QÜESTIIÓ, Vol 17, 1 pp. 39-74, 1993

APLICACIÓN DE LAS DISTANCIAS EN ESTADÍSTICA

C.M. CUADRAS y J. FORTIANA

Universitat de Barcelona

Este artículo ofrece una panorámica actualizada de la relevancia en Estadística de ideas geométricas basadas en el concepto de distancia. Sus aplicaciones se agrupan según cuatro áreas: Estimación Puntual, Contrastes de Hipótesis, Representación de Conjuntos y Predicción basada en Distancias.

Applying distances in Statistics.

Key words: Mahalanobis Distance, Rao Distance, Principal Coordinate Analysis, Geometric Representation of Finite Sets, Distance–based Prediction.

1. INTRODUCCIÓN

Desde su principio, la Estadística moderna ha dependido de la Teoría de la Probabilidad, del Análisis, la Teoría de la Medida y del Algebra. La metodología estadística no podría avanzar sin los recursos que proporcionan estas áreas de la Matemática.

También desde los principios, la Geometría, y especialmente las propiedades topológicas derivadas del concepto de distancia, han desempeñado un papel im-

---

C.M. Cuadras y J. Fortiana. Departament d'Estadística. Facultat de Biología. Avgda. Diagonal, 645. 08028 Barcelona.

–Article rebut el gener de 1993.

–Acceptat l'abril de 1993.

portante en Estadística, aunque su incorporación como elemento de trabajo es más reciente.

Sus primeros usos están latentes en el test Ji–cuadrado de K. Pearson y en el test t de Student, donde las discrepancias entre *observado* y *esperado* se miden mediante un estadístico que en el fondo es una distancia. Tales ejemplos, y muchos otros, son casos particulares de la distancia debida a Mahalanobis [52]

$$(1) \qquad (x-y)' \cdot \Sigma^{-1} \cdot (x-y),$$

donde x,$y \in \mathbb{R}^p$, y Σ es una matriz de covarianzas adecuada.

La distancia (1) interviene en la propia definición de la distribución normal multivariante, en Análisis Discriminante, en la T^2 de Hotelling, en la detección de *outliers*, etc., e incluso, como se ve en la sección , interviene en cualquier contraste de hipótesis.

Nos parece oportuno citar los trabajos de Hotelling [41] y Weyl [75], pioneros en la aplicación de la Geometría Diferencial al contraste de hipótesis: Dado el modelo de regresión no lineal

$$y_i = \beta f_i(\theta) + e_i, \qquad (i = 1, \ldots, n),$$

donde las $f_i(\theta)$ son funciones conocidas que dependen de un parámetro θ y los errores $e_1, \ldots, e_n$ son variables aleatorias independientes igualmente distribuidas (iid), con distribución $N(0, \sigma^2)$, consideremos la hipótesis nula $H_0 : \beta = 0$. El estadístico Λ de razón de verosimilitud equivale a

$$W = \max_{\theta} \frac{\left(\sum_i f_i(\theta)\, y_i\right)^2}{\sum_i f_i^2(\theta) \sum_i y_i^2}.$$

Sin embargo, puesto que θ no es identificable cuando $\beta = 0$, no es factible aplicar la teoría asintótica sobre la distribución de Λ, ni los estadísticos equivalentes de Wald y de Rao, asintóticamente distribuidos como Ji–cuadrado. Véase Rao [68, pág. 417] y la sección 3.1.

Empleando las notaciones: $f(\theta) = (f_1(\theta), \ldots, f_n(\theta))$, $\gamma(\theta) = f(\theta)/\|f(\theta)\|$, $y = (y_i, \ldots, y_n)$, $\langle \cdot, \cdot \rangle$ para el producto escalar, y $U = y/\|y\|$, la región de rechazo toma la forma $\{\max_\theta \langle \gamma(\theta), U \rangle \geq W^2\}$, y puede ser descrita utilizando términos estrictamente geométricos, como el de distancia geodésica, relativos a la esfera unidad en el espacio $\mathbb{R}^n$. Véase Knowles y Siegmund [49].

Este ejemplo, nada trivial, es sólo una muestra de los innumerables campos de la Estadística y el Análisis de Datos en los que es crucial el concepto de distancia. En este trabajo presentamos, junto a aplicaciones recientes de

dicho concepto, una revisión de ciertos aspectos de otros más clásicos, como continuación de [16, 25], por lo que algunos temas reaparecen por razones de coherencia. Organizaremos la exposición considerando los siguientes apartados:

- Estimación puntual
- Contraste de hipótesis
- Representación de conjuntos
- Modelos de predicción

2. ESTIMACIÓN PUNTUAL

2.1. En modelos lineales

La utilización más clara y elegante del concepto de distancia se consigue en el estudio del modelo lineal

$$y = X \cdot \beta + e,$$

donde la estimación del vector paramétrico β es aquel $\widehat{\beta}$ tal que $\widehat{y} = X\widehat{\beta}$ verifica que ${R_0}^2 = \|y - \widehat{y}\|^2$ es mínimo. Además, si $e \sim N(0, \sigma^2 I_n)$, entonces se verifica que ${R_0}^2/\sigma^2 \sim {\chi^2}_{n-r}$, siendo $r = \operatorname{rang}(X)$, resultado básico del Análisis de la Varianza.

Sea $\Psi = P \cdot \beta = (\psi_1, \ldots, \psi_q)'$ un vector de funciones paramétricas estimables, es decir, $\mathcal{F}(P) \subset \mathcal{F}(X)$, donde la notación $\mathcal{F}(\cdot)$ indica el subespacio generado por las filas de una matriz. La hipótesis $H_0 : \Psi = \Psi_0$ se decide mediante

$$F = \frac{(\widehat{\Psi} - \Psi_0)' \cdot (P \cdot (X'X)^- \cdot P')^{-1} \cdot (\widehat{\Psi} - \Psi_0)}{{R_0}^2} \times \frac{n-r}{q},$$

siendo $\widehat{\Psi} = P \cdot \widehat{\beta}$ la estimación Gauss–Markov de Ψ. Nótese que el numerador de F es una distancia tipo Mahalanobis entre $\widehat{\Psi}$ y Ψ_0.

