2}$$

y $a = 1 - \gamma$, $b = \pi/\sqrt{6}$.

d. Distribución logística

$$p\ (x \mid \alpha,\beta) = \frac{1}{4\ \beta}\ sech^2\ \left(\frac{x-\alpha}{2\ \beta}\right) \qquad x,\alpha \in R\ ,\ \beta > 0$$

La distancia entre los puntos (α_1, β_1) y (α_2, β_2) es

$$R\,(1,2) = \frac{\sqrt{b}}{3}\ log\ \frac{1 + \delta(1,2)}{1 - \delta(1,2)}$$

donde

$$\delta(1,2) = \left[\ \frac{(3/b)\ (\alpha_2 - \alpha_1)^2 + (\beta_2 - \beta_1)^2}{(3/b)\ (\alpha_2 - \alpha_1)^2 + (\beta_2 + \beta_1)^2}\ \right]^{1/2}$$

$b = \pi^2 + 3$.

5.4. Distancias entre distribuciones multivariantes

La distancia de Rao para diversas distribuciones multivariantes ha sido estudiada por Bhattacharyya (1946), Atkinson y Mitchell (1981), Burbea y Rao (1982*a,b*), Oller y Cuadras (1982*a*, 1983, 1985), Ruiz-Rivas y Cuadras (1988). Actualmente se intenta encontrar la distancia de Rao para distribuciones multivariantes en las que ni μ ni Σ están fijos.

a. Multinomial (N fijo)

$$p(x \mid \theta) = \frac{N\ !}{x_1\ !\ \ldots\ x_n\ !}\ (\theta_1)^{x_1}\ \ldots\ (\theta_n)^{x_n}$$

$$x_i \geq 0 \quad 0 < \theta_i < 1$$

$$\sum_{i=1}^{n}\ x_i = N\ ;\ \sum_{i=1}^{n}\ \theta_i = 1$$

La distancia de Rao entre los puntos $(a_1, \ldots, a_n)$, $(b_1, \ldots, b_n)$ es

$$R(1,2) = 2\ \sqrt{N}\ \ arcos\ (\ \sum_{i=1}^{n}\ \sqrt{a_i\ b_i}\ )$$

es decir, hemos obtenido la distancia de Hellinger-Battacharyya inicialmente propuesta por Bhattacharyya (1946) (ver sección 7.2).

b. Multinomial negativa (*r* fijo)

$$p(x \mid \theta) = \frac{\Gamma\ (x_1 + \ldots + x_n + r)}{x_1\ !\ \ldots\ x_n\ !\ \Gamma(r)}\ (\theta_1)^{x_1} \ldots\ (\theta_n)^{x_n}\ (\theta_{k+1})^{r}$$

siendo $\sum_{i=1}^{n} \theta_i < 1$, $\theta_{n+1} = 1 - \sum_{i=1}^{n} \theta_i$

La distancia entre $(a_1, \ldots, a_n, a_{n+1})$ y $(b_1, \ldots, b_n, b_{n+1})$ es

$$R\ (1,2) = 2 \sqrt{r}\ cos\ h^{-1} \left(\frac{1 - \sum_{i=1}^{n} \sqrt{a_i\, b_i}}{\sqrt{a_{k+1}\, b_{k+1}}} \right)$$

c. Normal multivariante (Σ fijo)

La distancia de Rao entre $N_n(\mu_1, \Sigma)$ y $N_n(\mu_2, \Sigma)$, donde $\mu_1, \mu_2 \in R^n$ y Σ permanece fijo es

$$R(1,2) = \{ (\mu_1 - \mu_2)'\ \Sigma^{-1}\ (\mu_1 - \mu_2) \}^{1/2}$$

es decir, es la conocida distancia de Mahalanobis (1936).

d. Normal multivariante (μ fijo)

La distancia entre $N_n(\mu_o, \Sigma_1)$ y $N_n(\mu_o, \Sigma_2)$, donde μ_o permanece fijo, es

$$R\ (1,2) = \left(\frac{1}{2} \sum_{i=1}^{n} log^2\ \lambda_i \right)^{1/2}$$

donde $\lambda_1, \ldots, \lambda_n$ son los valores propios de Σ_2 respecto de Σ_1, es decir, las soluciones de la ecuación en determinantes

$$| \Sigma_2 - \lambda\ \Sigma_1 | = 0$$

e. Normal multivariante

La distancia entre dos normales $N_n(\mu_1, \Sigma_1)$, $N_n(\mu_2, \Sigma_2)$ es un problema todavía no completamente resuelto. Aunque se han podido integrar las geodésicas, el problema algebraico de determinar las constantes de integración para enlazar dos puntos de la variedad todavía no se ha podido resolver. Sin embargo, se conoce la solución para algunos casos particulares. Además se ha podido demostrar que existe una isometría entre la variedad y el grupo $P_{n+1}(R)$ de todas las matrices definidas positivas de orden $n+1$, con la métrica

$$< A\ ,\ B\ > = tr\ (A\ B')$$

Considerando entonces que se puede obtener una distancia para una subvariedad de $P_{n+1}(R)$, se ha conseguido la siguiente cota inferior para la distancia de Rao

$$d(1\ ,\ 2) = \left(\frac{1}{2}\sum_{i=1}^{n+1} log^2\ (\lambda_i)\ \right)^{1/2}$$

donde $\lambda_1, \ldots, \lambda_n, \lambda_{n+1}$ son los valores propios de S_2 respecto S_1, siendo

$$S_i = \begin{pmatrix} \Sigma_i + \mu_i\,\mu_i' & \mu_i \\ \mu_i' & 1 \end{pmatrix} \qquad i = 1,2$$

Véase Burbea (1986), Calvo (1988), Skovgaard (1984).

5.5. Una distancia intrapoblacional

Supongamos ahora que el vector aleatorio X toma valores en una población Ω. La correspondencia (29) permite definir una distancia entre los individuos de Ω caracterizados por un punto θ de Θ. En efecto, podemos caracterizar Ω a través del espacio dual E_θ^*, siendo $E_\theta = T_\theta$ el espacio tangente introducido anteriormente, haciendo corresponder a $\omega \in \Omega$ la forma lineal $\omega^* \in E_\theta^*$ tal que $\omega^*(Z) = Z(\omega)$, para toda variable aleatoria $Z \in E_\theta$. Como la métrica en E_θ^* inducida por la métrica en E_θ, tiene como matriz asociada G^{-1} respecto a la base dual de $Z_1, \ldots, Z_n$, podemos definir una distancia entre individuos

$$R(\omega_1, \omega_2) = d_{E_\theta^*}(\omega_1^*\ ,\ \omega_2^*) \tag{31}$$

donde $d_{E_\theta^*}$ es una distancia euclídea local.

Por ejemplo, consideremos una población $N_p(\mu;\Sigma)$, con Σ fijo. Entonces E_μ está generada por el vector aleatorio $Z = \Sigma^{-1}(X-\mu)$ y la métrica en E_μ viene dada por Σ^{-1}, luego la métrica en E_μ^* viene dada por Σ. La distancia (al cuadrado) entre dos individuos con valores $\mathbf{z}_1 = \Sigma^{-1}\ (x_1 - \mu)$, $\mathbf{z}_2 = \Sigma^{-1}\ (x_2 - \mu)$, es

$$(x_1-x_2)'\ \Sigma^{-1}\ \Sigma\ \Sigma^{-1}\ (x_1-x_2) = (x_1-x_2)'\ \Sigma^{-1}\ (x_1-x_2)$$

es decir, coincide con la distancia de Mahalanobis (12).

Consideremos ahora $n+1$ sucesos mutuamente excluyentes $A_1, \ldots, A_n, A_{n+1}$ de probabilidades $p_1, \ldots, p_n, p_{n+1}$ y la función de densidad

$$f(x_1, \ldots, x_n / p_1, \ldots, p_n) = p_1^{x_1} \ldots p_n^{x_n} \left(1 - \sum_{i=1}^{n} p_i\right)^{1 - \sum_{i=1}^{n} x_i}$$

$x_i \in \{0,1\}$, $0 < p_i < 1$, $\quad i=1, \ldots, n$,

es decir, $x_i = 1$ si $\omega \in A_i$, $x_i = 0$ en caso contrario. Indicando

$$x_{n+1} = 1 - \sum_{i=1}^{n} x_i \qquad p_{n+1} = 1 - \sum_{i=1}^{n} p_i$$

el tensor métrico G sobre $\Theta = \{ (p_1, \ldots, p_n \mid \sum_{i=1}^{n} p_i < 1 \}$ es

$$g_{ij} = \frac{\delta_{ij}}{p_i} + \frac{1}{p_{n+1}} \qquad i,j = 1, \ldots, n$$

Con algo de esfuerzo se obtiene entonces que la distancia entre un individuo ω_1 que presenta la característica A_i, es decir, $\omega_1 \in A_i$, y otro individuo $\omega_2 \in A_j$, es

$$R(\omega_1,\omega_2) = \sqrt{(1 - \delta_{ij}) \left(\frac{1}{p_i} + \frac{1}{p_j} \right)}$$

siendo δ_{ij} la delta de Kronecker. Finalmente, si consideramos k particiones independientes de Ω

$$A_1^{(i)}, \ldots, A_{n_i}^{(i)}, A_{n_i+1}^{(i)} \qquad i = 1, \ldots, k$$

la distancia entre ω_1 y ω_2 tales que

$$\omega_1 \in A_{i_1}^{(1)} \cup \ldots \cup A_{i_k}^{(k)} \qquad \omega_2 \in A_{j_1}^{(1)} \cup \ldots \cup A_{j_k}^{(k)}$$

viene dada por

$$R(\omega_1, \omega_2) = \sum_{\alpha=1}^{k} (1 - \delta_{i_\alpha j_\alpha}) \left(\frac{1}{p_{i_\alpha}} + \frac{1}{p_{j_\alpha}} \right) \tag{32}$$

siendo $p_{i_\alpha} = P\,(A_i^{(\alpha)})$. La distancia (32) puede ser utilizada para diferenciar individuos de una población conocida la presencia o ausencia de características cualitativas.

6. DIVERGENCIAS

Las medidas no paramétricas de divergencia entre distribuciones de probabilidad se definen como expresiones funcionales (a menudo relacionadas con la teoría de la información), que miden el grado de discrepancia entre

dos distribuciones cualesquiera, no necesariamente pertenecientes a una misma familia paramétrica. Después de los trabajos pioneros de Pearson (prueba *ji*-cuadrado) y Hellinger (la famosa distancia de Hellinger, publicada en 1909), otros autores han estudiado divergencias (Shannon, Kullbach y Leibler, Renyi, etc.). La divergencia aplicada a distribuciones de probabilidad serían introducidas por Csiszar (1963, 1967, 1972, 1975), estudiadas en diferentes versiones por Matusita (1955, 1964), Havrda y Charvat (1967), Vajda (1972) y generalizadas por Burbea y Rao (1982 *a,b*).

Las divergencias tienen aplicaciones en inferencia estadística y en procesos estocásticos. Véase Bishop *et al.* (1975), Liese y Vajda (1987).

6.1. Distribución multinomial

Sea $p = (p_1, \ldots, p_n)$ el vector de probabilidades correspondiente a una distribución multinomial. Un funcional ϕ - entropia es

$$H_\phi (p) = \Sigma\, p_i\, \phi (p_i) \tag{33}$$

donde ϕ es una función estrictamente convexa tal que $\phi(1) = 0$. H_ϕ es una función sobre la clase de distribuciones multinomiales n-dimensionales que es máxima cuando los p_i son iguales y alcanzan el valor mínimo (cero) cuando algún $p_i = 1$. H_ϕ mide el grado de discrepancia con la distribución de máxima entropía, y ha sido ampliamente utilizada como medida de diversidad.

Sean $p = (p_1, \ldots, p_n)$, $q = (q_1, \ldots, q_n)$ dos distribuciones multinomiales. La divergencia entre p y q se puede medir como la discrepancia entre el cociente $x_i = q_i / p_i$ y 1. Basándonos en el significado de (33), definimos una divergencia entre p y q, llamada ϕ-divergencia de Csiszar (1972), como el valor esperado de $x_1, \ldots, x_n$

$$C_\phi (p, q) = \Sigma\, p_i\, \phi (q_i / p_i) \tag{34}$$

Por la desigualdad de Jensen se tiene

$$C_\phi (p, q) = \Sigma\, p_i\, \phi (x_i) \geq \phi (\Sigma\, p_i\, x_i) = \phi (1) = 0$$

alcanzándose el valor 0 si y sólo si $p=q$. (34) se puede tomar como una medida de disimilaridad entre p y q, pero en general no es una distancia, pues no siempre es simétrica, o si lo es, puede no cumplir la desigualdad triangular. Sin embargo, tiene dos interesantes propiedades: $C_\phi(p, q)$ aumenta cuando se considera una partición más fina, y bajo la hipótesis $C_\phi (p, q) = 0$, el estadístico

$$\mathbf{V} = \frac{2\ N_1\ N_2}{(N_1+N_2)\ \phi\ ''\ (1)}\ C_o\ (\hat{p}\ ,\ \hat{q}) \qquad (35)$$

sigue (asintóticamente) la distribución *ji*-cuadrado con $n-1$ grados de libertad, siendo $\hat{p}_i$, $\hat{q}_i$ las frecuencias relativas muestrales para muestras de tamaños N_1, N_2 (Takeuchi *et al*, 1982).

El cuadro 3 contiene diversas formas de (34) según diferentes expresiones de $\phi(x)$, incluyendo, en su caso, la distancia genética que da lugar (ver sección 7.2). Un caso importante es $\phi(x) = -log\ x$. Entonces H_o es la famosa entropia de Shannon y (34) es

$$I_o\ (p,q) = \Sigma\ p_i\ log\ (p_i\ /\ q_i) \qquad (36)$$

conocida como medida de información de Kullback-Leibler. (36) mide la ganancia de información al pasar de la distribución p a la q, y ha sido utilizada en estadística, especialmente en estadística bayesiana (Bernardo, 1981, 1987). La simetrización de (36)

$$J_o\ (p,q) = I_o\ (p,q) + I_o(q,p)$$

es el invariante de Jeffreys, también llamada L-divergencia.

Obsérvese que para $\phi(x) = |\ 1-x\ |$ se obtiene $\Sigma\ |p_i - q_i\ |$ distancia que ha sido utilizada en genética (Prevosti, *et al*, 1975). Sin embargo, en este caso no se puede utilizar (35) porque $\phi\ ''(1)$ no existe.

Por otra parte, si consideramos las entropias $H_o(p)$, $H_o(q)$ y la entropia correspondiente a la mixtura $\lambda p + (1-\lambda)q$, entonces mediante la diferencia de Jensen

$$J_o(p,q) = H_o\ (\lambda p + (1\ -\lambda)\ \mathbf{q}) - \lambda\ H_o(p) - (1\ -\lambda)\ H_o(\mathbf{q})$$

obtenemos una distancia, llamada J-divergencia, entre p y q. Por ejemplo, utilizando la entropia de Gini-Simpson

$$H_o\ (p) = 1\ -\ \Sigma\ p_i^2$$

se obtiene la distancia

$$2\ \lambda(1\ -\lambda)\ \Sigma\ (p_i\ -\ q_i)^2$$

que ha sido utilizada en genética por Nei (1971). Véase Rao (1982). Tanto la J-divergencia como la L-divergencia son estudiadas, con más generali-

dad, en la siguiente sección. Por otra parte Pérez et al. (1986) prueban que la entropia de Gini-Sampson puede ser estimada (en poblaciones finitas) más fácilmente que la de Shannon, por lo que recomiendan la primera para estimar la diversidad.

CUADRO 3

Algunas ϕ-divergencias: $C_\phi(p,q) = \Sigma\, p_i\, \phi\,(q_i / p_i)$

$\phi(x)$	$C(p,q)$	NOMBRE	GENETICA
$-\log x$	$\Sigma\, p_i \log(p_i/q_i)$	Kullback-Leibler	
$\frac{(1 - x^{\alpha-1})}{(1-\alpha)}$	$1 - \Sigma\, q_i^{\alpha-1}\, p_i^{\alpha}$	Havrda-Charvat	
$2(1 - \sqrt{x})$	$\Sigma\,(\sqrt{p_i} - \sqrt{q_i})^2$	Bhattacharyya	Cavalli-Sforza
$\lvert 1 - x \rvert$	$\Sigma\, \lvert p_i - q_i \rvert$		Prevosti
$\frac{(x-1)^2}{(x+1)}$	$\Sigma\, \frac{(p_i - q_i)^2}{(p_i+q_i)}$		Balakrishnan-Sanghvi

Salicrú y Cuadras (1988) prueban que todo funcional ϕ-entropia (33) puede interpretarse como una medida de Csiszar (34) entre p y la distribución de máxima entropia $e = (1/n, \ldots, 1/n)$. Por ejemplo, para la entropia $H_\alpha(p)$ de Havrda-Charvat, en la que

$$\phi(x) = (\alpha - 1)^{-1}\,(1 - x^{\alpha-1}) \qquad \alpha > 1$$

se verifica

$$H_\alpha(p) = C_f(p,e) = \sum_{i=1}^{n} f\left((n\,p_i)^{-1}\right) p_i$$

para la función

$$f(x) = (\alpha - 1)^{-1}\,[1 - (n\,x)^{1-\alpha}]$$

Por otra parte, la minimización de $C_\phi(p,\hat{f})$, donde $\hat{f}$ representa el vector de frecuencias relativa y ϕ se elige adecuadamente, es equivalente a ciertos procedimientos clásicos en el tratamiento estadístico de datos multinomiales. Por ejemplo, tomando $\phi(x) = -\ln x$ hallar la estimación máximo verosí-

mil de p es equivalente a hallar $\hat{p}$ que minimiza $C_\phi(p,\hat{f})$. Tomando $\phi(x) = (x-1)^2$, entonces minimizar $C_\phi(p,\hat{f})$ es equivalente al método de la mínima *ji*-cuadrado

$$\min_{\mathbf{p}} \sum_{i=1}^{n} \frac{(\hat{f}_i - p_i)^2}{p_i} = C_\phi(\hat{p},\hat{f})$$

Véase Bishop *et al.* (1975).

6.2. Distribuciones absolutamente continuas

Sea $p(x)$ una función de densidad de un vector multivariante con distribución absolutamente continua y soporte en χ. Sea Φ una función real, dos veces derivable, sobre un intervalo T_Φ tal que $[0,1] \subset T_\Phi \subset [0,\infty]$. Se define el funcional Φ entropia

$$H_\Phi(p) = -\int_\chi \Phi[p(x)]\, dx \tag{38}$$

La J-divergencia entre dos distribuciones p,q, con respecto H_Φ, se define como la diferencia de Jensen

$$J_\Phi(p,q) = H_\Phi\left(\frac{p+q}{2}\right) - [H_\Phi(p) + H_\Phi(q)]/2 \tag{39}$$

La K-divergencia y la L-divergencia se definen como

$$K_\Phi(p,q) = \int_\chi (p-q)[\Phi(p)/p - \Phi(q)/q)]\, dx \tag{40}$$

y (suponiendo $T_\Phi = R^+$)

$$L_\Phi(p,q) = \int_\chi [p\,\Phi(q/p) + q\,\Phi(p/q)]\, dx \tag{41}$$

Finalmente, se define la M-divergencia como

$$M_\Phi^2(p,q) = \int_\chi (\sqrt{\Phi}(p) - \sqrt{\Phi}(q))^2\, dx \tag{42}$$

Todas estas definiciones pueden ser generalizadas facilmente escribiendo $d\mu$ en vez de dx, donde μ es una medida aditiva σ-finita y χ es un espacio medible Lebesgue.

Las J,K,L,M-divergencias son siempre simétricas. La M-divergencia es no negativa. Las condiciones para que las demás sean no-negativas son:

a) $J_\Phi(p,q) \geq 0$ si y sólo si $\Phi(u)$ es convexa en T_Φ.

b) $K_\Phi(p,q) \geq 0$ si y sólo si $\Phi(u)/u$ es creciente en T_Φ.

c) $L_\Phi(p,q) \geq 0$ si y sólo si $u\Phi(u^{-1}) + \Phi(u)$ es no negativa en R^+.

Para otras propiedades generales, véase Burbea y Rao (1982 *a,b*).

Estudiemos ahora casos particulares de las L-divergencias, en especial aquellas que están relacionadas con la función

$$\begin{aligned} \Phi_\alpha(u) &= (\alpha - 1)^{-1}(u^\alpha - u) && \alpha \neq 1, \\ &= u \log u && \alpha = 1 \end{aligned} \tag{43}$$

En este caso indicaremos J_α, K_α, L_α.

1) Tomando la función

$$f^*(u) = \Phi(u) + u\,\Phi(1/u)$$

vemos que una L-divergencia coincide con la f^* - divergencia de Csiszár

$$L_\Phi(p,q) = C_{f^*}(p,q) = \int_\chi p\, f^*(p/q)\, dx \tag{44}$$

2) Tomando $\alpha=1$ en (43) obtenemos

$$K_1(p,q) = L_1(p,q) = \int_\chi (p-q)(\log p - \log q)\, dx \tag{45}$$

que es la divergencia de Jeffreys-Kullback-Leibler, que juega un papel destacado en inferencia estadística.

3) Tomando $\alpha=2$ en (43) obtenemos

$$J_2(p,q) = 2\int_\chi (p-q)^2\, dx$$

4) Para $\Phi(u) = u$ en (42) y tomando la raíz cuadrada obtenemos

$$M(p,q) = \left[\int_\chi (\sqrt{p} - \sqrt{q})^2\, dx\right]^{\frac{1}{2}} \tag{46}$$

que es la distancia de Matusita (1955), ampliamente utilizada en inferencia estadística y teoría de la decisión (Matusita, 1964). $M(p,q)$ esta relacionada con la afinidad entre p y q.

$$\rho(p,q) = \int_\chi \sqrt{p}\,\sqrt{q}\, dx$$

Se verifica

$$M^2(p,q) = 2\ [\ 1 - \rho(p,q)\ ]$$

Para el caso particular de dos distribuciones normales $N(\mu_i,\Sigma_i)$, i=1,2, la afinidad $\rho(p,q)$ es

$$\frac{|\ \Sigma_1\ \Sigma_2\ |^{1/4}}{|\ (\Sigma_1 + \Sigma_2)\ /\ 2\ |^{1/2}}\ exp\ [\ -\ (\ \sum_{i,j=1}^{2}\ (\mu_i-\mu_j)'\ \Sigma_j^{-1}\ (\Sigma_i^{-1} + \Sigma_j^{-1})\ \Sigma_i^{-1}\ \mu_i\ )\ ]$$

Si $\Sigma_1 = \Sigma_2 = \Sigma$ entonces

$$\rho(p,q) = exp\ [\ -\frac{1}{8}\ (\mu_1-\mu_2)'\ \Sigma^{-1}\ (\mu_1-\mu_2)\ ]$$

5) $M(p,q)$ está relacionada con la distancia de Hellinger

$$H_\alpha(f,g) = |\ \int_\chi\ (f^{1/\alpha} - g^{1/\alpha})^\alpha\ dx\ |\quad \alpha \geq 1 \tag{47}$$

que verifica $0 \leq H_\alpha \leq 1$. Por otra parte, si consideramos el espacio de las funciones de cuadrado integrable sobre un soporte χ, con el producto escalar

$$< f,g > = \int_\chi f\ g\ dx$$

obtenemos un espacio de Hilbert en el que $\sqrt{\ H_2}$ es la distancia entre dos funciones. Además, para la esfera de radio unidad

$$E = \{\ f\ |\ f = \sqrt{p},\ p \text{ es densidad de probabilidad}\ \}$$

entonces $M(p,q) = H_2(p,q)^{\frac{1}{2}}$ representa la cuerda, mientras que

$$\mathbf{B}\ (p,q) = \arccos\ < \sqrt{p},\ \sqrt{q}\ > = \arccos\ \rho(p,q) \tag{48}$$

representa el arco que une los puntos p,q sobre la esfera E. (48) es una distancia geodésica sobre E, que en general será más pequeña que la distancia de Rao definida para una clase de densidades $p(x,\theta)$, parametrizada por θ, y que constituyen una subvariedad S de E. Asimismo, la métrica diferencial en S inducida por la distancia de $M(p,q)$ da también la distancia de Rao, como vemos en la sección siguiente. La relación entre las tres distancias es

$$M(p,q) \leq B(p,q) \leq\ R(p,q)$$

6.3. Métricas diferenciales a partir de divergencias

En la sección 5.1. introduciamos una distancia geodésica sobre un modelo estadístico $S = \{ p(X,\theta) \}$, donde $\theta \in \Theta$, utilizando el elemento de arco (25) y la matriz de información de Fisher. De forma análoga, utilizando la métrica diferencial definida por el Hessiano de una divergencia $D_\Phi(p,q)$, donde $p(x,\theta)$ es una familia paramétrica, a lo largo de una dirección del espacio tangente de Θ ,

$$ds^2_\Phi(\theta) = d^2 \{ D_\Phi(p, p) \}(\theta) \tag{49}$$

podemos construir una geometría riemanniana sobre S. Por ejemplo, para la J-divergencia tenemos

$$d^2 \{ J_\Phi(p,p) \}(\theta) = 1/4 \int_\chi \Phi''(p) [dp(\theta)]^2 dx$$

y como

$$dp(\theta) = \sum_{i=1}^{n} \frac{\partial p}{\partial \theta_i} d\theta_i$$

podemos tomar el elemento de arco

$$ds^2_\Phi(\theta) = 1/4 \sum_{i,j=1}^{n} g^\Phi_{ij}(\theta)\, d\theta_i\, d\theta_j$$

siendo

$$g^\Phi_{ij}(\theta) = \int_\chi \Phi''(p) \frac{\partial p}{\partial \theta_i} \frac{\partial p}{\partial \theta_j} dx$$

La matriz $(g^\Phi_{ij}(\theta))$ define un tensor covariante, y si Φ es convexa en T_Φ entonces define una métrica riemanniana sobre Θ. La distancia geodésica entre θ_A y θ_B es la que minimiza (26).

Se obtienen expresiones análogas para el elemento de arco para la K, L y M-divergencias, tomando:

$$g^\Phi_{ij}(\theta) = \int_\chi [\Phi(p)/p]' \frac{\partial p}{\partial \theta_i} \frac{\partial p}{\partial \theta_j} dx \qquad \text{(K-divergencia)}$$

$$g^\Phi_{ij}(\theta) = \int_\chi p^{-1} \frac{\partial p}{\partial \theta_i} \frac{\partial p}{\partial \theta_j} dx \qquad \text{(L-divergencia)}$$

$$g^\Phi_{ij}(\theta) = \int_\chi [(\sqrt{\Phi(p)})']^2 \frac{\partial p}{\partial \theta_i} \frac{\partial p}{\partial \theta_j} dx \qquad \text{(M-divergencia)}$$

Observaciones:

1) Para la L-divergencia se verifica

$$g_{ij}^{\Phi}(\theta) = E_{\theta}\left(\frac{\partial}{\partial \theta_i} \log p\,(X,\theta) \; \frac{\partial}{\partial \theta_j} \log p\,(X,\theta) \right)$$

luego es fácil ver que

$$ds_{\Phi}^{2}\,(\theta) = 2\Phi''(1)\, ds^{2}(\theta)$$

Si $\Phi''(1) > 0$, la métrica coincide (salvo una constante) con la métrica informacional o distancia de Rao. Encontramos un resultado similar para la distancia de Matusita.

2) Para la clase de funciones Φ_α definidas en (43) se obtiene

$$g_{ij}^{(\alpha)}(\theta) = \int_{\chi} p^{\alpha} \; \frac{\partial}{\partial \theta_i} \log p \; \frac{\partial}{\partial \theta_j} \log p \, dx \qquad (50)$$

que da lugar a la métrica informacional de orden α. En este caso, las cuatro métricas coinciden (salvo constantes). En particular, si $\alpha=1$, en todos los casos obtenemos la métrica informacional o distancia de Rao.

3) Las distancias obtenidas son todas invariantes por transformaciones admisibles de los parámetros. Para ciertas funciones Φ las distancias son además invariantes frente a transformaciones admisibles de las variables aleatorias. Por ejemplo, para la K-divergencia se cumple para $\Phi(u) = au \log(u) + bu + c$. En realidad se verifica esta propiedad para aquellas funciones Φ tales que $(g_{ij}^{\Phi}(\theta))$ es la matriz de información de Fisher. En efecto (Cuadras, *et al.*, 1985; Oller y Cuadras, 1987) la invarianza para las variables es una cualidad que prácticamente sólo se cumple para la distancia de Rao.

Para más información sobre este tema, véase Burbea y Rao (1982*a,b*), Salicrú (1987). La construcción de medidas paramétricas de información sobre funciones de densidad $p(x,\theta)$ a partir de medidas no paramétricas, había sido planteada de manera análoga por diversos autores (Kagan, Vajda, Aggarwal y Boeke). Véase Ferentinos y Papaionnau (1981).

7. ALGUNAS APLICACIONES

7.1. Biología

La aplicación de las distancias estadísticas a la biología, especialmente antropología y genética, son muy numerosas. Con la obtención de distancias entre poblaciones, especies, razas geográficas, etc., se han abordado problemas de sistemática, filogenia y clasificación taxonómica.

Pearson (1926) utiliza un coeficiente de semejanza racial para diferenciar razas humanas (ver sección 3.1.). Pero la distancia de Mahalanobis (14) es la más utilizada, especialmente combinada con el análisis canónico de poblaciones. Pueden verse aplicaciones a la biología sistemática en Seal (1964), Reyment (1973), Petitpierre y Cuadras (1977), Cirer (1987).

La utilización de distancias basadas en coeficientes de similaridad, combinadas con el análisis de coordenadas principales y el análisis de conglomerados, han significado una importante herramienta metodológica en Botánica, Zoología, Microbiología y Ecología. Los trabajos de Escarré (1973), Cantón y Sancho (1976) son bien representativos en este sentido.

En ecología se han utilizado también las distancias basadas en la métrica de Canberra, que presenta ciertas ventajas. Véase Lance y Williams (1967), Legendre y Legendre (1979) y una interesante aplicación en Del Castillo (1986).

Para una visión general del tema se recomienda consultar Constandse (1972), Goodman (1972), Sneath y Sokal (1973), Cuadras (1980). Respecto de la dimensión significativa o número de "clusters" significativos, véase Cuadras (1987).

7.2. Genética

Las llamadas distancias genéticas entre poblaciones son distancias estadísticas que se calculan sobre datos basados en frecuencias genéticas en loci polimórficos. Se trata, por lo tanto, de medidas que cuantifican la diferencia genética entre poblaciones en términos de las frecuencias alélicas de diferentes loci, es decir, de distancias "genotípicas", que se distinguen de otras (como el índice de semejanza racial de K. Pearson) que se considerarían distancias "fenotípicas".