2.2. Divergencia de Kullback–Leibler

La divergencia de Kullback–Leibler entre dos funciones de densidad p, q con respecto a una medida μ,

$$K(p,q) = \int p \log(p/q)\, d\mu, \tag{2}$$

juega un importante papel en el llamado problema de *la especificación* en inferencia estadística. Supongamos, para concretar, que μ es la medida de Lebesgue, y sea $\Gamma = \{p(x,\theta),\ \theta \in \Theta\}$ un modelo estadístico. La verdadera función de densidad es $p(x,\theta_0)$, donde θ_0 es el verdadero valor del parámetro. La divergencia entre $p(x,\theta_0)$ y $p(x,\theta)$ es

$$K\left(p(x,\theta_0),\, p(x,\theta)\right) = \int p(x,\theta_0)\log p(x,\theta_0)\,dx - \int p(x,\theta_0)\log p(x,\theta)\,dx.$$

El valor de θ que minimiza esta divergencia proporciona la densidad que más se acerca a la verdadera y corresponde al máximo de la integral

$$\int p(x,\theta_0)\log p(x,\theta)\,dx, \tag{3}$$

es decir, al máximo del valor esperado de $\log p(x,\theta)$.

Dada una muestra aleatoria simple $x_1,\ldots,x_n$ de una v.a. con densidad $p(x,\theta_0)$, una estima de (3) se obtiene mediante

$$\frac{1}{n}\sum_{i=1}^{n}\log\, p(x_i,\theta).$$

El valor $\widehat{\theta}$ que maximiza este promedio es la estimación máximo verosímil (ML) de θ. Quizás sea menos conocida la siguiente propiedad: Supongamos que la verdadera densidad es q, pero $q \notin \Gamma$. ¿Qué significaría entonces la estimación ML de θ? La divergencia entre q y $p(x,\theta)$ es ahora

$$\int q(x)\,\log\, q(x)\,dx - \int q(x)\,\log p(x,\theta)\,dx,$$

y el *verdadero valor* θ_0 del parámetro θ se puede definir como aquel θ_0 tal que $p(x,\theta_0) \in \Gamma$ es la densidad más próxima a q de acuerdo con la divergencia (2). θ_0 es entonces solución de

$$E_q\left[\partial \log p(x,\theta)/\partial\,\theta\right] = 0, \tag{4}$$

y se dice que q es *consistente* con θ_0. Veamos ahora qué ocurre con el estimador ML $\widehat{\theta}$ obtenido considerando el modelo Γ. Suponiendo las usuales condiciones de regularidad, sea

$$\begin{aligned} Z(x,\theta) &= \partial \log p(x,\theta)/\partial\theta, \\ J(\theta) &= E_q(Z\cdot Z'), \\ H(\theta) &= -E_q\left(\partial Z/\partial\,\theta\right). \end{aligned} \tag{5}$$

En un entorno de θ_0, $Z(x,\theta) = Z(x,\theta_0) + (\theta - \theta_0)\,(\partial Z/\partial\theta)_{\theta_0} + \cdots$, y si $x_1, \ldots, x_n$ son iid como q, entonces

$$\frac{1}{n}\sum Z(x_i,\theta) = \frac{1}{n}\sum Z(x_i,\theta_0) + (\theta - \theta_0)\,\frac{1}{n}\sum \left(\partial Z/\partial\theta\,(x_i,\theta)\right)_{\theta_0}.$$

Haciendo tender $n \to \infty$, teniendo en cuenta (4) y (5), obtenemos la identidad asintótica

$$\frac{1}{n}\sum Z(x_i,\theta) = 0 - (\theta - \theta_0)\,H(\theta_0),$$

que prueba que $\widehat{\theta}$, el estimador ML que anula $\sum Z(x_i,\theta) = 0$, converge a θ_0 en probabilidad. Además, por el teorema del valor medio podemos escribir

$$\sum Z(x_i,\theta) - \sum Z(x_i,\widehat{\theta}) = \sum \left(\partial Z(x_i,\theta)/\partial\theta\right)_{\theta^\star}(\theta - \widehat{\theta}),$$

donde $\theta^\star$ es un punto entre θ y $\widehat{\theta}$. Puesto que $\sum Z(x_i,\widehat{\theta}) \to 0$, $\widehat{\theta} \to \theta_0$, y $(1/n)\sum \partial Z(x_i,\theta)/\partial\theta \to H(\theta)$, tenemos de nuevo la identidad asintótica $(1/n)\sum Z(x_i,\theta) = (\widehat{\theta} - \theta_0)\,H(\theta_0)$, es decir,

$$(1/\sqrt{n})\sum Z(x_i,\theta_0) = \sqrt{n}\,(\widehat{\theta} - \theta_0)\,H(\theta_0).$$

Por el teorema central del límite, $(1/\sqrt{n})\sum Z(x_i,\theta_0)$ es asintóticamente normal de media $E_q\,(Z(x,\theta_0)) = 0$, que es la condición (4), y matriz de covarianzas $J(\theta_0)$. Finalmente tenemos que

$$\sqrt{n}\,(\widehat{\theta} - \theta_0) \overset{a}{\sim} N\left(0, H^{-1}(\theta_0)\cdot J(\theta_0)\cdot H^{-1}(\theta_0)\right).$$

Es decir, el estimador ML $\widehat{\theta}$ es asintóticamente normal y estimador consistente de θ_0, el valor del parámetro más próximo a q respecto la divergencia de Kullback–Leibler.

Ventajas y aplicaciones de la estimación ML del *verdadero valor* θ_0 pueden verse en [48], para el estudio de la robustez del test de razón de verosimilitud tomando densidades alternativas, en [71], para la obtención de intervalos de confianza robustos, en [42], para estimar parámetros en modelos bivariantes de supervivencia, y en [19], para el problema de la estimación de parámetros relativos a densidades multivariantes cuando sólo se conocen las marginales.