Dados n sucesos mutuamente excluyentes $A_1,...,A_n$ una distancia genética es una medida de divergencia entre dos distribuciones de probabilidad $p = (p_1,...,p_n)'$, $q = (q_1,...,q_n)'$. Si se conoce una matriz de covarianzas Σ, asociada a una distribución $a = (a_1,...,a_n)'$, es decir,

$$\sigma_{ij} = \begin{cases} a_i (1 - a_j) & i = j \\ - a_i a_j & i \neq j \end{cases}$$

entonces podemos utilizar una distancia de Mahalanobis singular

$$(p - q)' \ \Sigma^{-} (p - q) \tag{51}$$

siendo Σ^{-} una g-inversa de Σ. Sin embargo, Σ depende de la distribución de $A_1,...,A_n$, que es distinta en cada población. Podríamos entonces, tomar, por ejemplo, $a_i = (p_i + q_i) / 2$, con lo cual se llega a la expresión

$$2 \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{(p_i + q_i)} \tag{52}$$

que ya había sido propuesta en términos practicamente iguales por Sanghvi (1953). Otros autores han propuesto diferentes variantes de (52) que difieren en la forma de estimar Σ (Steinberg *et al.*, 1966; Balakrishnan y Sanghvi, 1968; Kurcynski, 1970).

Otro enfoque, quizás más razonable dadas las propiedades discutidas en la sección 5.3., consiste en definir una distancia geodésica entre p y q, es decir, proporcional a

$$\arccos \left(\sum_{i=1}^{n} \sqrt{p_i Q_i} \right) \tag{53}$$

Esta es la distancia de Bhattacharyya (1946), cuya interpretación geométrica es un arco de circunferencia máxima entre dos puntos de una esfera unidad en R^n. Puede probarse también que (53) viene a ser una aproximación asintótica de una distancia de Mahalanobis (Mardia *et al.*, 1979). Véase también (48).

La distancia (53) ha sido aplicada a la Genética (a sugerencia de R.A. Fisher) por Edwards y Cavalli-Sforza (1964) directamente o tomando la cuerda en lugar del arco en Cavalli-Sforza y Edwards (1967). También se han utilizado distancias proporcionales a

$$\sum_{i=1}^{n} | p_i - q_i | \tag{54}$$

en Prevosti *et al.* (1975) y Thorpe (1979) (véase sección 6.1.).

Todas estas distancias son, de hecho, distancias geométricas en espacios de dimensión igual al número de alelos en un locus. Pero si hay una mutación, debemos añadir una dimensión mientras que si un alelo se

extingue, debemos sustraer una dimensión. Como considerar el conjunto, practicamente ilimitado, de posibles alelos en un locus complicaría excesivamente el problema, Nei (1971, 1972) propone una distancia genética para estimar el número de sustituciones de alelos por locus

$$D = - \log J_{12} / (J_1 J_2)^{\frac{1}{2}}$$

siendo J_1, J_2 y J_{12} los valores esperados de Σp_i^2, Σq_i^2 y $\Sigma p_i q_i$. Obsérvese que D es una distancia entre poblaciones basada en la medida de diversidad de Gini-Simpson $1-\Sigma p_i^2$ (sección 6.1.).

En genética se han utilizado también, otros tipos de distancias. Frommel y Holzhütter (1985) consideran una distancia entre aminoácidos inversamente proporcional a la probabilidad de reemplazamiento mutuo. Coll, Cuadras y Egozcue (1980) utilizan una distancia del tipo de Mahalanobis para situar los cromosomas humanos en el plano metafísico.

Los inconvenientes y ventajas de las distancias genéticas han sido objeto de polémica (Balakrishnan y Sanghvi, 1968; Fitch y Neel, 1969; Edwards, 1971; Goodman, 1972; Prevosti, 1974; Nei, 1987). Las tres últimas referencias (Prevosti, Goodman, Nei) contienen un amplio estudio sobre las distancias genéticas. Véase también Constandse (1972).

7.3. Psicología

La medida de la proximidad entre objetos psicológicos y su representación geométrica, se consigue a través del concepto de distancia y de disimilaridad, hasta el punto de que su estudio y aplicaciones han desembocado en una rama del análisis de datos con fuerte personalidad: el llamado "multidimensional scaling" (MDS). La versión métrica del MDS, en el que se supone que la matriz de distancias psicológicas es euclídea (Teorema 1), fue desarrollada por Torgerson (1958). Sin embargo, las distancias entre objetos psicológicos son, a menudo, el resultado de medidas subjetivas del tipo: 0=idéntico, 1=muy parecido, 2=bastante parecido, 3=poco parecido, 4=muy diferente. Las distancias resultantes suelen ser no euclídeas, lo que motivó a Shepard (1962 *a,b*) y Kruskal (1964 *a,b*) a desarrollar métodos para convertir las distancias en euclídeas por transformación monótona de las mismas, de modo que se preservara la preordenación entre los objetos a representar. Este es el MDS no métrico, tantas veces utilizado en psicometría. Véase aplicaciones en Romney *et al.* (1972), Wish y Carroll (1982), Dunn-Rankin (1983). Para una exposición teórica del MDS véase De Leeuw y Heiser (1982), Cuadras *et al.* (1985).

La ordenación de objetos a lo largo de un continuo psicológico, sugiere una interesante aplicación del concepto distancia en Psicología. Supongamos que un grupo de sujetos tienen que ordenar n objetos $A_1,...,A_n$ de acuerdo con la cierta escala de valores $\theta_1,...,\theta_n$. Sea

$$p_{ij} = P\ (A_i > A_j) \quad i,j = 1,...,n$$

la proporción de sujetos que prefieren A_i sobre A_j en el sentido de que $\theta_i > \theta_j$. El modelo de Thurstone (1927) supone que

$$p_{ij} = \int_{-\infty}^{\theta_i - \theta_j} \Phi(y)\, dy$$

siendo $\Phi(y)$ la función de densidad normal standard. Obsérvese que si $\theta_i > \theta_j$ entonces $p_{ij} > 0.5$, mientras que si $\theta_i < \theta_j$ entonces $p_{ij} < 0.5$. La estimación de la escala $\theta_1,...,\theta_n$ presenta cierto grado de complejidad (Coombs *et al.*, 1981). Una alternativa consiste en definir la distancia (Davison, 1983)

$$d(A_i,A_j) = |\, p_{ij} - 0.5\, |$$

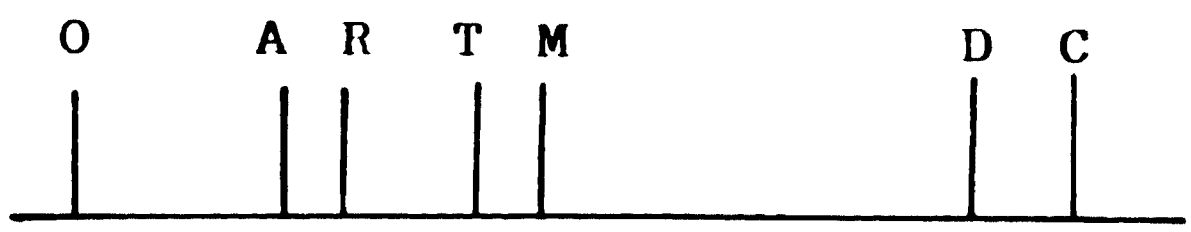
que es función monótona de $|\, \theta_i - \theta_j\, |$. La representación de los objetos $A_1,...,A_n$ mediante MDS, a lo largo de la primera dimensión, proporciona la escala deseada. Como $|\, p_{ij} - 0.5\, |$ está acotado, una generalización razonable es

$$d(A_i,A_j) = |\, \Psi^{-1}\, (p_{ij})\, | \tag{55}$$

donde Ψ es la función de distribución normal standard.

La tabla 2 contiene las frecuencias sobre 262 estudiantes al comparar los defectos de los profesores de Estadística. Aplicando MDS sobre la distancia no euclídea (55) se obtiene la ordenación ilustrada en la Figura 9.

Figura 9



Obsérvese que D y C destacan como peores defectos sobre los demás.

TABLA 2

	O	D	R	M	A	C	T
O	–	71	159	121	150	54	119
D	191	–	196	156	193	138	175
R	103	66	–	89	112	83	107
M	141	106	173	–	160	83	111
A	112	69	150	102	–	75	91
C	208	124	179	179	187	–	175
T	143	87	155	151	171	87	–

O = Falta de orden en las explicaciones.
D = Conoce poco la materia (no sabe resolver dudas).
R = Poca o mala relación con los alumnos.
M = No sabe motivar a los alumnos.
A = Falta de amenidad en las clases.
C = Poca claridad al explicar o escribir.
T = Exceso de teoría (pocos ejemplos y aplicaciones).

7.4. Arqueología

Supongamos que estamos interesados en ordenar cronológicamente n objetos arqueológicos $A_1,...,A_n$. Podemos imaginar que los n objetos están situados sobre una curva m-dimensional $x=x(t)$, donde t representa el tiempo. En otras palabras, a cada objeto le asignamos unas coordenadas euclídeas

$$A_i : (x_1(t_i), x_2(t_i),...,x_m(t_i)) \quad i = 1,..., m$$

donde t_i representa el tiempo cronológico relativo a A_i.

Los objetos presentarán una ordenación cronológica

$$A_{i_1} < A_{i_2} < \ldots < A_{i_n}$$

si se verifica

$$t_{i_1} < t_{i_2} < \ldots < t_{i_n}$$

Este problema, aparentemente complicado, se puede resolver mediante una matriz de distancias. En efecto, supongamos que en relación a ciertas características cualitativas y cuantitativas, podemos definir una matriz de distancias $\Delta = (\delta_{ij})$, donde δ_{ij} es la distancia entre A_i y A_j. Entonces es de esperar que la distancia será pequeña para objetos próximos en el tiempo y grande para objetos lejanos en el tiempo. La representación de los objetos por MDS permitirá su ordenación cronológica. Generalmente, la representación 2-dimensional adopta la forma de herradura (Kendall, 1971).

Spaulding (1971) propone el siguiente ejemplo. Se desean ordenar 5 herramientas cortantes A, B, C, D y E que han sido fabricadas utilizando piedra, bronce o hierro de acuerdo con la matriz de incidencia:

	Piedra	Bronce	Hierro
A	0	1	0
B	1	1	0
C	0	1	1
D	0	0	1
E	1	0	0

Aplicando análisis de coordenadas principales a la matriz de distancias calculada utilizando (5), donde s_{ij} es el coeficiente de similaridad de Jaccard, se obtiene la representación de la figura 8, que sugiere que la datación relativa de las herramientas es

$$E < B < A < C < D$$

que concuerda con el orden cronológico natural: piedra, piedra-bronce, bronce, bronce-hierro, hierro. Véase también Orton (1988).

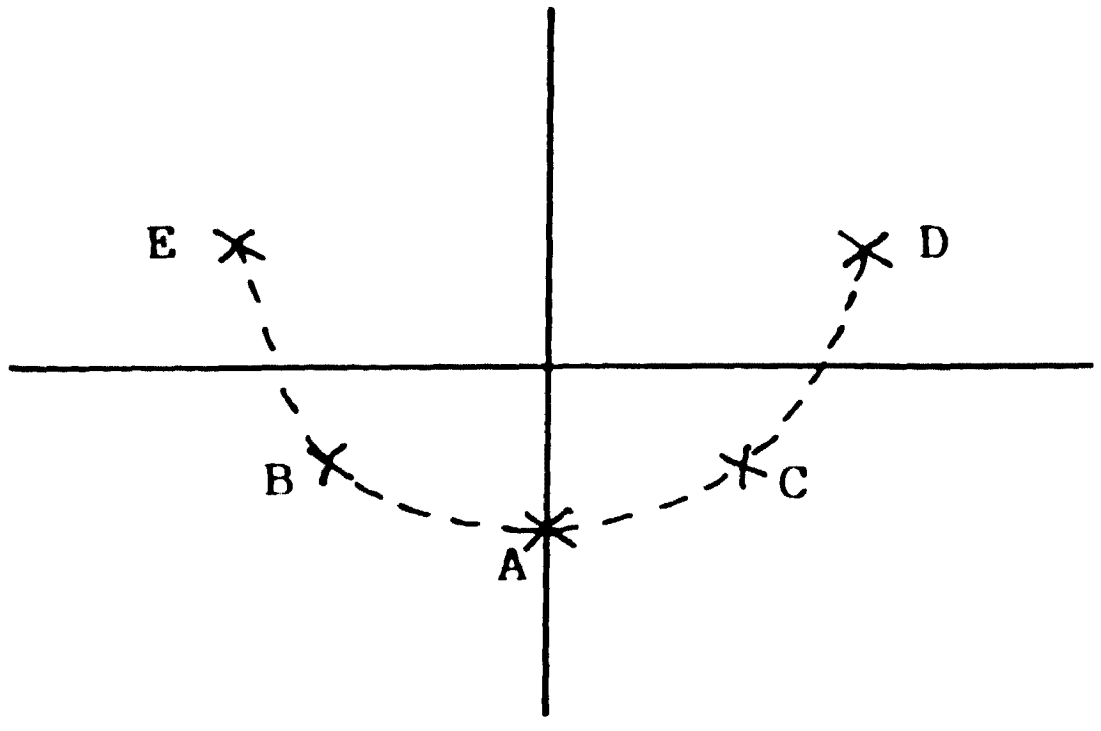


Fig. 10.– Ordenación cronológica de 5 herramientas teniendo en cuenta la presencia de diversos materiales (piedra, bronce, hierro).

7.5. Lingüística

El análisis de las dimensiones semánticas latentes en un conjunto de palabras, es otra interesante aplicación de las distancias estadísticas. En este caso no se trata de encontrar una dimensión lineal (como en el modelo de Thurstone, sección 7.3.), o una dimensión curvilínea (como en el caso de la ordenación cronológica), sino diversas dimensiones que permitan explorar y ordenar el conjunto de palabras estudiadas.

Partiendo de una matriz de distancias sobre 23 adjetivos del castellano relacionados con las nociones de peso y extensión espacial, Manzano y Costermans (1976), aplicando MDS, obtienen 6 ejes que permiten ordenar los adjetivos a lo largo de otras tantas dimensiones semánticas. En los extremos de cada eje se situan dos adjetivos opuestos, comunicados por un gradiente de adjetivos intermedios.

Otros ejemplos donde se aplican distancias estadísticas para explorar dimensiones y estructuras semánticas (nombres de colores, familia de verbos "to have", nombres de profesiones, etc.), pueden verse en Romney *et al.* (1972). Véase también Morgan (1981).

7.6. Manova y comparación de experimentos

Consideremos el modelo lineal del análisis multivariante de la varianza

$$Y = X\ B + E$$

donde Y(*nxp*) es una matriz de datos, X(*nxm*) es una matriz de diseño, B(*mxp*) es una matriz de parámetros, E(*nxp*) es una matriz de errores aleatorios. E contiene *n* filas estocásticamente independientes, cada una de ellas con distribución $N_p(0, \Sigma)$.

El concepto de distancia puede ser útil para estudiar diversos aspectos de MANOVA. Por ejemplo, consideremos *q* funciones paramétricas estimables multivariantes (*fpem*)

$$\psi_i = P_i' B \qquad i = 1,...,q$$

donde los vectores fila $P_i' = (p_{i1},...,p_{im})$ son combinación lineal de las filas de X. Como es sabido, existe entonces un estimador insesgado y de dispersión mínima para cada Ψ_i (Teorema de Gauss-Markov). Generalizando la distancia de Mahalanobis (14) entre poblaciones, Cuadras (1974) define la distancia (al cuadrado) entre *fpem*

$$M^2(i,j) = (\Psi_i - \Psi_j)'\ \Sigma^{-1}(\Psi^i - \Psi^j) \tag{56}$$

La distancia (56), combinada con técnicas de reducción de la dimensión en análisis de datos, permite representaciones euclídeas de los niveles de un determinado factor en un diseño multifactorial, con aplicaciones a Farmacología (Vallejo *et al.*, 1975; Peris *et al.*, 1975; Ballús *et al.*, 1980), la Sistemática (Cuadras, 1981 *a*) y la Agricultura (Oller y Cuadras, 1982*b*). Véase también Cuadras (1977, 1981*b*).

Una segunda aplicación consiste en definir distancias entre dos modelos lineales $Y_i = X_i B_i + E_i$, $i=1,2$. Cuadras y Rios (1986) y Rios y Cuadras (1986) proponen diversas distancias, estudiando diferentes casos (univariante, heterocedástico, diferente matriz de diseño, multivariante, etc.) que

relacionan con ciertos contrastes de hipótesis. La distancia (al cuadrado) para el caso $X_1 = X_2 = X$ es

$$L^2 = tra \{ \Sigma^{-1} (B_1 - B_2)' X X' (B_1 - B_2) \} \qquad (57)$$

Véase también Burbea y Oller (1988) y algunas aplicaciones en Cuadras *et al.*, (1985), Rios y Oller (1988). La comparación de experimentos asi como la equivalencia entre experimentos mediante distancias, ha sido estudiada por Le Cam (1975).

7.7. Regresión cualitativa

Supongamos que deseamos plantear la regresión múltiple de una variable cuantitativa Y sobre p variables cualitativas (binarias, categóricas, ordinales, etc.), y que disponemos de una muestra de n individuos. Un posible camino, que evitaría la asignación de valores cuantitativos arbitrarios para las variables cualitativas asi como los problemas de colinealidad, consiste en obtener una matriz de distancias euclídeas $\Delta = (\delta_{ij})$, donde δ_{ij} es la distancia entre los individuos i,j calculada a partir de la información entre ambos contenida en las p variables cualitativas. Para calcular δ_{ij} podemos utilizar (6) (variables binarias), (32) (variables categóricas) o el coeficiente general propuesto por Gower (1971).

Sea ahora la matriz X(nxm) verificando (18), obtenida a partir de la descomposición espectral de B (teorema 1). Con la matriz X convertimos la información cualitativa sobre cada individuo en la información cuantitativa contenida en las filas de X, es decir, cada fila $x_i = (x_{i1},...,x_{im})$ de X resume la información cualitativa sobre el individuo i en relación con los demás individuos, verificándose $\delta^2_{ij} = (x_i - x_j)'(x_i - x_j)$.

Si $y = (y_1,...,y_n)'$ es el vector de observaciones de la variable Y, proponemos el modelo de regresión múltiple

$$y_i = \mu + x_{i1}\,\beta_1 + ... + x_{im}\,\beta_m + e_i \qquad i=1,...,n$$

Se puede entonces demostrar (Cuadras, 1988) lo siguiente:

a) Si las variables son binarias y el coeficiente de similaridad utilizando es el de Sokal y Michener (sección 2.1.), entonces el método propuesto y la predicción obtenida mediante regresión múltiple clásica coinciden.

b) El coeficiente de determinación de Y sobre las variables cualitativas es

$$R^2 = Y' X \Lambda^{-1} X' Y \,/\, n\, s_y^2 \qquad (58)$$

donde $\Lambda = diag\ (\lambda_1,...,\lambda_m)$ contiene los valores propios de B.

c) Consideremos ahora el problema de predecir el valor y_{n+1} de la variable dependiente, conocidas las características cualitativas de un nuevo individuo $n+1$. Entonces podremos calcular las distancias del individuo $n+1$ a los demás individuos:

$$\delta_{1n+1},\ \delta_{2n+1},\ \ldots,\ \delta_{nn+1}$$

Indicando $d = (\delta^2_{1n+1}\ \ldots\ \delta^2_{nn+1})'$, $b = (b_{11}\ \ldots\ b_{nn})'$, siendo $b_{ii} (i=1,...,n)$ los elementos diagonales de B, y siendo finalmente B^- una g-inversa de B, la predicción es

$$y_{n+1} = \bar{y} + \frac{1}{2}\ (b - d)'\ B^-\ y \qquad (59)$$

7.8. Contrastes de hipótesis

Ciertas medidas de disimilaridad o divergencia entre distribuciones son útiles para construir contrastes de hipótesis. La más conocida es

$$D_n = \sup_{-\infty < x < \infty} |\ S_n(x) - F(x)\ |$$

que mide, para una muestra aleatoria simple de tamaño n, la discrepancia entre la función de distribución empírica $S_n(x)$ y la teórica. D_n interviene en el test de Kolmogorov-Smirnov de bondad de ajuste de los datos a una distribución.

Los contrastes sobre las medias en poblaciones normales multivariantes que utilizan la T^2 de Hotelling, están basados en la distancia de Mahalanobis. Así, en muestras de tamaño N, la hipótesis $H_o: \mu = \mu_o$ se decide mediante el estadístico

$$T^2 = N\ (\bar{x} - \mu_o)'\ S^{-1}\ (\bar{x} - \mu_o)$$

mientras que para el contraste $H_o: \mu_1 = \mu_2$ se utiliza

$$T^2 = \frac{N_1 N_2}{N_1 + N_2}\ (\bar{x} - \bar{y})'\ S^{-1}\ (\bar{x} - \bar{y})$$

En el caso univariante, ambos contrastes son equivalentes al conocido test t de Student. En general, las distancias estadísticas pueden aplicarse para construir un contraste que sirva para comparar dos distribuciones F,G.

Sea $\delta(F,G)$ una distancia que vale cero si $F \equiv G$. Supongamos que existe un estadístico V que es función de una estimación $\hat{\delta}(F,G)$ cuya distribución es conocida cuando $\delta(F,G) = 0$. Entonces las distribuciones son distintas si V es significativo. En el caso paramétrico se puede utilizar la distancia de Rao, pudiéndose demostrar (Oller, 1983) que

$$V = \frac{N_1 N_2}{N_1+N_2} \hat{\delta}^2(F,G)$$

sigue (asintóticamente) la distribución *ji*-cuadrado con *p*(=número de variables) grados de libertad. En el caso no paramétrico se puede utilizar una divergencia. Por ejemplo, dadas dos distribuciones univariantes F,G, para la divergencia

$$\delta(F,G) = \int_{R^2} (F(x) - G(y))^2 \, d\left(\frac{F(x)+G(y)}{2} \right)$$

existe un U-estadístico para estimar $\delta(F,G)$ (Cuadras, 1986). Análogamente, el estadístico U de Mann-Whitney mide la discrepancia entre $P(X < Y)$ y el valor $\frac{1}{2}$.

Para otros aspectos sobre este tema véase Rao(1982). Obsérvese que, en un contexto similar, se pueden detectar "outliers" utilizando distancias.

7.9. Asociación estocástica y máxima correlación

Ciertas divergencias permiten medir el grado de dependencia estocástica entre dos variables aleatorias X,Y, con distribución $H(x,y)$ y marginales $F(x)$, $G(y)$. Por ejemplo,

$$\theta = \int_{R^2} (H(x,y) - F(x)\, G(y))^{\alpha} \, d\mu$$

donde $d\mu$ puede ser $dxdy$, $dF(x)dF(y)$ o $dH(x,y)$. En el caso $\alpha = 1$, $d\mu = dxdy$, θ es la covarianza entre X,Y. También es posible relacionar θ con los coeficientes de correlación por rangos de Kendall y el grado de correlación de Spearman (Cuadras, 1985).

Fréchet (1957) considera una clase de distancias entre X, e Y

$$d_H(X,Y) = E_H [f(|X - Y|)] \qquad (60)$$

donde f es una función creciente subaditiva en R^+ con $f(0) = 0$. Dadas las distribuciones marginales F,G, el problema de encontrar las distribuciones

conjuntas H tales que (60) es un valor extremo, ha sido considerado por diversos autores (Hoeffding, Fréchet, Bass, Dall'Aglio, Cambanis, Tchen, etc.).

Por ejemplo, para $f(u)=u^2$ se verifica

$$d_{H+}(X,Y) \leq d_H(X,Y) \leq d_{H-}(X,Y)$$

donde

$$H^+(x,y) = min \{ F(x), G(y) \}$$

$$H^-(x,y) = min \{ F(x) + G(y) - 1, 0 \}$$

son las distribuciones, llamadas cotas de Frechet, cuyas marginales son F y G. H^- , H^+ son las distribuciones que dan mínima y máxima correlación entre X , Y (Hoeffding, 1940). Las cotas extremas de (60) para $f(u)=u^\alpha$, $\alpha \geq 1$, han sido estudiados por Dall'Aglio (1972). En general, este problema conecta con el de la construcción de distribuciones bivariantes con marginales, dadas, y tiene interesantes aplicaciones en programación lineal, mecánica cuántica, simulación estadística, biometría, etc. Véase Ruiz-Rivas et al. (1979), Cuadras y Augé (1981), Cuadras (1985), Sánchez (1986), Ruiz-Rivas y Cuadras (1988).

Finalmente la distancia de Rao puede sernos útil para definir una medida de asociación entre dos vectores aleatorios $X=(X_1,...,X_p)$, $Y=(Y_1,...,Y_q)$. Supongamos que la distribución es $N_{p+q}(\mu , \Sigma)$ con

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad ran\ \Sigma_{12} = r$$

Consideremos la distribución $N_{p+q}(\mu, \Sigma_o)$, siendo

$$\Sigma_o = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

Desde luego, si $\Sigma = \Sigma_o$, X es independiente de Y. La distancia de Rao entre ambas distribuciones (Carmona y Cuadras, 1987) es

$$R(X,Y) = [\frac{1}{2} \sum_{i=1}^{r} (log(1 - \rho_i^2))^{\frac{1}{2}} - \sum_{i=1}^{r} log(1+\rho_i) (1-\rho_i)]^{\frac{1}{2}}$$

donde $\rho_1 \geq \ldots \geq \rho_r$ son las correlaciones canónicas entre X,Y. Entonces R(X,Y) puede interpretarse como un índice de asociación estocástica entre X e Y, que puede generalizarse a cualquier otra familia de distribuciones.

SUMMARY

STATISTICAL DISTANCES

This paper is concerned with the application of distance functions to statistics and data analysis. Closed form expressions of distances and similarity coefficients between individuals and populations are exposed and discussed. Some applications to biology, genetics, psichology, archaeology, linguistics, manova, regression and stochastic association are also included.

Key words: Mahalanobis distance, Rao distance, ultrametric distance, similarity coefficients, measures of divergence.

AMS 1980: 62H25; 62H30; 62P99.

8. BIBLIOGRAFIA

AMARI, S. (1985). *Differential geometrical methods in statistics.* Lecture notes in statistics, 28. Springer Verlag, Berlín.

ARCAS, A. (1987). *Sobre la representación de un conjunto mediante arboles aditivos.* Questiio, 11 (2), 39-50.

ARCAS, A. y CUADRAS, C. M. (1987). *Métodos geométricos de representación mediante modelos en árbol.* Pub. de Bioest. y Biomat., 20, Universidad de Barcelona.

ATKINSON, C. y MITCHELL, A. F. S. (1981). *Rao's distance measure.* Sankhya, 43 A, 345-365.

BALAKRISHNAN, V. y SANGHVI, L. D. (1968). *Distance between populations on the basis of attribute data.* Biometrics, 24, 859-865.

BALLUS, C., CUADRAS, C. M., MALGA, A., SANCHEZ-TURET, M. y VALLVE, C. (1980). *Estudio de dos ansioliticos (Diazepam y Clobazam) mediante una prueba de conducción de automóviles.* Rev. Dep. Psiquiatria, Fac. Medicina, Barcelona, 7, 107-122.

BARTHELEMY, J. P. y GHENOCHE, A. (1988). *Les arbres et les representations des proximites.* Masson, Paris.

BENZECRI, J. P. (1965). *Problemes et Methodes de la Taxinomie.* Pub. Inst. Statistique Univ. Paris, Rennes et Paris.

BENZECRI, J. P. (1976). *L'Analyse des Donnees I. La Taxonomie. L'Analyse des Donnees.* II. *L'Analyse des Correspondances.* Dunod, Paris.

BERNARDO, J. M. (1981). *Bioestadística. Una perspectiva Bayesiana.* Vicens-Vives, Barcelona.

BERNARDO, J. M. (1987). *Approximations in statistics from a decision-theoretical viewpoint.* Probability and Bayesian Statistics (R. Vierte, Ed.) 53-60, Plenum, N. York.

BHATTACHARYYA, A. (1946). *On a measure of divergence between two multinomial populations.* Sankhya, 7, 401-406.

BISHOP, Y. M. M., FIENBERG, S. E. y HOLLAND, P. W. (1975). *Discrete multivariate analysis; Theory and Practice.* Mit Press, Cambridge, Mass.

BUNEMAN, P. (1971). *The recovery of trees from measures of dissimilarity*. En: Mathematics in the Archaeological and Historical Sciences (F. R. Hodson, D. G. Kendali y P. Tautu, Eds.), 387-395, Edinburgh University Press.

BURBEA, J. (1986). *Informative geometry of probability spaces*. Expositiones mathematicae, 4, 347-378.

BURBEA, J. y OLLER, J. M. (1988). *Information metric for univariate linear elliptic models*. Stat. and Decisions, 6, 209-221.

BURBEA, J. y RAO, C. R. (1982*a*). *Entropy differential metric, distance and divergence measures in probability spaces: a unified approach*. J. of Multivariate Analysis, 12, 575-596.

BURBEA, J. y RAO, C. R. (1982*b*). *Differential metrics in probability spaces*. Prob. math. statist., 3, 115-132.

CAILLIEZ, F. (1983). *The analytical solution of the additive constant problem*. Psychometrika, 48 (2), 305-308.

CAILLIEZ, F. y PAGES, J. P. (1976). *Introduction a l'analyse des donnees*. Smash, Paris.

CALVO, M. (1988). *Sobre la geometría informacional del modelo normal multivariante. Aplicaciones a la Biología*. Tesis doctoral, Universidad de Barcelona.

CANTON, E. y SANCHO, J. (1976). *Análisis numérico de un grupo de Pseudomonas aeróbicos*. Microbiol. Española, 29, 59-73.

CARMONA, F. y CUADRAS, C. M. (1987). *Measures of multivariate association based on the Rao's distance*. Analyse statistique des grandes tableaux et donnees d'enquete (T. Aluja, M. Marti, Eds.), 181-184, Barcelona.

CARROLL, J. D. (1976). *Spatial, non-spatial and hybrid models for scaling*. Psychometrika, vol. 41(4), 439-463.

CAVALLI-SFORZA, L. L. y EDWARDS, A. W. F. (1967). *Phylogenetic analysis: Models and estimation procedures*. Evolution, 21, 550-570.

CIRER, A. M. (1987). *Aplicación de técnicas estadísticas multivariantes a las poblaciones del Lacertido Podarcis Pityusensis (Bosca, 1883)*. Rev. Esp. de Herpetología, 2, 145-163.

CLARK, P. F. (1952). *An extension of the coefficient of divergence for use with multiple characters*. Copeia 1952, 61-64.