2.3. El método de la mínima distancia

Es un método de estimación promovido por J. Wolfowitz en una serie de artículos que culminaron en [76]. Supongamos que la función de distribución de un vector aleatorio es $G \in \Gamma = \{F_\theta,\ \theta \in \Theta\}$. Sean $x_1, \ldots, x_n$ iid como G, y sea

G_n la función de distribución empírica. Si $\delta(G_n, F_\theta)$ es una medida de distancia entre G_n y $G = F_\theta$, el método de la mínima distancia (MD) consiste en tomar como estimación de θ el valor

$$\widehat{\theta} \text{ tal que } \quad \delta(G_n, F_{\widehat{\theta}}) = \inf_{\theta \in \Theta} \delta(G_n, F_\theta).$$

MD es útil como método alternativo de estimación cuando otros métodos no son aplicables. Como distancia se suele tomar la de Kolmogorov

$$\delta_K(G_n, F_\theta) = \sup_{-\infty < x < \infty} |G_n(x) - F_\theta(x)|,$$

o la de Cramér–von Mises

$$\delta_C(G_n, F_\theta) = \int_{-\infty}^{+\infty} [G_n(x) - F_\theta]^2 \, w_\theta(x) \, dF_\theta(x).$$

MD proporciona estimadores que convergen en probabilidad a θ y tienen propiedades de robustez en el caso de desviaciones locales del modelo. Incluso, si $G \notin \Gamma$, tomando δ_C con $w_\theta(x) = 1/f_\theta(x)$, el estimador MD proporciona una estimación $\widehat{\theta}$ tal que $F_{\widehat{\theta}}$ es una proyección $\mathcal{L}^2$ de G_n en Γ (Véase [64]).

MD es especialmente útil en la estimación no paramétrica de funciones (de densidad, de distribución, de regresión, etc.). Supongamos, por ejemplo, que $f(x)$ es la función de densidad. Un resultado clásico es que no existe estimador "razonable" de $f(x)$, en el sentido de que el estimador $\widehat{f}_n(x)$ verifique la igualdad $E\left(\widehat{f}_n(x)\right) = f(x) \ \forall x$, (cfr. [65]). Así, la teoría clásica de la estimación no es aplicable, existiendo razones para considerar estimadores tipo núcleo

$$\widehat{f}_n(x) = \frac{1}{n\, h_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right),$$

donde $h_n \to 0$ para $n \to \infty$, y K es una densidad de probabilidad, por ejemplo

$$K(x) = \begin{cases} 1/2 & \text{si } |x| \leq 1, \\ 0 & \text{si } |x| > 1. \end{cases}$$

Bajo ciertas condiciones se prueba que $\widehat{f}_n(x)$ converge uniformemente a $f(x)$. Un criterio de proximidad en la estimación de $f(x)$ se basa en la distancia $\mathcal{L}^1$, $\delta(\widehat{f}_n, f) = \int_{-\infty}^{+\infty} |\widehat{f}_n(x) - f(x)| \, dx$, pues empleando esta distancia y el estimador tipo núcleo, se verifica que $\delta(\widehat{f}_n, f) \xrightarrow{\text{c.s.}} 0$ para toda f (Devroye and Gyrfi [27]).

Finalmente, el método MD es también útil para estimar θ en el modelo de regresión lineal $y = A(x)' \cdot \theta + e$ donde y es un vector aleatorio y $A(x)$ es un funcional arbitrario, (por ejemplo, $A(x) = (1, x, \ldots, x^k)'$ en regresión polinómica), tomando la distancia de Cramér–von Mises. Véase González Manteiga [34].

3. CONTRASTE DE HIPÓTESIS

El concepto de distancia subyace en la mayor parte de contrastes de hipótesis, jugando la distancia de Mahalanobis un papel muy destacado.

3.1. Distancia de Mahalanobis

El ejemplo paradigmático es el contraste $H_0 : \mu = \mu_0$ para una distribución normal p–variante $N_p(\mu, \Sigma)$, con Σ desconocido. Tanto la razón de verosimilitud como el principio de unión–intersección (véase [53]) nos llevan a considerar el estadístico T^2 de Hotelling

$$T^2 = n\,(\overline{x} - \mu_0)' \cdot S^{-1} \cdot (\overline{x} - \mu_0),$$

donde $\overline{x}$, S son la media y covarianza de una muestra de tamaño n. Así, el test T^2 está basado en la distancia de Mahalanobis entre $\overline{x}$ y μ_0 y por tanto es equivalente al test F ([53, 58]). Para una perspectiva bayesiana de este test, en el caso $H_0 : \mu = 0$, con $\Sigma = \sigma^2 I$, véase [32].

Análogamente, supongamos que $x_1, \ldots, x_n$ son iid según $N_p(\mu_1, \Sigma)$, que $y_1, \ldots, y_n$ son iid según $N_p(\mu_2, \Sigma)$, y consideremos el contraste $H_0 : \mu_1 = \mu_2$. También los criterios clásicos nos llevan al estadístico

$$T^2 = \frac{nm}{m+n}\,(\overline{x} - \overline{y})' \cdot S^{-1} \cdot (\overline{x} - \overline{y}),$$

donde $\overline{x}$, $\overline{y}$, S son los estimadores usuales de μ_1, μ_2, Σ ([53, 58]), es decir, a la T^2 de Hotelling, que es también proporcional a la estimación de la distancia de Mahalanobis entre μ_1 y μ_2.

Más generalmente, consideremos el modelo lineal $Y = X \cdot B + E$ donde Y es $n \times p$, X es $n \times m$, la matriz de parámetros B es $m \times p$ y E es $n \times p$. Se supone que las filas de E son iid $N_p(0, \Sigma)$, con $r = \operatorname{rang}(\Sigma)$. Sea $\Psi' = (\psi_1, \ldots, \psi_p) = P' \cdot B$ una función paramétrica estimable multivariante, $\widehat{\Psi}$ el estimador Gauss–Markov

$$\widehat{\Psi} = P' \cdot \widehat{B} = P' \cdot (X'X)^{-} \cdot X' \cdot Y,$$

y $\widehat{\Sigma} = (n-r)^{-1}\,(Y - X\widehat{B})' \cdot (Y - X\widehat{B})$ la estimación centrada de Σ. Entonces, el contraste de hipótesis $H_0 : \Psi = \Psi_0$, donde Ψ_0 es conocido, se puede decidir mediante el estadístico

$$(\widehat{\Psi} - \Psi_0)' \cdot \widehat{\Sigma}^{-1} \cdot (\widehat{\Psi} - \Psi_0),$$

que es una distancia tipo Mahalanobis y cuya distribución bajo H_0 es también proporcional a una F (cfr. [14, 15]).

En un contexto parecido, la distancia entre dos modelos lineales $Y_i = X \cdot B_i + E_i$, $(i = 1, 2)$, se puede definir como

$$L^2 = \text{tr}\left\{\Sigma^{-1} \cdot (B_1 - B_2)' \cdot X' \cdot X \cdot (B_1 - B_2)\right\}, \tag{6}$$

que puede justificarse como una distancia de Mahalanobis entre dos distribuciones normales $N_p(I_p \otimes X \cdot B_i, \Sigma \otimes I_n)$, $(i = 1, 2)$. Como $L^2 = 0$ si y sólo si $X \cdot B_1 = X \cdot B_2$, la distancia (6) puede servirnos para contrastar la hipótesis $H_0 : X \cdot B_1 = X \cdot B_2$. Para más detalles y generalizaciones, véase [69].

Finalmente, supongamos que la densidad de probabilidad de un vector aleatorio X es $p(x, \theta)$, parametrizado por $\theta \in \Theta$, y que se cumplen las condiciones de regularidad ordinarias. Consideremos la hipótesis compuesta $H_0 : \theta \in \Theta_0 \subset \Theta$. Dada una muestra $x_1, \ldots, x_n$, el procedimiento clásico iniciado por Neyman y Pearson [59] para decidir acerca de H_0 utiliza la razón de verosimilitud

$$\Lambda = \sup_{\theta \in \Theta_0} \mathcal{L} \Big/ \sup_{\theta \in \Theta} \mathcal{L},$$

siendo $\mathcal{L} = \prod_{i=1}^n p(x_i, \theta)$ la función de verosimilitud. Para n grande, el criterio se basa en el estadístico $U = -2 \log \Lambda$ que, bajo H_0, sigue asintóticamente una distribución ji-cuadrado χ^2_{q-r}, siendo $q = \dim(\Theta)$, y $r = \dim(\Theta_0)$.

Un criterio alternativo se debe a Rao [67]. (Véase, por ejemplo, [68]). Se basa en los *efficient scores*

$$Z_i(\theta) = \frac{\partial}{\partial \theta} \log p(x_i, \theta),$$

y en el comportamiento de $V_\theta = (1/\sqrt{n}) \sum_{i=1}^n Z_i(\theta)$. Se verifica que $E(V_\theta) = 0$ y, además, si $\widehat{\theta}$ es el estimador máximo verosímil de $\theta \in \Theta$, entonces $V_{\widehat{\theta}} = 0$. Obsérvese que $\mathcal{F}_\theta = E(Z_i(\theta) \cdot Z_i'(\theta))$ es la matriz de información de Fisher y también la matriz de covarianzas de $Z_i(\theta)$. Puede entonces probarse que la distribución asintótica de $V_\theta' \cdot \mathcal{F}_\theta \cdot V_\theta$, para cada valor de $\theta = (\theta_1, \ldots, \theta_q)$, es χ^2_q.

Rao propone el estadístico $S = V_{\theta^\star}' \cdot \mathcal{F}_{\theta^\star}^{-1} \cdot V_{\theta^\star}$, siendo $\theta^\star$ la estimación máximo verosímil de θ dentro de Θ_0.

Podemos poner $V_{\theta^\star} = \sqrt{n}\,(1/n) \sum_{i=1}^n Z_i(\theta^\star) = \sqrt{n} \cdot \overline{Z}_{\theta^\star}$, y como bajo H_0, $\mathcal{F}_{\theta^\star}$ puede considerarse una estimación de $\mathcal{F}_{\theta_0}$, donde θ_0 representa el verdadero valor del parámetro, tenemos que la proximidad de $V_{\theta^\star}$ a $V_{\theta_0} = 0$ favorece la hipótesis nula. Podemos medir esta proximidad mediante la distancia de Mahalanobis entre $\overline{Z}_{\theta^\star}$ y la media esperada 0,

$$\left(\overline{Z}_{\theta^\star} - 0\right)' \cdot \mathcal{F}_{\theta^\star}^{-1} \cdot \left(\overline{Z}_{\theta^\star} - 0\right) = n\,S.$$

Ahora bien, según se muestra en [68], se cumple la igualdad asintótica

$$U = -2\log\Lambda \stackrel{a}{=} V_{\theta^\star}{}' \cdot \mathcal{F}_{\theta^\star}{}^{-1} \cdot V_{\theta^\star},$$

de modo que la razón de verosimilitud, el estadístico más utilizado en contrastes de hipótesis, resulta ser asintóticamente equivalente a una distancia de Mahalanobis, por ser $\mathcal{F}_{\theta^\star}$ la estimación de una matriz de covarianzas.

Como ilustración, consideremos la hipótesis nula $H_0 : \theta = \theta_0$ para una v.a. con densidad exponencial $p(x,\theta) = \theta^{-1}\exp(-\theta^{-1}x)$, $x > 0$, $\theta \in \Theta = \mathbb{R}_+$. La razón de verosimilitud es

$$\Lambda = \left(\frac{\overline{x}}{\theta_0}\right)^n \exp\left[n\left(1 - \frac{\overline{x}}{\theta_0}\right)\right].$$

Por otra parte, el estadístico de Rao es

$$S = V_{\theta_0}{}' \cdot \mathcal{F}_{\theta_0}{}^{-1} \cdot V_{\theta_0} = \frac{\sqrt{n}}{\theta_0{}^2}(\overline{x} - \theta_0)\,\theta_0{}^2\,\frac{\sqrt{n}}{\theta_0{}^2}(\overline{x} - \theta_0) = n\,\frac{(\overline{x} - \theta_0)^2}{\theta_0{}^2},$$

donde $\overline{x}$ es la media muestral en muestras de tamaño n. Claramente la distribución asintótica de S es ${\chi^2}_1$ y más simple que

$$-2\log\Lambda = -2n\left[\log\left(\frac{\overline{x}}{\theta_0}\right) + \left(1 - \frac{\overline{x}}{\theta_0}\right)\right].$$

La equivalencia asintótica se deduce fácilmente de que, para n grande, podemos suponer $-1 < (\overline{x} - \theta_0)/\theta_0 \leq 1$, así que vale el desarrollo de Taylor

$$\log\left(\frac{\overline{x}}{\theta_0}\right) = \log\left(1 + \left(\frac{\overline{x}}{\theta_0} - 1\right)\right) = \frac{(\overline{x} - \theta_0)}{\theta_0} - \frac{(\overline{x} - \theta_0)^2}{2\theta_0{}^2} + \cdots,$$

y de aquí resulta $-2\log\Lambda \stackrel{a}{=} S$.

3.2. Distancia de Matusita

Sean F_1, F_2 funciones de distribución, y sean f_1, f_2 las funciones de densidad respecto una cierta medida μ, que supondremos es la medida de Lebesgue. La distancia de Matusita se define como

$$\delta_M^2(F_1, F_2) = \int \left\{\sqrt{f_1(x)} - \sqrt{f_2(x)}\right\}^2 dx = 2(1 - \rho), \tag{7}$$

donde $\rho = \int \sqrt{f_1(x)\,f_2(x)}\,dx$ es la llamada *afinidad* entre F_1 y F_2.