COLL, M. D., CUADRAS, C. M. y EGOZCUE, J. (1980). *Distribution of human chromosomes on the methaphase plate: Symmetrical arrangement in human male cells*. Genet. Res., 36, 219-234.

CONSTANDSE, T. S. (1972). *Coefficients of biological distance*. Anthrop. pub., Oosterhout. Humanities Press, N. York.

COOMBS, C. H., DAWES, R. M. y TVERSKY, A. (1981). *Introducción a la Psicología Matemática*. Alianza universidad textos, Madrid.

COOPER, L. G. (1972). *A new solution to the additive constant problem in metric multidimensional scaling*. Psychometrika, 37, 311-322.

CORTER, J. E. (1982). *ADDTREE/P: a PASCAL program for fitting additive trees based on Sattah and Tversky's ADDTREE algorithm*. Behavior research and instrumentation, 14(3), 353-354.

CRITCHLEY, F. (1985). *Dimensionality theorem in multidimensional scaling and hierarchical cluster analysis*. Fourth int. symp. data analysis and informatics. Vol. 1. Versailles: Inst. nat. de recherche en inform. et en autom, 85-110.

CSISZAR, I. (1963). *Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten*. Publ. Math. Inst. Hungar. Acad. Sci. 8, Ser. A, 85-108.

CSISZAR, I. (1967). *Information-type measures of difference of probability distributions and indirect observations*. Studia Sci. Math. Hungar. 2, 299-318.

CSISZAR, I. (1972). *A class of measures of informativity of observation channels*. Periodica Math. Hungar. 2, 191-213.

CSISZAR, I. (1975). *I-divergence geometry of probability distributions and minimization problems*. Ann. Probab. 3, 146-158.

CUADRAS, C. M. (1974). *Análisis discriminante de funciones paramétricas estimables*. Trab. estad. inv. oper., 25(3), 3-31.

CUADRAS, C. M. (1977). *Sobre la reducción de la dimensión en análisis estadístico multivariante*. Trab. estad. I. Op., 28, 63-76.

CUADRAS, C. M. (1980). *Metodes de representacio de dades i la seva aplicacio en Biologia*. Col. Soc. Catalana Biología, 13, 95-133.

CUADRAS, C. M. (1981*a*). *Métodos de Análisis Multivariante*. Eunibar, Barcelona.

CUADRAS, C. M. (1981*b*). *Análisis y representación multidimensional de la variabilidad*. Inter. sym. concept. meth. paleo., Barcelona, 287-297.

CUADRAS, C. M. (1983). *Análisis algebráico sobre distancias ultramétricas*. Actas 44 per. de sesiones del Instituto Internacional de Estadística, Madrid. cont. libres, vol. II, 554-557.

CUADRAS, C. M. (1985). *Sobre medidas de dependencia estocástica invariantes por transformaciones monótonas*. Hom. F. D'A. Sales, cont. cient., Fac. Matem. Univ. Barcelona, 28-47.

CUADRAS, C. M. (1986). *Problemas de Probabilidades y Estadística. Vol. 2*. PPU, Barcelona.

CUADRAS, C. M. (1987). *Dimensionality and number of clusters in multivariate analysis*. Analyse statistique des grandes tableaux et donnees d'enquete (T. Aluja, M. Marti, Eds.), 53-67, Barcelona.

CUADRAS, C. M. (1988). *Métodos estadísticos aplicables a la reconstrucción prehistórica*. Munibe. 6, 25-33.

CUADRAS, C. M. y AUGE, J. (1981). *A continuous general multivariate distribution and its properties*. Comm. in stat., theor. meth., A10(4), 339-353.

CUADRAS, C. M. y CARMONA, F. (1983). *Dimensionalitat euclidiana en distancies ultrametriques*. Questiio, 7(1), 353-358.

CUADRAS, C. M., OLLER, J. M., ARCAS, A. y RIOS, M. (1985). *Métodos Geométricos de la Estadística*. Questiio, 9(4), 219-250.

CUADRAS, C. M. y OLLER, J. M. (1987). *Eigenanalysis and metric multidimensional scaling on hierarchical structures*. Questiio, 11(3), 37-58.

CUADRAS, C. M. y RIOS, M. (1986). *A distance between multivariate linear models and its properties*. II catalan intern. symp. on stat. Vol. 2, cont. paper, 81-84, Barcelona.

CUADRAS, C. M. y RUIZ-RIVAS, C. (1980). *Una contribución al análisis de proximidades*. Pub. Secc. Matem. Univ. Au. Barcelona, 22, 103-106.

D'ANDRADE, R. G., QUINN, N. R., NERLOVE, S. B. y ROMNEY, A. K. (1972). *Categories of Disease in American-English and Mexican-Spanish*. En: Multidimensional Scaling. Vol. 2. Applications (A. Kimball Ronney, Ed.) Seminar Press, N. York.

DALL'AGLIO, G. (1972). *Frechet classes and compatibility of distribution functions*. Symposia mathematica, 9, 131-150, Academic Press, N. York.

DAVISON, M. L. (1983). *Multidimensional scaling*. J. Wiley, N. York.

DAWID, A. P. (1975). *Discussion on Professor Efron's paper (1975)*. Ann. Statist., 3, 1231-1234.

DE LEEUW, J. y HEISER, W. (1982). *Theory of multidimensional scaling*. En: Handbook of Statistics. Vol. 2, (P. R. Krishnaiah, L. N. Kanal, Eds.), North-Holland pub. co., Amsterdam.

DEL CASTILLO, M. (1986). *Nueva aproximación metodológica al estudio de la biogeografía de los peces epicontinentales*. Oecología Acuática, 8, 71-94.

DUNN-RANKIN, P. (1983). *Scaling Methods*. Lea, Hillsdale, N. Jersey.

EDWARDS, A. W. F. (1971). *Distances between populations on the basis of genic frequencies*. Biometrics, 27, 873-881.

EDWARDS, A. W. F. y CAVALLI-SFORZA, L. L. (1964). *Reconstruction of evolutionary trees*. En: Phenetic and Philogenetic classification (V. H. Heywood, J. Mcneile, Eds.), 67-76. The Systematics Assoc. Publ. no. 6, London.

EFRON, B. (1975). *Defining the curvature of a statistical problem (with applications to second order efficiency)*. The Annals of Statistics, vol. 3, no. 6, 1189-1242.

ESCARRE, A. (1972). *Essai d'application des methodes de la Taxonomie Numerique a l'etude d'une chenaie dans la Vallee de Burunda (Navarre)*. Inv. Pesq., 36(1), 7-14.

FERENTINOS, K. y PAPAIOANNAU, T. (1981). *New parametric measures of information*. Information and control, 51, 193-208.

FITCH, W. M. y NEEL, J. V. (1969). *The philogenetic relationships of some Indian tribus of Central and South America*. Amer. J. Human genetics, 21, 384-397.

FRECHET, M. (1957). *Sur la distance des deux lois de probabilite*. Publ. inst. stat. Univ. Paris, 6, 185-198.

FROMMEL, C. y HOLZHUTTER, H. G. (1985). *An estimate of the effect of point mutation and natural selection on the rate of aminoacids replacement in proteins*. J. of molecular evolution, 21, 233-257.

GOODMAN, M. M. (1972). *Distance analysis in biology*. Sys. zool., 21(2), 174-186.

GOWER, J. C. (1966). *Some distance properties of latent root and vector methods in multivariate analysis*. Biometrika, 53, 315-328.

GOWER, J. C. (1967). *A comparison of some methods of cluster analysis*. Biometrics, 23, 623-637.

GOWER, J. C. (1971*a*). *Statistical methods of comparing diferent multivariate analysis of the same data*. En: Mathematics in the archaeological and historical sciences (Hodson, F. R., Kendall, D. B. y Tautu, P. Eds.) Edinburgh University Press, 138-149.

GOWER, J. C. (1971*b*). *A general coefficient of similarity and some of its properties*. Biometrics 27, 857-871.

GOWER, J. C. (1982). *Euclidean distance geometry*. Math. scientist, 7, 1-14.

GOWER, J. C. y BANFIELD, C. F. (1975). *Goodness-of-fit criteria for hierarchical classification and their empirical distributions its relation with the external variables*. En: Proceedings of the 8th inter. biometric conference, 347-361.

GOWER, J. C. y LEGENDRE, P. (1986). *Metric and Euclidean Properties of Dissimilarity coefficients*. J. of Classification, 3, 5-48.

HAVRDA, J. y CHARVAT, F. (1967). *Quantification method in classification processes: Concept of structural alpha-entropy*. Kybernetika, 3, 30-35.

HOLMAN, E. W. (1972). *The relation between hierarchical and euclidean models for psychological distances*. Psychometrika, 37, 417-423.

JACCARD, P. (1900). *Contribution au probleme de l'immigration post-glaciaire de la flore alpine*. Bull. Soc. Vaudoise sci. nat. 36: 87-130.

JARDINE, C. J., JARDINE, N. y SIBSON, R. (1967). *The structure and construction of taxonomic hierarchies*. Math. Biosci. 1, 173-179.

JARDINE, N. y SIBSON, R. (1971). *Mathematical Taxonomy*. J. Wiley, N. York.

JOHNSON, S. C. (1967). *Hierarchical clustering schemes*. Psychometrika, 32, 241-254.

KRUSKAL, J. B. (1964*a*). *Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis*. Psychometrika, 29, 1-27.

KRUSKAL, J. B. (1964*b*). *Nonmetric multidimensional scaling: a numerical method.* Psychometrika, 29, 115-129.

KULCZYNSKI, S. (1928). *Die Pflanzenassoziationen der Pieninen.* Bull. Int. Acad. Polonaise sci. et lettres. Classe sci. math. et nat., serie B, suppl. II (1927): 57-203.

KURCYNSKI, T. W. (1970). *Generalized distance and discrete variables.* Biometrics, 26, 525-534.

LANCE, G. N. y WILLIAMS, W. T. (1967). *Mixed-data classificatory programs. I. Agglomerative systems.* Aust. computer J.1: 15-20.

LE CAM, L. (1975). *Distance between experiments.* En: A survey of statistical design and linear models (J. N. Srivastava, Ed.) North-Holland, pub. co., Amsterdam, 383-396.

LEGENDRE, L. y LEGENDRE, P. (1979). *Ecologie Numerique.* Masson, Paris.

LEGENDRE, P., DALLOT, S. y LEGENDRE, L. (1985). *Succession of species within a community: chronological clustering with applications to marine and freshwath zooplankton.* American naturalist, 125, 257-258.

LIESE, F. y VAJDA, I. (1987). *Convex statistical distances.* BSB, Teubner, Leipzig.

LINGOES, J. C. (1971). *Some boundary conditions for a monotone analysis of symmetric matrices.* Psychometrika, 36, 195-203.

MAHALANOBIS, P. C. (1936). *On the generalized distance in statistics.* Proc. nat. inst. sci. India, 2(1), 49-55.

MANZANO, M. y COSTERMANS, J. (1976). *Dos métodos para el estudio psicológico del léxico: su aplicación a algunos adjetivos de la lengua española.* Latino americana de psicol., 8(2), 171-191.

MARDIA, K. V. (1977). *Mahalanobis distances and angles.* En: Multivariate Analysis-IV (P. R. Krishnaiah, ed.), 495-511. North-Holland pub. co., Amsterdam.

MARDIA, K. V. (1978). *Some properties of classical multidimensional scaling.* Comm. stat., A7 (13), 1233-1241.

MARDIA, K. V.; KENT, J. T. y BIBBY, J. M. (1979). *Multivariate analysis.* Academic Press, London.

MATUSITA, K. (1955). *Decision rules, based on the distance for problems of fit, two samples, and estimation.* Ann. math. statist. 26, 631-640.

MATUSITA, K. (1964). *Distance and decision rules.* Ann. inst. statist. math. 16, 305-320.

MITCHELL, A. F. S., (1988). *Statistical Manifolds of Univariate Elliptic Distributions.* International statistical review, 56, 1, 1-16.

MORGAN, B. J. T. (1981). *Three applications of methods of cluster-analysis.* Statistician, 30, 205-223.

NEI, M. (1971). *Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identity.* Amer. natur. 105: 385-98.

NEI, M. (1972). *Genetic distance between populations.* Amer. natur. 106: 283-92.

NEI, M. (1987). *Molecular evolutionary genetics.* Columbia University Press, New York.

OHSUMI, N. y NAKAMURA, T. (1981). *Some properties of monotone hierarchical dendrogram in numerical classification.* Proc. inst. statist. mathem. (Tokio), 28(1), 117-133.

OLLER, J. M. (1983). *Utilización de métricas Riemannianas en Análisis de Datos Multidimensionales y su aplicación a la Biología.* Pub. de Bioestadística y Biomatemática, 11, Universidad de Barcelona.

OLLER, J. M. (1987). *Information metric for extreme value and logistic probability distributions.* Sankhya, 49 A, 17-23.

OLLER, J. M. y CUADRAS, C. M. (1982*a*). *Defined Distances for some probability distributions.* En: Proceed. II world conf. math. serv. man (A. Ballester, D. Cardus, E. Trillas, eds.) Un. pol. de las Palmas, 563-565.

OLLER, J. M. y CUADRAS, C. M. (1982*b*). *Representación canónica en MANOVA: Aplicación a una clase de Diseño anidado*. Questiio, 6(3), 221-229.

OLLER, J. M. y CUADRAS, C. M. (1983). *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, secc. III, 32-36.

OLLER, J. M. y CUADRAS, C. M. (1985). *Rao's distance for negative multinomial distributions*. Sankhya, 47 A, 75-83.

OLLER, J. M. y CUADRAS, C. M. (1987). *Sobre ciertas condiciones que deben verificar las distancias en espacios probabilísticos*. Actas XV reunión seio, Vol. 2, 503-509, Universidad de Oviedo.

ORLOCI, L. (1967). *An agglomerative method for classification of plant communities*. J. Ecol., 55, 193-205.

ORTON, C. (1988). *Matemáticas para arqueólogos*. Alianza Editorial, Madrid.

PEARSON, K. (1926). *On the coefficient of racial likeness*. Biometrika 18, 337-343.

PEREZ, R., GIL, M. A. y GIL, P. (1986). *Estimating the uncertainty associated with a variable in a finite population*. Kybernetes, 15, 251-256.

PERIS, A., ROMEU, J. y CUADRAS, C. M. (1975). *Valoración de la actividad del Ludiomil a través de un ensayo multicéntrico*. Arch. Neurobio. 38(5), 471-484.

PETITPIERRE, E. y CUADRAS, C. M. (1977). *The canonical analysis applied to the taxonomy and evolution of the genus Timarcha Latr. (Col. Crysomelidae)*. Mediterránea, 1, 13-28.

PREVOSTI, A. (1974). *La distancia genética entre poblaciones*. Miscellanea Alcobe, Univ. Barcelona, 109-118.

PREVOSTI, A., OCAÑA, J. y ALONSO, G. (1975). *Distances between populations of Drosophila Susbobscura based on chromosome arrangement frequencies*. Theor. appl. genetics, 45, 231-241.

PRUZANSKY, S., TVERSKY, A. y CARROLL, J. D. (1982). *Spatial versus tree representations of proximity data*. Psychometrika, 47(1), 3-24.

RAO, C. R. (1945). *Information and the accuracy attainable in the estimation of statistical parameters*. Bull, Calcutta math. soc., 37, 81-91.

RAO, C. R. (1948). *On the distance between two populations*. Sankhya, 9, 246-248.