La distancia (7), introducida por Matusita [54], aunque es también conocida como distancia de Hellinger, ha sido aplicada en problemas de estimación, decisión y análisis discriminante. Por ejemplo, la hipótesis $H_0 : F_1 = F_2$ es equivalente a $H_0 : \delta_M^2(F_1, F_2) = 0$. Se acepta H_0 si $\delta_M^2(F_1, F_2) \leq \eta_\epsilon$, donde $\eta_\epsilon > 0$ depende del nivel de significación ϵ y de los tamaños muestrales m, n. La decisión se toma empleando la distancia $\delta_M^2(S_1, S_2)$ entre las funciones de distribución empíricas.

Matusita [55] discute extensamente la utilización de la distancia (7) en el caso normal $N(\mu, \Sigma)$. Consideremos algunos ejemplos:

1) La hipótesis $H_0 : \mu = \mu_0$ se decide a través de $\delta_M^2(F, S_n)$, donde F es $N(\mu_0, \Sigma)$, y S_n es $N(\overline{x}, S)$, siendo $\overline{x}$ y S la media y covarianza muestrales.

2) La hipótesis $H_0 : \Sigma = \Sigma_0$ se decide calculando la distancia, o lo que es lo mismo, la afinidad entre $N(\mu, \Sigma)$ y $N(\mu, \Sigma_0)$,

$$\rho = \left|{\Sigma_0}^{-1} \Sigma^{-1}\right|^{1/4} \cdot \left|1/2 \left({\Sigma_0}^{-1} + \Sigma^{-1}\right)\right|^{-1/2}.$$

3) La hipótesis de que $X = (x_1, \ldots, x_p)$, con distribución $N(\mu, \Sigma)$, verifica que los vectores aleatorios $x_1, \ldots, x_p$ son estocásticamente independientes, es decir, que $\Sigma = \mathrm{diag}\,(\Sigma_{11}, \ldots, \Sigma_{pp})$, se decide calculando el supremo

$$\rho = \sup_{\Sigma \in M_0} \left(\left|\Sigma^{-1} S^{-1}\right|^{1/4} \left|1/2 \left(\Sigma^{-1} + S^{-1}\right)\right|^{-1/2} \right),$$

siendo M_0 la clase de matrices con cajas en la diagonal y cero en el resto.

Sin embargo, tanto la distancia de Matusita como otras de formulación parecida, en el caso de normalidad multivariante, vienen a ser funciones crecientes de la distancia de Mahalanobis. Una ventaja de la distancia de Matusita es que puede ser aplicada a variables discretas [30], y a variables mixtas [50]. De todos modos, sus aplicaciones se centran más bien en el área del Análisis Discriminante (véase [51]).

3.3. Distancia de Rao

Aunque introducida por Rao [66] hace bastante tiempo, ha sido estudiada más recientemente por Atkinson y Mitchell [2], Burbea y Rao [6], Oller y Cuadras [61],[62], Burbea y Oller [7], y otros.

Un modelo estadístico $\{p(x, \theta), \theta \in \Theta\}$, con estructura de variedad diferenciable procedente de la inclusión de Θ en algún $\mathbb{R}^n$, se dota de la estructura riemanniana cuya métrica en el punto $p(x, \theta)$ se expresa por la matriz de información de Fisher $\mathcal{F}_\theta$. La *distancia de Rao* $\delta_R(F, G)$ entre dos distribuciones

F y G pertenecientes a una misma familia paramétrica, es la distancia geodésica entre los correspondientes puntos de la variedad. Se conoce para bastantes distribuciones [16], aunque el caso normal multivariante ha sido sólo en parte resuelto [9].

La distancia de Rao puede ser utilizada, como la de Matusita, en el contraste de $H_0 : F = G$, en su forma equivalente $H_0 : \delta_R(F,G) = 0$. Bajo condiciones de regularidad generales se demuestra (cfr. [8]) que $V = n_1\, n_2\, \widehat{\delta_R^2}(F,G)/(n_1 + n_2)$ sigue asintóticamente una $\chi^2{}_p$, siendo p el número de parámetros y $\widehat{\delta_R^2}$ una estimación máximo verosímil de δ_R^2.

Como ejemplo de aplicación, consideremos el modelo lineal normal $Y \sim N(X \cdot \beta, \sigma^2 I_n)$, con $\theta = (\beta, \sigma) \in \mathbb{R}^m \times \mathbb{R}_+$. Dada una matriz H de hipótesis demostrable, una región crítica para decidir sobre $H_0 : H \cdot \beta = 0$, es de la forma $W = \{x \in \mathbb{R}^n : \delta_R(\widehat{\gamma}, H) > \eta_\epsilon\}$, siendo $\widehat{\gamma} = (\widehat{\beta}, \widehat{\sigma})$ la estimación ML de (β, σ), y $\delta_R(\widehat{\gamma}, H) = \inf\{\delta_R(\widehat{\gamma}, \gamma) : \gamma \in \Theta_H\}$ la distancia de Rao entre $\widehat{\gamma}$ y la subvariedad $\Theta_H = \{\gamma = (\beta, \sigma) : H\,\beta = 0\}$. Puede probarse que este test equivale al F clásico.

Un estudio más general de este test mediante la distancia de Rao sobre la familia de densidades elípticas

$$p(x, \beta, \sigma) = \Gamma(n/2)\, \pi^{-n/2}\, |\Sigma_0|^{-1/2}\, \sigma^{-n}\, F\left(\sigma^{-2}(y - X\,\beta)'\, \Sigma_0^{-1}\, (y - X\,\beta)\right),$$

donde F es una función no negativa sobre $\mathbb{R}_+$ satisfaciendo la condición de normalización, Σ_0 y X son matrices fijas, se debe a Burbea y Oller [7]. Véase también [63].

Aunque este planteamiento y el de Matusita son muy parecidos, conviene observar que si F y G pertenecen a una misma familia paramétrica, se cumple que $\delta_M(F,G) \leq \delta_R(F,G)$, es decir, la distancia de Rao tiene mayor poder de separación que la de Matusita, que en cambio es aplicable en un contexto no paramétrico.

Esto se debe a que la distancia de Rao aprovecha el conocimiento de una parametrización: Consideremos la variedad diferenciable de dimensión infinita

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

esfera unidad (o espacio proyectivo) del espacio $\mathcal{L}^2$ de las funciones de cuadrado integrable, dotado de la estructura diferenciable inducida por la estructura natural de espacio de Hilbert con producto escalar $\langle f, g\rangle = \int f\, g\, dx$. Entonces δ_R es la longitud de una curva contenida en la subvariedad de dimensión finita $\Theta \subset \mathcal{E}$, mientras que δ_M es la longitud de la línea recta entre dos puntos de $\mathcal{E}$.

No obstante, justo es añadir que las distancias de Matusita, Rao, y otras medidas de divergencia, coinciden localmente [6]. Véase [57] para el problema de la estimación de la distancia de Rao.

4. REPRESENTACIÓN DE CONJUNTOS

La representación de un conjunto finito U de objetos, indivíduos o estímulos constituye una de las más interesantes aplicaciones de la Estadística basada en la topología asociada a una distancia. Las aplicaciones abarcan muchos campos: Arqueología, Ecología, Genética, Psicología, Sociopolítica, etc. Dedicamos esta sección a las representaciones más usuales de un conjunto finito de elementos, a saber:

1. Representación Euclídea,
2. Representación Ultramétrica (en forma de dendrograma),
3. Representación Cuadripolar (en forma de árbol aditivo),
4. Representación de Robinson (en forma de árbol piramidal).

Haremos especial énfasis en el punto (1), puesto que proporciona una forma general de predicción. En lo sucesivo designaremos convencionalmente $U = \{1, 2, \ldots, n\}$.

Definición 1 *Una* matriz de disimilaridades $\boldsymbol{\Delta} = (\delta_{ij})$ *es una matriz real simétrica* $n \times n$ *cuyos elementos* δ_{ij} *satisfacen* $\delta_{ij} = \delta_{ji} \geq \delta_{ii} = 0$, $\forall i, j \in U$.

Se conocen muchos métodos para construir disimilaridades. Aqui partimos de una $\boldsymbol{\Delta}$ obtenida aplicando uno de dichos métodos, y nos centraremos más en sus propiedades y en el tipo de representación de U que permiten.

Definición 2

1. $\boldsymbol{\Delta}$ *es* Euclídea *si existe una configuración de puntos en un espacio euclídeo* $\mathbb{R}^p$ *cuyas interdistancias coincidan con las contenidas en* $\boldsymbol{\Delta}$, *es decir, si existen* $x_1, \ldots, x_n \in \mathbb{R}^p$ *tales que* $\delta_{ij}{}^2 = (x_i - x_j)' \cdot (x_i - x_j)$, $\forall i, j \in U$.

2. $\boldsymbol{\Delta}$ *es* ultramétrica *si* $n \geq 3$ *y para todas las ternas* $i, j, k \in U$ *se verifica que* $\delta_{ij} \leq \max\{\delta_{ik}, \delta_{jk}\}$.

3. $\boldsymbol{\Delta}$ *es* cuadripolar *si* $n \geq 4$, *y para todas las cuaternas* $i, j, k, l \in U$ *se verifica la llamada desigualdad aditiva o axioma de los cuatro puntos:* $\delta^+{}_{ij} \leq \max\{\delta^+{}_{ik}, \delta^+{}_{jk}\}$, *siendo* $\delta^+{}_{ij} = \delta_{ij} + \delta_{kl}$, $\delta^+{}_{ik} = \delta_{ik} + \delta_{jl}$ *y* $\delta^+{}_{jk} = \delta_{jk} + \delta_{il}$.

4. Δ *es* de Robinson *si* $n \geq 3$, *y para todas las ternas* $i, j, k \in U$ *con* $i \leq j \leq k$ *se verifica que* $\max\{\delta_{ij}, \delta_{jk}\} \leq \delta_{ik}$.

Pasamos ahora a justificar cada una de estas definiciones en el campo de las aplicaciones.

4.1. Representación Euclídea

Existen numerosísimas aplicaciones de este tipo de representación, y son clásicas en Análisis Multivariante. El siguiente teorema es fundamental para todo lo que sigue. La demostración puede encontrarse en [15, 53, 73].

TEOREMA 4.1 *Sea* $\Delta = (\delta_{ij})$ *una matriz* $n \times n$ *de disimilaridades sobre un conjunto finito* U. *Consideremos la matriz* $A = (a_{ij})$, *siendo* $a_{ij} = -\frac{1}{2}\,{\delta_{ij}}^2$, *y* $B = H \cdot A \cdot H$, *donde* $H = I_n - \frac{1}{n}\,\mathbf{1}_n \cdot \mathbf{1}_n'$ *es la matriz* centradora de datos, *con* $\mathbf{1}_n$ *representando el vector* $n \times 1$ *cuyos elementos son todos iguales a* 1. $\|\cdot\|$ *indica la norma euclídea usual.*

Δ *es euclídea si, y sólo si* B *es semidefinida positiva.*

En caso afirmativo, U *puede ser representado por* $x_1, \ldots, x_n \in \mathbb{R}^p$, *siendo* $p = \operatorname{rang}(B)$, *de modo que* ${\delta_{ij}}^2 = \|x_i - x_j\|^2$, $\forall\, i, j \in U$

La solución habitual del Análisis de Coordenadas Principales (Torgerson [74], Gower [35]) parte de la descomposición $B = V \cdot \Lambda \cdot V'$, donde Λ es la matriz diagonal de valores propios de B y V es ortogonal.