RAO, C. R. (1954). *On the use and interpretation of distance function in statistics*. Bull. int. stat. inst., 34, 90-97.

RAO, C. R. (1982). *Diversity and dissimilarity coefficients: A unified approach*. Theor. Pop. Biology, 21(1), 24-43.

RAPOPORT, A. y FILLENBAUM, S. (1972). *An Experimental Study of Semantic Structures*. En: Multidimensional Scaling. II. (A. K. Romney, R. N. Shepard y S. B. Nerlove, eds.), 96-131. Seminar Press, New York.

REYMENT, R. A. (1973). *The discriminant function in systematic biology*. En discriminant analysis, 311-337 (T. CACOULLOS, ed.). Academic Press, New York.

RIOS, M. y CUADRAS, C. M. (1986). *Distancia entre Modelos Lineales Normales*. Questiio, 10(2), 83-92.

RIOS, M. y OLLER, J. M. (1988). *Rao distance between multivariate linear normal models and application to the classification of response curves*. Sometido a : Biometrics.

ROMNEY, A. K.; SHEPARD, R. N. y NERLOVE, S. B. (Eds.) (1972). *Multidimensional scaling. Theory and applications in the behavioral sciences. Vol. II. Applications*. Seminar Press, New York.

RUIZ-RIVAS, C. y CUADRAS, C. M. (1988). *Inference properties of a one-parameter bivariate family of distributions with given marginals*. J. of Multivariate Analysis, 27 (2), 447-456.

RUIZ-RIVAS, C., USON, M. T., CUADRAS, C. M. (1979). *Alguns aspectes i aplications de la construccio de distribucions bivariants*. Questiio, 3(3), 121-127.

RUSSELL, P. F. y RAO, T. R. (1940). *On habitat and association of species of anopheline larvae in south-eastern Madras*. J. Malar. Inst. India 3: 153-178.

SALICRU, M. (1987). *Medidas de divergencia en análisis de datos*. Tesis doctoral, Universidad de Barcelona.

SALICRU, M. y CUADRAS, C. M. (1987). *Funciones de entropia asociadas a medidas de Csiszar*. Questiio, 11(3), 3-12.

SANCHEZ, P. (1986). *Construcción de distribuciones con marginales multivariantes dadas*. Questiio, Vol. 10, N.º 3, 113-141

SATTATH, S y TVERSKY, A. (1977). *Additive similarity trees*. Psychometrika, 42(3), 319-345.

SCHOENBERG, I. J. (1935). *Remarks to Maurice Frechet's article'sur la definition axiomatique d'une classe d'espaces vectorieelles distancies applicables vectoriellement sur l'espace de Hilbert'*. Ann. Math. 36, 724-732.

SEAL, H. L. (1964). *Multivariate statistical analysis for biologists*. Methuend and Co. Ltd., London.

SEBER, G. A. F. (1984). *Multivariate Observations*. J. Wiley, New York.

SHEPARD, R. N. (1962*a*). *The analysis of proximities: multidimensional scaling with an unknown distance function. I.* Psychometrika, 27, 125-140.

SHEPARD, R. N. (1962*b*). *The analysis of proximities: multidimensional scaling with an unknown distance function. II.* Psychometrika, 27, 219-246.

SKOVGAARD, L. T. (1984). *A Riemannian Geometry of the Multivariate Normal Model*. Scand. J. statist., 11, 211-223.

SNEATH, P. H. A. y SOKAL, R. S. (1973). *Numerical taxonomy*. W. H. Freeman and Co., San Francisco.

SOKAL, R. R. y MICHENER C. D. (1958). *A statistical method for evaluating systematic relationships*. Univ. Kansas sci. bull. 38. 1.409-1.438.

SORENSEN, T. (1948). *A method of stablishing groups of equal amplitude in plant sociology based on similarity of species content and its aplication to analyses of the vegetation on Danish commons*. Biological SKR., 5, 1-34.

STEINBERG, A. G., BLEIBTREU, H. K., KURCYNSKI, T. W. y MARTIN, A. O. (1966). *Genetic studies on an inbred human isolated*. En: Third int. Congress Human Genetics, Chicago (J. F. Crow. Ed.), 267-289.

TAKEUCHI, K., YANAI, H. y MUKHERJEE, B. N. (1982). *The Foundations of Multivariate Analysis*. Wiley El., New Delhi.

THORPE, J. P. (1979). *Enzyme variation in taxonomy: The estimation of sampling errors in measurement of interspecif genetic similarity*. Biol. J. Linn. Soc., 11, 369-386.

THURSTONE, L. L. (1927). *A law of comparative judgment*. Psychological Review, 34, 273-286.

TORGERSON, W. S. (1958). *Theory and methods of scaling*. J. Wiley, New York.

VAJDA, I. (1972). *On the f-divergence and singularity of probability measures*. Period. math. hungar. 2, 223-234.

VALLEJO, J., PORTA, A. y CUADRAS, C. M. (1975). *Incidencias de los psicofármacos en la evolución reumatológica del enfermo poliartrítico crónico*. Medicina clínica, 64(9), 452-457.

WATERMAN, M. S., SMITHY, T. F., SINGH, M., y BEYER, W. A. (1977). *Additive evolucionary trees*. J. Theor. Biol., 64, 199-213.

WISH, M y CARROLL, J. P. (1982). *Multidimensional Scaling and its applications*. En: Handkbook of statistics, Vol.2 (P. R. Krisnaiah, L. N., Kanal, Eds.), 317-345. North-Pub. Co., Amsterdam.

COMENTARIOS

BELEN CASTRO IÑIGO
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco (Bilbao)

El concepto de distancia juega un papel muy importante dentro de las técnicas de análisis de datos, puesto que en las mismas, se trata de representar individuos respecto a variables, en base a distancias o similaridades definidas sobre las matrices de datos.

Por ejemplo, el "análisis de componentes principales" y el "análisis canónico" utilizan variables cuantitativas, haciéndose uso en el primer caso, de una distancia euclídea y en el segundo de la distancia de Mahalanobis. Poco queda que añadir sobre estas distancias al artículo del profesor Cuadras.

El "análisis de coordenadas principales" se realiza sobre variables cualitativas y con distancias relacionadas con similaridades. Una técnica más general que ésta y que a su vez trabaja también con índices de similaridad, es el "análisis de proximidades" (MDS). Si bien en el artículo se enumeran índices de similaridad sobre variables cuantitativas y cualitativas, se puede mencionar la existencia de algunos aplicables a tablas mixtas, con variables tanto binarias como cualitativas y cuantitativas. Entre ellos destaca el de Gower (1971):

$$s_{ij} = \sum_{k=1}^{p} s_{ijk} \,/\, \sum_{k=1}^{p} w_{ijk}$$

w_{ijk} es 1 ó 0 dependiendo de si la comparación considerada es válida para k y, excepto en el caso de dicotómicas suele ser 0 cuando se desconozca el valor de k en uno o ambos individuos. En dicotómicas w_{ijk} es 0 cuando k está ausente en ambos individuos.

s_{ijk} para binarias coincide con el de Jaccard; en el caso de nominales es 1 si los individuos son iguales en k y 0 en caso contrario; para datos intervalo $s_{ijk} = 1 - |x_{ik} - x_{jk}| / R_k$ con x_{ik} el valor del individuo i en k y R_k el recorrido de k.

Los índices de similaridad adquieren también protagonismo en el campo de los procesos estocásticos. Algunos (como el de Marley (1981)) se han obtenido a partir de las medidas discutidas por Shepard y Arabie (1979) y Tversky (1977). Así, por ejemplo, si $X_J(t)$, para $J \subset T = \{x_1, x_2, \ldots, x_n\}$, siendo T el conjunto de estímulos usados en el experimento, denota el número de veces que el evento J ocurre en el tiempo t, se puede suponer que la similaridad entre x_i y x_j está basada en:

$$s_{ij}(t) = \frac{1}{t}\left\{\theta \sum_{i,j \in J \subset T} X(t) - \alpha \sum_{\substack{i \in J \subset T \\ j \notin J \subset T}} X(t) - \beta \sum_{\substack{i \notin J \subset T \\ j \in J \subset T}} X_J(t)\right\}$$

donde θ, α, β son constantes.

Con relación a los índices de similaridad es preciso mencionar que en muchas aplicaciones aparece un claro elemento estocástico en los atributos para los individuos comparados. Esto ocurre cuando la similaridad de los organismos individuales se determina en base a caracteres morfológicos y psicológicos. Situaciones de este tipo ocurren, no sólo en biología, sino también en arqueología y crítica literaria. Los índices más utilizados como el de Jaccard, Sokal y Michener resuelven este problema muy pobremente. Por otro lado, distancias como Sokal (1961), Sokal y Sneath (1963) podrían tratarse desde el punto de vista probabilístico si la distribución de los atributos individuales es conocida. Ghent (1963) utilizó el coeficiente τ de Kendall como índice de similaridad entre comunidades de plantas y animales, y aunque podría llevarse a un tratamiento probabilístico, su aplicabilidad resultaría escasa donde los atributos son inconmensurables. Goodall (1966) desarrolló un índice basado en probabilidad, que puede aplicarse a atributos cualitativos, ordenados, métricos, pudiéndose efectuar combinaciones de similaridades respecto a los diferentes atributos. Este índice se comparó con el de Sorensen (1948) sobre datos de ocho comunidades de tierra de pasto de Colorado.

Otra técnica a comentar dentro del análisis de datos es el "análisis factorial de correspondencias" (AFC) que es apropiado para representar tablas de frecuencias a las que se les aplica una distancia euclídea, la *ji*-cuadrado. Dicha distancia, que a pesar de sus propiedades dentro de esta técnica ha sido omitida en el artículo, se define para dos distribuciones L_1, L_2 sobre el espacio de las distribuciones de probabilidad P_J con centro f_J como:

$$\| L_J^1 - L_J^2 \|_{f_J^2} = \sum_{j \in J} (L_j^1 - L_j^2)^2 / f_j$$

y adquiere su máxima aplicabilidad en AFC sobre dos poblaciones H_i y $H_{i'}$ en relación a los caracteres $A_1, \ldots, A_n$, teniendo la forma:

$$d^2(i, i') = \sum_{j \in J} (f_j^i - f^{i'}_j)^2 / f_j$$

con f_j^i = prob $\{ A_j / H_i \}$ y f_j = frecuencia de A_j

Autores como Piris (1986) han contribuido al estudio de la naturaleza de la distancia *ji*-cuadrado.

Dentro del campo de la Estadística Matemática el concepto de distancia adquiere una importancia fundamental. El profesor Cuadras resume de una manera excelente los aspectos en los que dicho instrumento se presenta de una manera más notable en este campo. Quisiera recalcar que la mayor parte de la literatura estadística reciente maneja las distancias para la búsqueda de estimadores robustos. Huber (1964) propuso una clase de M estimadores como soluciones a un problema minimax de este tipo.

Ante el mal comportamiento de estos procedimientos en situaciones en las que propiedades como invarianza y simetría no se presentaban, fue necesario efectuar una extensión de los mismos.

Wolfowitz (1957) publicó un paper introduciendo el método de mínima distancia (MD), dando resultados consistentes y proporcionando ejemplos de su uso. Knüsel (1969) examinó consideraciones sobre la robustez de este método. Littell y Rao (1975) y Rao, Schuster, y Littell (1975) consideraron con más detalle el uso de la distancia de Kolmogorov para la estimación MD. Holm (1976) sugirió la estimación MD como el método más natural para algunos problemas de robustez.

Parr y Shucany (1980) estudiaron el comportamiento de los estimadores (MD) con respecto a distancias como:

— La distancia ponderada de Kolmogorov en R

$$D_\psi(K,L) = \sup_{x \in R} | K(x) - L(x) | \psi (L(x))$$

— La distancia ponderada de Cramer-von Mises

$$W_\psi^2 (K,L) = \int_{-\infty}^{\infty} (K(x) - L(x))^2 \psi(L(x)) \, dL(x)$$

— La distancia del intervalo maximal de probabilidad de Kniper.

$$V(K,L) = \sup_{-\infty < a < b < \infty} | (K(b) - K(a)) - (L(b) - L(a)) |$$

— $Z^2_{a,b}(K,L) = a \int_{-\infty}^{\infty} (K(x) - L(x))^2 \, dL(x) + b \left(\int_{-\infty}^{\infty} (K(x)) - L(x)) \, dL(x) \right)^2$

— una clase de discrepancias incluyendo los casos de Cramer-von Mises $W^2(K,L)$, de Watson $U^2(K,L)$ y Chapman para K y L funciones de distribución definidas en un subconjunto (común) de R.

Autores que en la actualidad trabajan sobre este tema serían Donoho y Liu (1988).

Dentro de estos estimadores de mínima distancia destaca el basado en la distancia de Hellinger (MHD). Este estimador ha sido conocido por ser eficiente de primer orden en un modelo paramétrico multinomial (Rao (1963)). Más recientemente Beran (1977) argumentó que un estimador MHD para un modelo paramétrico contínuo debería ser robusto y mostró que es asintóticamente eficiente para un modelo con soporte compacto. Stather (1981), Tamura y Boos (1986) dan extensiones a modelos con soporte compacto. Simpson (1987) enfoca la estimación MHD para datos discretos, donde el modelo permite tener soportes infinitamente numerables. Este estimador ha sido aplicado a estudios de mutagenecidad química.

Para acabar deseo felicitar al Profesor Cuadras por este magnífico artículo, que debe servir para reflexionar sobre este instrumento tan importante y utilizado dentro de la Estadística y en muchas ocasiones quizá poco conocido.

BIBLIOGRAFIA

DONOHO, D. L. y LIU, R. C. (1988) *Pathologies of some minimum distance functionals*. The Annals of Statistic, Vol. 16, N.º 2

DONOHO, D. L. y LIU, R. C. (1988) *The "automatic" robustness of minimum distance estimators*. The Annals of Statistic, Vol. 16, N.º 2.

GOODALL, D. W. (1966) *A new similarity index based on probability*. Biometrics, Vol. 22.

JAMBU, M. y LEBAUX, M. O. (1983) *Cluster Analysis and Data Analisis*. North-Holland.

MARLEY, A. A. J. (1981) *Multivarite stochastic processes compatible with "aspect" models of similarity and choise*. PSYCHOMETRIKA, Vol. 46, N.º 4.

PIRIS, J. M. (1986) *Distancia Khi-cuadrado y diversidad sociopolítica.* Tesis doctoral.

TAMURA, R. N. y BOOS, P. (1986) *Minimum Hellinger Distance Estimation for Multivariate Location and Covariance.* Journal American Statistical Association, Vol. 81.

WOLFOWITZ, J. (1957) *The Minimum Distance Method.* Annals of Mathematical Statistics, Vol. 28.

M.ª PILAR MARTIN-GUZMAN
Departamento de Economía Aplicada
Universidad Autónoma de Madrid

Este artículo tiene un gran interés como presentación coherente y muy actualizada del estado de la cuestión en el estudio de las distancias estadísticas.

Es este un tema de incidencia creciente en la estadística aplicada. La descripción de aplicaciones posibles que, sin pretensiones de exhaustividad, incluye el autor en el apartado 7 nos da una idea de su gran utilidad. Añadiré que las nociones de distancia y similaridad son también instrumentos básicos en el análisis económico, y en especial en un sector del mismo actualmente en auge: la economía regional.

La presentación que realiza el autor de los distintos tipos de distancias que pueden definirse imponiendo alternativamente propiedades diversas me parece especialmente interesante, por su relación con algunos resultados básicos del análisis multivariante, sobre todo en las cuestiones de agrupación y seriación. Así, el conglomerado jerárquico formado con elementos cuyas distancias sean ultramétricas es único, en tanto que una distancia euclidea no ultramétrica puede dar lugar a múltiples opciones de formación jerárquica de conglomerados, con resultados en general distintos.

Finalmente, quisiera hacer observar que la definición de similaridad que el autor adopta no es la única existente. En la literatura estadística es frecuente que se acepte como similaridad una aplicación positiva y simétrica que satisfaga además alguna condición de monotonicidad. Pero el autor de este artículo impone un axioma especialmente restrictivo: que $S_{ii} = 1$. De acuerdo con su definición, el coeficiente de Russell y Rao, definido como

$$\frac{a}{p}$$

es decir, número de caracteres presentes comunes dividido por el número de características, puede no ser una similaridad.

La eliminación del axioma $S_{ii} = 1$ permite incluir en el concepto de similaridad a los coeficientes habitualmente utilizados como tal, pero entonces las transformaciones

$$\delta_{ij} = 1 - S_{ij}$$

y

$$\delta_{ij} = \sqrt{1 - S_{ij}}$$

pueden no conducir finalmente a una distancia, ya que podría ser $\delta_{ii} > 0$ para algún *i*. La transformación de Gower, en cambio

$$\delta_{ij} = \sqrt{S_{ii} + S_{jj} - 2S_{ij}}$$

no presenta tal problema, y esta es otra de sus grandes ventajas.

JOSE M. GARCIA-SANTESMASES
(Universidad Complutense, MADRID)

Tema difícil de encontrarlo recogido, aunque sea parcialmente, queda ilustrado no sólo por la bibliografía que aporta, sino también por su versatilidad y extensión.

El enfoque dado al tema es doblemente atractivo. Por un lado presenta los últimos resultados obtenidos por el autor y su equipo que juntamente con otros ofrecen un útil estado del arte sobre este tema.

Por otro ofrece un amplio conjunto de aplicaciones algunas de ellas bien comentadas, otras simplemente referenciadas que lo hacen particularmente útil desde el punto de vista práctico.

Aunque es difícil la elección, creo que el tema del MDS merece un apartado separado y no a través de una aplicación.

Especialmente interesante tanto por su lectura como por los resultados que presenta, por su fondo y su forma, es el capítulo dedicado a las ultramétricas. Echo en falta algoritmos para la obtención de éstas a partir de desemejanzas como pueden ser los métodos subdominantes. Jardine y Sibson (1971), Sánchez (1977).

No parece natural, sin embargo, en un artículo de este tipo, la ausencia de una referencia explícita a la distancia de la Chi-dos o de Benzecrí, Benzecrí

(1976), Cuadras (1981), Lebart (1977) por ser ésta especialmente utilizada en aplicaciones sociológicas que por otra parte también están ausentes en el capítulo de aplicaciones.

En mi opinión, en un artículo de estas características es necesario un comentario sobre los aspectos computacionales del tema, como son referencias a paquetes de programas que implementen alguna de las distancias que se presentan, SPAD, BMDP, Clustan, etc., y las dificultades de implementación, temas éstos especialmente útiles desde un punto de vista práctico, Jambu (1977), Wishart (1978).

En cualquier caso mi felicitación al autor del artículo y al director de la revista por proporcionarnos una aproximación a un tema de tan dificil localización.

BIBLIOGRAFIA

BENZECRÍ, J.P. (1976). «L'Analyse Des Données I, La taxonomie. L'Analyse des Données II, L'Analyse des Correspondances». *Dunod,* París.

BMDP, Statistical Software, (1985). *Dixon Ed.* University of California Press.

CUADRAS, C.M. (1981). «Métodos de Análisis Multivariante». *Eunibar,* Barcelona.

LEBART, L., MORINEAU, A. y TABARD, N. (1977). «Techniques de la Description Stastistique». *Dunod,* París.

JAMBU, M., LEBEAUX, M.O. (1978). «Classification automatique pour L'Analyse des Données», Tomo II. *Logiciels,* Dunod.

JARDINE, N. y SIBSON, R. (1971). «Mathematical Taxonomy». *J. Wiley,* New York.

SÁNCHEZ, M. (1977). «Tratamiento estadístico de Datos». *Centro de cálculo de la Universidad Complutense,* Madrid.

SPAD.N CISIA. 25 Avenue de L'Europe 92310, Sèvres, (France).

WISHART, D. (1978). «Clustan User manual». *Report n.º 47, Program library unit.* Edinburgh University.

DANIEL PEÑA

Escuela Técnica Superior de Ingenieros Industriales
Universidad Politécnica de Madrid

Cuando la Revista solicitó a Carlos Cuadras un artículo de revisión del concepto y aplicaciones de la noción de distancia en Estadística, estábamos seguros de poder contar con un trabajo de calidad. La respuesta ha

sobrepasado nuestras expectativas, y quiero felicitar al autor por su profunda y rigurosa presentación de este tema, al que el Profesor Cuadras ha hecho contribuciones relevantes. Este comentario pretende únicamente resaltar que: 1) como señala el autor en su trabajo, cualquier problema de inferencia estadística puede replantearse como un problema de distancias y 2) en mi opinión, las distancias de Kullback-Leibler –para distancias entre distribuciones– y la de Mahalanobis –para vectores de datos– ocupan un lugar especialmente destacado en Estadística.

Comenzando con la distancia de Kulback-Leibler (KL), es bien conocido que la estimación máximo-verosímil resulta al minimizar esta distancia entre la verdadera distribución y la estimada. En efecto, sea $f(x; \Theta)$ un modelo paramétrico que satisface las condiciones habituales de regularidad y sea Θ_0 el verdadero valor del vector desconocido de parámetros. La distancia KL entre un modelo estimado $f(x; \Theta)$ y el modelo verdadero $f(x; \Theta_0)$ es

$$\int ln\ f(x; \Theta_0)\ f(x; \Theta_0)dx - \int ln\ f(x; \Theta)\ f(x; \Theta_0)dx$$

como el primer término es constante, minimizar la distancia equivale a maximizar el valor promedio de $ln\ f(x; \Theta)$, cuyo estimador con los datos muestrales es

$$\frac{1}{n} \sum ln\ f(x; \Theta)$$

y obtenemos el método de máxima verosimilitud.

Esta distancia aparece de una manera natural tanto en el enfoque clásico (Kullback, 1978) como en el Bayesiano (Bernardo 1979a, 1979b) no sólo ligada a problemas de estimación, sino también a la construcción de herramientas diagnósticas (Geisser y Johnson 1983, Guttman y Peña 1988), la selección de modelos (Akaike 1974, Peña y Arnaiz 1981) o el desarrollo de métodos robustos de estimación (Peña y Guttman 1989).

Es fácil comprobar que la distancia KL entre dos poblaciones normales multivariantes $N(\mu_1, \Sigma)$, $N(\mu_2, \Sigma)$ con la misma matriz de covarianzas Σ es:

$$\frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$$

que es la distancia de Mahalanobis entre las medias de ambas distribuciones. Por tanto, en la hipótesis de Normalidad, la distancia KL se reduce de forma natural a la distancia de Mahalanobis. Todos los problemas de inferencia en poblaciones normales y, por extensión, los problemas de inferencia asintótica con familias regulares, están basados en esta distancia. En efecto, la otra distancia básica para datos cuantitativos, la X^2 de Pearson, que en su versión más conocida es:

$$X^2 = (\mathbf{O} - \mathbf{T})' \mathbf{M}^{-1} (\mathbf{O} - \mathbf{T}) = \sum \frac{(O_i - T_i)^2}{e_i} \qquad (1)$$

donde **O** es un vector de k frecuencias observadas, **T** un vector de k frecuencias teóricas tal que $E[\mathbf{O}] = \mathbf{T}$ y M una matriz diagonal cuyos términos son los componentes del vector T, puede escribirse (véase Kendall y Stuart 1979, p. 381) como una distancia de Mahalanobis entre un vector de dimensión k–1, **X** obtenido eliminando uno cualquiera de los componentes de O, y su vector de medias μ (obtenido análogamente de T):

$$X^2 = (\mathbf{X} - \mu) \sum^{-1} (\mathbf{X} - \mu) \qquad (2)$$

donde $\sum$ es la matriz de varianzas y covarianzas del vector **X**.

Los contrastes de medias en poblaciones normales pueden clasificarse en dos grupos: en el primero conocemos el parámetro de escala σ^2 en la matriz de varianzas y covarianzas. El contraste es entonces de la forma:

$$(\hat{\Theta} - \Theta)\ \mathrm{M}^{-1}\ (\hat{\Theta} - \Theta)\ \sigma^2 \qquad (3)$$

donde $\hat{\Theta}$ es un vector de medias estimadas tal que $E(\hat{\Theta}) = \Theta$ y $Var(\hat{\Theta}) = \sigma^2 \mathbf{M}$. Por ejemplo, si Θ es escalar, la expresión (3) se reduce a $n(X - \mu)^2/\sigma^2$, que es el contraste clásico de la media. Si Θ es un vector de medias y la matriz de varianzas y covarianzas de los datos $\sum$ es conocida, (3) se reduce a:

$$(X - \mu)' \sum^{-1} (\mathbf{X} - \mu)$$

que será una X^2 con grados de libertad iguales a la dimensión de μ.

Como los estimadores de poblaciones normales son lineales, llamando **X** al vector de datos, en general podemos escribir:

$$\hat{\Theta} = \mathbf{A}\ \mathbf{X} \qquad E(\hat{\Theta}) = \mathbf{A}\ \mu$$

y la distancia (3) puede tambien escribirse:

$$X^2 = (\mathbf{X} - \mu)'\mathbf{A}'\mathbf{M}^{-1}\mathbf{A}\ (\mathbf{X} - \mu)\ \sigma^2 \qquad (4)$$

que equivale a una distancia de Mahalanobis entre los datos y sus medias. Por ejemplo, en el caso más simple de una población normal con media desconocida pero varianza conocida, el contraste clásico de la media resulta ahora:

$$(\mathbf{X} - \mu)'\mathbf{A}'\mathbf{A}\ (\mathbf{X} - \mu)\ \sigma^2$$

donde la matriz cuadrada **A'A** tiene rango uno y todos los componentes igual a n^{-2}. Es bien conocido (Peña 1987, pp. 234-235) que cualquier contraste asintótico (razón de verosimilitudes, Lagrange o Wald) equivale a una distancia de Mahalanobis entre el vector de parámetros que se contrasta y su estimación, diferenciándose los contrastes únicamente en las aproximaciones utilizadas para escribir la matriz de varianzas y covarianzas de los estimadores.

Cuando σ^2 es desconocida, el contraste compara la distancia de Mahalanobis entre el estimador y el parámetro con otra distancia calculada a partir de los datos para estimar el efecto de escala, obteniendo el contraste general de la F:

$$F = \frac{(\hat{\Theta} - \Theta)'\ \mathbf{M}^{-1}\ (\hat{\Theta} - \Theta)\ \dfrac{1}{k}}{(\mathbf{X} - \hat{\mu})\ \sum^{-1}\ (\mathbf{X} - \hat{\mu})\ \dfrac{1}{n-k}} \qquad (5)$$

donde Θ tiene dimensión k y matriz de varianzas/covarianzas M σ^2, $\hat{\mu}$ es la estimación de las medias del vector de datos **X** que tiene matriz de varianzas covarianzas $\sigma^2\ \sum$. Estos son los contrastes básicos de los modelos lineales donde se supone independencia y, en consecuencia, $\sum$=I. La idea es, sin embargo, completamente general y puede aplicarse para cualquier vector de variables normales con matriz de varianzas covarianzas $\sum\ \sigma^2$. Por ejemplo, para muestras de una población multivariante se obtiene la distribución T^2 de Hotelling.

En los contrastes de varianza para una población normal de nuevo nos encontramos la distancia de Mahalanobis.

$$X^2 = \frac{(X - \hat{\mu})'\Sigma^{-1}\ (X - \hat{\mu})}{\sigma^2}$$

por ejemplo, para una población normal univariante $\hat{\mu} = \bar{X}$, $\Sigma = I$ y el contraste se reduce al bien conocido resultado $\Sigma(X_i - \bar{X})^2/\sigma^2$. La comparación de varianzas de dos poblaciones es de nuevo una comparación entre distancias de Mahalanobis.

En resumen, cualquier problema de test de hipótesis en Estadística puede interpretarse en términos de distancias: Los contrastes que resultan en una distribución X^2 utilizan una cierta distancia de Mahalanobis entre dos vectores, y los que se basan en la distribución F (o T^2 de Hotelling o, por supuesto *t* de Student, como caso particular) comparan las distancias de Mahalanobis entre dos vectores cuyas dimensiones respectivas son los grados de libertad de la F.

Análogamente, cualquier problema de estimación paramétrica resulta al minimizar una cierta distancia. Por ejemplo, los métodos de estimación robusta minimizan, en lugar de la distancia euclidea $(Y - X\beta)'Y - X\beta)$, donde Y es un vector de datos, X una matriz de variables y β un vector de parámetros, una distancia de Mahalanobis del tipo

$$(Y - X\beta)' \Sigma^{-1}(\beta) (Y - X\beta)$$

donde Σ depende de los parámetros desconocidos y, por tanto, la minimación de la función requiere un procedimiento iterativo (mínimos cuadrados generalizados iterativos).

REFERENCIAS

AKAIKE, I. (1974). "A new Look at the Statistical Model Identification". *IEEE transactions on Automatic Control,* 19,6, 716-722.

BERNARDO, J.M. (1979a). "Excepted utility as expected information". *Annals of Statistics,* 7, 686-690.

BERNARDO, J.M. (1979b). "Reference posterior distributions for Bayesian Inference". *Journal of Royal Statistical Society B,* 41, 113-147 (con discusión).

GUTTMAN, I. and PEÑA, D. (1988). "Outliers and influence: Evaluation by posteriors of Parameters in the Linear model". *Bayesian Statistics* 3, Bernardo, J.M. et al (editors). Oxford University Press.

JOHNSON, W. y GEISSER, S. (1983). "a predictive view of the detection and Characterization of Influential Observations in Regression Analysis". *Journal of American Statistical Association,* 78, 381, 137-144.

KENDALL, M y STUART, A. (1977). *The Advanced Theory of Statistics* (Vol. 1). Charles Griffin.

KULBACK, S. (1978). *Information Theory and Statistics*. Dover.

PEÑA, D. (1987). *Estadística, Modelos y Métodos* (Vol. 1). Alianza Universidad Textos.

PEÑA, D. y ARNAIZ, G. (1981). "Criterios de selección de modelos ARIMA". *Trabajos de Estadística y de I.O.* 32, 1, 70-93.

PEÑA, D. and GUTTMAN, I. (1989). "Optimal Collapsing of mixture distributions in robust recursive estimation". *Communication in Statistics, Theory and Methods,* (in press).

J. M. PRADA SANCHEZ
Dpto. de Estadística e Investigación Operativa
Universidad de Santiago

El trabajo de Carles M. Cuadras, que contiene resumida parte de su labor investigadora durante los últimos años, constituye, a mi juicio, una aportación fundamental para el estadístico actual. Presenta una exposición didáctica, rigurosa y exhaustiva, sobre la noción de distancia estadística, e incluye muchas e interesantes aplicaciones de la misma en diversos campos, así como una completa bibliografía al respecto.