La matriz X, consistente en las p columnas no nulas de $V \cdot \Lambda^{1/2}$ verifica que $X \cdot X' = B$, por lo que sus filas constituyen la configuración euclídea deseada, representando el elemento i–ésimo de U por el punto $x'_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$. Las columnas de X (*ejes principales*) se interpretan como variables, de modo que la propia X puede pensarse como una matriz de "datos" para los puntos que representan U en $\mathbb{R}^p$. Estas columnas son vectores propios de B, así que podemos escribir la configuración:

$$
\begin{array}{cc|cccc|c}
 & & \lambda_1 & \lambda_2 & \cdots & \lambda_p & \\
 & 1 & x_{11} & x_{12} & \cdots & x_{1p} & x'_1 \\
U & 2 & x_{21} & x_{22} & \cdots & x_{2p} & x'_2 \\
 & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
 & n & x_{n1} & x_{n2} & \cdots & x_{np} & x'_n
\end{array}
\qquad (\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0)
$$

Esta representación euclídea de U en dimensión reducida goza de excelentes propiedades ([53, pp. 399–400 y 406–407]):

a) Como $X' \cdot \mathbf{1} = 0$, los datos de X son centrados, es decir, se anulan las medias de las columnas. La igualdad $X' \cdot X = \Lambda$ equivale a que la varianza de cada columna de X es proporcional al correspondiente valor propio, y columnas distintas son incorrelacionadas.

b) Optimalidad: La resolución en cada dimensión $k \leq p$ es máxima entre todas las representaciones euclídeas de U en $\mathbb{R}^k$, es decir, si $x_1(k), \ldots, x_n(k)$ son las k primeras coordenadas principales, y $y_1(k), \ldots, y_n(k)$ son las coordenadas de otra representación euclídea de U en dimensión k, entonces

$$\sum_{i,j} \|y_i(k) - y_j(k)\|^2 \leq \sum_{i,j} \|x_i(k) - x_j(k)\|^2 = 2n\,(\lambda_1 + \cdots + \lambda_k).$$

Dando el nombre de *variabilidad geométrica* de U a $\operatorname{tr} B = \frac{1}{2n} \sum_{i,j=1}^{n} \delta_{ij}^2$, medida natural de la dispersión de este conjunto, vemos que la proporción de variabilidad explicada por las k primeras coordenadas principales es

$$P_k = \left(\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i \right) \times 100.$$

Cuando B no es semidefinida positiva, el comportamiento de $\boldsymbol{\Delta}$ se refleja en el siguiente resultado. Véase una demostración en [15, pág. 380].

TEOREMA 4.2 *Supongamos que B tiene $p > 0$ valores propios positivos y $q > 0$ valores propios negativos. Entonces existen $z_1, \ldots, z_n \in \mathbb{R}^p \oplus \boldsymbol{i}\, \mathbb{R}^q$, con $\boldsymbol{i} = \sqrt{-1}$, es decir, $z_j = (x_j, \boldsymbol{i}\, y_j)$, con $x_j \in \mathbb{R}^p$ y $y_j \in \mathbb{R}^q$, $(j = 1, \ldots, n)$, verificando que*

$$\delta_{j\,k}^2 = \|x_j - x_k\|^2 - \|y_j - y_k\|^2, \qquad \forall\, j, k = 1, \ldots, n.$$

Los puntos $z_1, \ldots, z_n$ cuyas distancias reproducen $\boldsymbol{\Delta}$ pueden representarse en forma de una matriz de datos, con una parte real X y una parte imaginaria Y:

		λ_1	λ_2	$\cdots$	λ_p	0	μ_1	μ_2	$\cdots$	μ_q	
	1	$x_{1\,1}$	$x_{1\,2}$	$\cdots$	$x_{1\,p}$	1	$y_{1\,1}$	$y_{1\,2}$	$\cdots$	$y_{1\,q}$	z_1
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U	i	$x_{i\,1}$	$x_{i\,2}$	$\cdots$	$x_{i\,p}$	1	$y_{i\,1}$	$y_{i\,2}$	$\cdots$	$y_{i\,p}$	z_i
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	n	$x_{n\,1}$	$x_{n\,2}$	$\cdots$	$x_{n\,p}$	1	$y_{n\,1}$	$y_{n\,2}$	$\cdots$	$y_{n\,p}$	z_n

siendo $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0 > \mu_1 \geq \mu_2 \geq \cdots \geq \mu_q$. Véase una ilustración de este caso en Oliva *et al.* [60].

4.2. Representación Ultramétrica

Las ultramétricas tienen un papel fundamental en el estudio de las clasificaciones jerárquicas, iniciado por C. Linneo en su famoso *Sistema Natural* y continuado, bajo una perspectiva matemática, por Benzécri [3], Jardine *et al.* [43, 44], Johnson [46], Hartigan [38], Sokal, Rohlf, Sneath y otros, creadores de la Taxonomía Numérica de las especies vegetales y animales.

Esta relación se debe a que una ultramétrica $\boldsymbol{\Delta}$ define una *jerarquía indexada* (C, α) en U, es decir, $C \subset \mathcal{P}(U)$, verificando que

1. $U \in C$, y $\{i\} \in C \quad \forall i \in U$.
2. $\forall c_1, c_2 \in C$, o bien $c_1 \cap c_2$ es $\emptyset$, o bien uno de los dos conjuntos c_1, c_2 está contenido en el otro.
3. Todo $c \in C$ es reunión de los elementos de C que contiene, o bien no contiene ningún otro elemento de C.
4. Existe una aplicación no negativa (*índice* de la jerarquía), $\alpha : C \longrightarrow \mathbb{R}$ tal que $\alpha(\{i\}) = 0$, y $\alpha(c) < \alpha(c')$ si $c \subset c'$.

Dada una matriz de disimilaridades $\boldsymbol{\Delta}$, para cada $r \in \mathbb{R}_+$, la relación binaria $i \sim_r j \iff \delta_{ij} \leq r$ es de equivalencia si y sólo si $\boldsymbol{\Delta}$ es ultramétrica. El conjunto de las clases de equivalencia correspondientes a todos los $r \in \mathbb{R}_+$, es una jerarquía indexada. Obsérvese que se obtienen clases distintas solamente para aquellos r que aparecen como elementos de $\boldsymbol{\Delta}$. Recíprocamente, una jerarquía indexada (C, α) sobre U define una $\boldsymbol{\Delta}$ ultramétrica, siendo $\delta_{ij} = \alpha(c_{ij})$, donde c_{ij} es la mínima clase de C que contiene $\{i\}$ y $\{j\}$.

La representación geométrica de U se realiza mediante un grafo llamado dendrograma. Por ejemplo, la matriz

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

sobre $U = \{a, b, c, d, e, f\}$ es ultramétrica. U puede representarse mediante el dendrograma de la figura 1, que visualiza la jerarquía $C = \{\{a\}_0, \ldots, \{f\}_0, \{a, b, c\}_1, \{d, e\}_2, \{a, b, c, d, e\}_4, U_5\}$, donde se ha indicado el índice de la jerarquía como subíndice.