Algunas otras aplicaciones de interés de las distancias estadísticas en los contextos de estimación, contrastes de hipótesis y otros son las siguientes:

- La estimación del parámetro θ de una población absolutamente contínua f_θ, al minimizar sobre el espacio paramétrico la distancia de Hellinger entre f_θ y un estimador paramétrico de f_θ. Esta idea, introducida por Beran (1977), fue extendida por Lasala (1981) en su tesis doctoral.
- La búsqueda de la ventana óptima en un problema de estimación de la densidad puede plantearse minimizando distancias estadísticas (Hellinger, Kullback), como puede verse en Hall (1983).
- Cristóbal, J.A., Faraldo, P. y González Manteiga, W. (1987), hacen una estimación parámetrica de la regresión minimizando sobre el espacio paramétrico la distancia L_2 entre el modelo y un estimador no paramétrico del mismo.
- Härdle, W. y Mammen, E. (1988, preprint), realizan contrastes de hipótesis acerca de un modelo de regresión utilizando la distribución Booptstrap de la distancia L_2 entre una estimación clásica del modelo y una estimación no paramétrica del mismo.
- Una posible aplicación interesante de las distancias estadísticas nos llevaría, en la técnica Bootstrap, a sustituir la distribución empírica por aquélla que menos «dista» de ella, entre las pertenecientes a una cierta clase de distribuciones (p. ej.: la clase de distribuciones simétricas).

REFERENCIAS

BERAN, R. (1977). «Minimum Hellinger distance estimates for parametric models». *Annals of Statistics,* Vol. 5, 3, 445-463.

CRISATOBAL, J.A., FARALDO, P. y GONZALEZ MANTEIGA, W. (1987). «A class of linear regresion parameter estimators constructed by nonparemetric estimation». *Annals of Statistics,* Vol. 15, 2, 603-609.

HALL, P. (1983). «Large sample optimality of least squares Cross-validation in density estimation». *Annals of Statistics,* Vol. 11, 4, 1156-1174.

HARDLE, W. y MAMMEN, E. (1988). «Compairing non parametric versus parametric regresion fits». *(preprint).*

LASALA, M.P. (1981). «Cuestiones notables sobre procedimientos robustos: Funcionales de mínima *g*-divergencia y sus estimadores asociados». *Tesis doctoral.* Universidad de Zaragoza.

Contestación

Agradezco los comentarios de B. Castro, M. P. Martín-Guzmán, J. M. García-Santesmases, D. Peña y J. M. Prada, porque no tan solo contienen observaciones interesantes sobre aspectos concretos, sino que además contienen ideas y referencias que enriquecen mi artículo. Mi contestación voy a llevarla a cabo incluyendo las respuestas en cuatro secciones.

1. SIMILARIDADES

Aunque en la sección 2.1 se imponía la restricción $0 \leq s_{ij} \leq 1$, en la misma sección se señala que una similaridad puede también tomar valores superiores a 1, y se destaca la importancia de la distancia.

$$\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}} \tag{1}$$

en lo que hace referencia a las propiedades métrica y euclídea, como queda bien patente en el cuadro 2. Sin embargo, tiene razón M.P. Martín-Guzmán. La restricción $s_{ii} = 1$ puede dejar de cumplirse para el coeficiente de Russell y Rao, definido como *a/p*.

En cuanto al coeficiente de similaridad para datos tanto cuantitativos como categóricos, se hace referencia al coeficiente de Gower (1971) en la sección 7.7. Conviene observar que este coeficiente puede escribirse también como

$$s_{ij} = \frac{\sum_{k=1}^{n_1} (1 - |x_{ik} - x_{jk}| / R_k) + a + \alpha}{n_1 + (n_2 - d) + n_3} \tag{2}$$

donde n_1 es el número de variables continuas, a y d son el número de coincidencias para las n_2 variables dicotómicas y α es el número de estados coincidentes para las n_3 variables multinomiales. Deben aplicarse algunas correcciones en el caso de datos faltantes. R_k es el rango de la *k-esima* variable continua.

Es fácil comprobar que s_{ij} se reduce al coeficiente de Jaccard si todas las variables son dicotómicas y al coeficiente de Sokal y Michener si todas son binarias. Nótese la distinción entre variables dicotómicas (en las que se resalta la importancia de las coincidencias positivas) y variables binarias.

La importancia de s_{ij} reside en que (1) da lugar siempre a una distancia euclídea (salvo, eventualmente, el caso de datos faltantes). Por otra parte, recientemente hemos utilizado esta distancia (con datos reales) para verificar el modelo de regresión propuesto en la sección 7.7, obteniendo buenos resultados con respecto al procedimiento clásico de cuantificar las variables cualitativas.

2. DISTANCIA *ji–cuadrado*

La no inclusión de esta distancia en el texto puede explicarse por la dificultad de definirla sin una clara referencia al Análisis de Correspondencias (AC), lo que hubiera comportado alargar un original ya de por sí bastante extenso. Aunque el AC es un tema que ya he tratado en Cuadras (1981, cap. 14), lo comentaré brevemente relacionándolo con otros métodos.

Sea $F = (f_{ij})$ una tabla de contingencia *rxs*. Podemos suponer que las frecuencias son relativas y suman 1. La distancia *ji–cuadrado* es una medida de la diferencia entre los perfiles de dos filas *i, k,* que se define por

$$\delta^2(i,k) = \sum_{j=1}^{s} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{kj}}{f_{k.}} \right)^2 \qquad (3)$$

Análogamente se define la distancia entre columnas.

El AC puede describirse a través de la descomposición singular

$$D_r^{-1/2}(F - E)D_c^{-1/2} = U \Lambda V' \qquad (4)$$

donde $UU'=I$, $V'V=I$, Λ es diagonal y contiene los valores singulares, $D_r = diag(f_{1.}, \ldots, f_{r.})$, $D_s = diag(f_{.1}, \ldots, f_{.s})$., y $E=D_r 1\,1' D_c$
Obsérvese que los elementos de E son $e_{ij} = f_{i.} f_{.j}$ (ó $f_{i.} f_{.j}/N$ si trabajamos con frecuencias absolutas), luego la descomposición singular (4) mide la discrepancia entre F y la matriz E que corresponde al caso de independencia entre filas y columnas.

La representación de las filas se consigue a través de la matriz de coordenadas euclídeas.

$$R = D_r^{-1/2} U \Lambda$$

y la de las columnas a través de

$$C = D_c^{-1/2} V \Lambda$$

Ambas matrices están relacionadas por

$$R = D_r^{-1} F \; C \; \Lambda^{-1}$$

$$C = D_c^{-1} F' R \; \Lambda^{-1}$$

permitiendo una representación simultánea de filas y columnas a lo largo de unos mismos ejes. Las distancias euclídeas entre filas (columnas) son iguales a las distancias *ji*-cuadrado. Utilizando adecuadamente el análisis de componentes principales, podemos representar y estudiar la dependencia entre filas y columnas de una tabla de contingencia.

El AC está estrechamente relacionado con otros métodos de análisis de datos categóricos.

1) Puede utilizarse como un complemento del modelo log-lineal. Así como este modelo es útil para detectar interacciones, AC permite representarlas graficamente (Van der Heijden y de Leeuw, 1985).

2) Se relaciona con el llamado modelo RC (Row-Column association model) de Goodman (Van der Heijden y Worsley, 1988).

3) Puede interpretarse como una solución "biplot" (Gabriel, 1971) aplicada a la matriz $D_r^{-1/2} F \; D_c^{-1/2}$ (Cuadras *et al.*, 1985).

4) Se puede introducir como un análisis de correlación canónica entre dos conjuntos de variables categóricas. Los valores singulares contenidos en Λ pueden ser interpretados como correlaciones canónicas (Cuadras, 1981, cap.22).

5) Se verifica

$$tra \; \Lambda^2 = \chi^2 / N$$

donde χ^2es el estadístico *ji*-cuadrado del test de independencia en tablas de contingencia. Existe así una evidente relación con la χ^2 interpretada como una medida de divergencia entre $F=(f_{ij})$ y $E=(f_{i.} \; f_{.j})$, tabla construida bajo la hipótesis de independencia.

6) Puede ser planteado como un método de RA ("reciprocal averaging"), que ya había sido propuesto por diferentes autores (Horst, Richardson, Fisher) y recuperado por Hill (1973) para el análisis de datos ecológicos.

7) Finalmente AC es prácticamente equivalente a los métodos conocidos como "dual scaling" y "optimal scaling". introducidos y estudiados por diversos autores (Fisher, Guttman, Kendall, Lancaster, Nishisato, etc.). Véase Takeuchi *et al.* (1982), Greenacre (1984).

3. ALGORITMOS DE CLASIFICACION, MDS Y COMPUTACION DE DISTANCIAS

La descripción de algoritmos de clasificación jerárquica, equivalentes a la construcción de una distancia ultramétrica a partir de otra distancia dada, así como la descripción del MDS, son tareas que desbordan totalmente las intenciones de este artículo. Me limitaré a resaltar, de nuevo, que existen dos tipos de estructuras: el par (Ω,δ), formado por un conjunto Ω y una distancia δ, que recoge (en algún sentido apropiado) las relaciones entre los datos, y el espacio geométrico modelo (V,d), gracias al cual se dispone de una representación reconocible de (Ω,δ). En general, es necesario algún criterio de aproximación para pasar del primero al segundo. Este paso se lleva a cabo mediante un algoritmo adecuado

$$(\Omega,\delta) \xrightarrow{\text{algoritmo}} (V,d)$$

Si el espacio (V,d) es ultramétrico se trata entonces de un algoritmo de clasificación jerárquica. Si el espacio es euclídeo, entonces necesitamos algunos de los métodos de MDS, como se comenta brevemente en la sección 4.1.

Por otra parte, los aspectos computacionales de las distancias estadísticas constituyen un aspecto importante del tema, que merece un estudio aparte que no sería completo sin una experimentación con diversos conjuntos de datos. Se trata también de una tarea que desborda las intenciones que se han pretendido en este artículo.

En general, los paquetes de programas sobre análisis multivariante de datos contienen rutinas que calculan distancias, bien explícitamente o de forma implícita en función del método empleado. CLUSTAN es el paquete más completo en este sentido, aunque también debemos mencionar BMDP, SPAD, etc., como comenta J. M. García-Santesmases. Además del conocido SPSS, otros paquetes de programas dignos de mención son: GENSTAT, MDS(X) y NTSYS.

4. DIVERGENCIAS VERSUS DISTANCIA DE RAO

Al tratar sobre las distancias estadísticas entre distribuciones de probabilidad, este artículo expone tanto las divergencias funcionales como la distancia geodésica o distancia de Rao. Sin embargo, se deja entrever una ligera preferencia hacia la distancia de Rao. En efecto, la distancia de Rao, por sus connotaciones geométricas y sus interesantes propiedades, viene a ser la generalización natural a distribuciones paramétricas cualesquiera de la distancia de Mahalanobis, que la engloba como caso particular.

Las medidas de divergencia, como la de Kullback-Leibler (KL), poseen una formulación más simple y gozan también de interesantes propiedades y aplicaciones, como bien expone D. Peña. De hecho, existe una fuerte relación entre la función de verosimilitud y KL cuando la distribución pertenece a la familia exponencial. La ventaja de considerar KL queda patente en el caso de que la verdadera distribución q no pertenezca a una determinada familia paramétrica. Supongamos que q realmente no pertenece a la familia exponencial p_θ, y sin embargo deseamos estimar θ. Entonces la función de verosimilitud es esencialmente una estimación de la divergencia KL entre la q y p_θ. Además el estimador máximo verosímil $\hat{\theta}_n$, que no estima el verdadero parámetro porque $q \neq p_\theta$, es un estimador consistente y asintóticamente normal de θ^*, siendo θ^* el valor del parámetro tal que $K(q, p_\theta)$ es mínimo. En otras palabras, el estimador máximo verosímil proporciona una estimación del valor del parámetro que da lugar a la distribución de la familia lo más próxima posible, respecto a KL, a la verdadera distribución. En un contexto similar podemos situar los comentarios de J. M. Prada.

No obstante, aunque KL y otras medidas de divergencia son medidas de discrepancia razonables entre dos distribuciones, en ciertos casos pueden ser más apropiadas otras medidas. Si bien faltaría una demostración más elocuente de la superioridad de la distancia de Rao, he aquí algunos argumentos en su favor, en tanto que queda como problema abierto compararla con otras distancias en un contexto más general.

1) La divergencia KL no es propiamente una distancia, pues no es simétrica, y aunque puede simetrizarse, no cumple la desigualdad triangular.

2) Las divergencias son medidas de discrepancia funcional perfectamente apropiadas en el caso no paramétrico. Cuando se conoce una parametrización, aprovechamos mejor la información que proporciona la muestra utilizando la distancia de Rao, que viene a medir el cambio de información entre los parámetros.

3) La distancia de Rao posee mayor poder de separación que la distancia de Matusita, extensamente utilizada en problemas de inferencia estadística, como se refleja en la relación (sección 6.2)

$$M(p,q) \leq B(p,q) \leq R(p,q)$$

4) Las divergencias y la distancia de Rao coinciden localmente (sección 6.3), lo que tiene como consecuencia que en ciertos casos se llegue a resultados equivalentes.

5) En el caso multinomial, el elemento de arco es

$$ds^2 = \sum (dp_i)^2 / p_i$$

La distancia de Rao se obtiene integrando el elemento de arco a lo largo de una curva geodésica. Obsérvese entonces la analogía formal entre ds^2 y la clásica fórmula del estadístico *ji*-cuadrado

$$\chi^2 = \sum (f_i - e_i)^2 / e_i$$

6) Todos los problemas de inferencia en modelos lineales normales univariantes pueden ser resueltos a través de la distancia de Rao (Burbea y Oller, 1988).

En cuanto a los numerosos problemas en los que interviene la distancia de Mahalanobis, pueden ser también abordados con mayor generalidad utilizando la distancia de Rao o la distancia relacionada introducida en la sección 5.5 (distancia entre individuos). Sin embargo, ocurre que si la estimación de los parámetros alcanza la cota de Cramer-Rao, al ser la matriz de información de Fisher una matriz de covarianzas, esta distancia viene a ser esencialmente la distancia de Mahalanobis.

5. REFERENCIAS

BURBEA, J. y OLLER, J. M. (1988). The information metric for univariate linear elliptic models. *Statistics & Decisions* 6, 209-221.

CUADRAS, C. M. (1981). *Métodos de Análisis Multivariante*. Eunibar, Barcelona.

CUADRAS, C. M., OLLER, J. M., ARCAS, A. y RIOS, J. M. (1985). Métodos geométricos de la Estadística. *Qüestiió* 9(4), 219-250.

GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453-467.

GENSTAT. A general statistical program. Rothamsted Experimental Station, Harpender, Hertfordshire.

GOWER, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27, 857-871.

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, Inc., London.

HILL, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination. *J. Ecol.* 61, 237-249.

MDS(X). Multidimensional Scaling Package. Univ. of Edinburgh.

NTSYS. Numerical Taxonomy System of Multivariate Statistical Programs. Dept. of Ecology and Evolution. The State University of New York at Stony Brook.

TAKEUCHI, K., YANAI, H. y MUKHERJEE, B. N. (1982). *The Foundations of Multivariate Analysis*. Wiley El., New Delhi.

VAN DER HEIJDEN, P. G. M. y de LEEUW, J. (1985). Correspondence analysis used complementary to loglinear analysis. *Psychometrika* 50, 429-447.

VAN DER HEIJDEN, P. G. M. y WORSLEY, K. J. (1988). Comment on "Correspondence analysis used complementary to loglinear analysis". *Psychometrika* 53, 287-291.