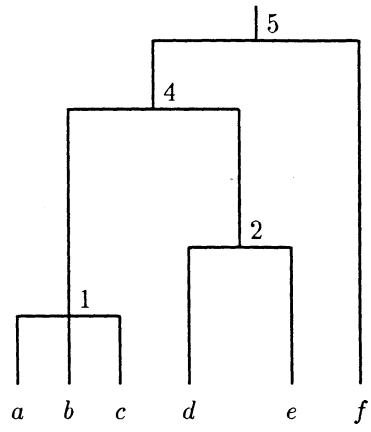

Figura 1.
Dendrograma representando la matriz ultramétrica $\boldsymbol{\Delta}$.

La representación de conjuntos, sea a través de coordenadas principales, sea a través de dendrogramas, es muy frecuente en las aplicaciones (Un ejemplo reciente en lingüística puede verse en [60]). Esta dualidad de representación impulsó a diversos especialistas a relacionarlas entre sí. Gower [36] conjeturó que toda distancia ultramétrica sobre U es euclídea, y propuso una medida del grado de ajuste de unos datos a una representación euclídea, que es la base de la llamada *representación procrusteana*. Tal conjetura fue demostrada por Holman [40], y desde entonces se han obtenido diversos resultados en esta línea que sintetizamos a continuación:

Sea $\boldsymbol{\Delta} = (\delta_{ij})$ una matriz ultramétrica sobre un conjunto finito U de n elementos.

Proposición 1 *Supongamos que $\delta_{ij} > 0$ para $i \neq j$. Entonces $\boldsymbol{\Delta}$ es euclídea $(n-1)$-dimensional.*

Véase [40], [37], [22].

Proposición 2 *Sea $h_1 = \min\{\delta_{ij} : \delta_{ij} > 0\}$. Entonces el mínimo valor propio de la matriz B definida en el teorema (4.1) es $\lambda_1 = \frac{1}{2}h_1^2$.*

Véase [15].

Proposición 3 *Existe una partición $U = U_0 + U_1 + \cdots + U_r$ tal que U_0 está formado por elementos aislados, y cada U_j, para $j = 1, \ldots, r$, es un cluster maximal de elementos equidistantes con distancia común h_j.*

Si μ_0 es el mayor valor propio de B, entonces $\mu_0 > \lambda_r^2 = \frac{1}{2}h_r^2 \geq \cdots \geq \lambda_1^2 = \frac{1}{2}h_1^2$, donde $\lambda_r \geq \cdots \geq \lambda_1$ son tambien valores propios de B.

Además, la matriz X descrita en el teorema (4.1) tiene también una partición según estos valores propios: $X = (X_0|X_1|\cdots|X_r)$, verificándose que cada matriz X_j proporciona una representación euclídea de U_j, para $j = 0, 1, \ldots, r$.

Véase [24].

Proposición 4 *U puede representarse perfectamente en dimensión 1. Es decir, existe una transformación monótona $d_{ij} = f(\delta_{ij})$ de los elementos de $\boldsymbol{\Delta}$, y un vector $t = (t_1, \ldots, t_{n-1})$, con $t_k \geq 0$, $1 \leq k \leq n-1$, verificando que*

$$d_{ij} = \sum_{k=i}^{j-1} t_k \quad \text{para } i < j.$$

Véase [11].

El teorema de Holman (proposición 1) viene a decir que la representación euclídea y la que utiliza un dendrograma son aparentemente opuestas, pues la primera exige dimensión reducida, mientras que la segunda necesita nada menos que dimensión $n-1$. La proposición (3) sirve para clarificar la relación entre ambos tipos de representaciones. La proposición (4) afirma que una transformación monótona de $\boldsymbol{\Delta}$ permite una ordenación euclídea unidimensional que puede ser utilizada como medio de definir el espaciado del eje horizontal del dendrograma.

4.3. Representación Cuadripolar

Si la motivación de las ultramétricas proviene de la necesidad de clasificar atendiendo a la similaridad actual de las especies, la motivación para las matrices cuadripolares tiene su origen en los llamados árboles evolutivos, que clarifican la filogenia de las especies (en lugar de especies podríamos considerar cualquier otro ejemplo).

Un grafo conexo sin ciclos, cuyos ejes tienen longitudes no negativas, y cuyos extremos son los elementos de U, recibe el nombre de *árbol aditivo*. Las longitudes de los caminos que unen los extremos de un árbol aditivo generan una matriz de distancias de tipo cuadripolar. Recíprocamente, si $\boldsymbol{\Delta}$ es cuadripolar, entonces U se puede representar mediante un único árbol aditivo (Buneman, [5]). En particular, la desigualdad aditiva equivale a que toda cuaterna i, j, k, $l \in U$ admite una representación como indica la figura 2.

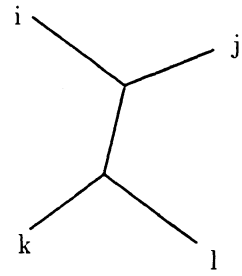

Figura 2.
Grafo de cuatro puntos verificando la desigualdad aditiva.

Un dendrograma es un caso particular de árbol aditivo, con la diferencia esencial de que un árbol aditivo genérico no tiene un punto raiz equidistante de los extremos, ni permite definir una jerarquía indexada. Puede citarse, sin embargo, la siguiente propiedad: Si $\boldsymbol{\Delta} = (\delta_{ij})$ es cuadripolar, existe entonces una matriz ultramétrica $D = (d_{ij})$, y una aplicación $\psi : U \longrightarrow \mathbb{R}$, tal que $\delta_{ij} = d_{ij} + \psi(i) + \psi(j)$, (Sattah y Tversky [72]).

La siguiente matriz sobre $U = \{a, b, c, d, e, f\}$,

$$\Delta_c = \begin{pmatrix} 0.0 & 4.5 & 5.0 & 8.0 & 11.0 & 12.5 \\ & 0.0 & 5.5 & 8.5 & 11.5 & 13.0 \\ & & 0.0 & 7.0 & 10.0 & 11.5 \\ & & & 0.0 & 11.0 & 12.5 \\ & & & & 0.0 & 4.5 \\ & & & & & 0.0 \end{pmatrix},$$

es cuadripolar, y el árbol aditivo que representa U viene en la figura 3. Si se tratara de una árbol evolutivo, las especies a y b tendrían un ancestro común, representado por el nodo n, no perteneciente a U.

Las distancias cuadripolares no son euclídeas en general. Se conoce la siguiente relación, (véase [4]):

Proposición 5 *Sea $\boldsymbol{\Delta}$ una matriz cuadripolar sobre U. Entonces $\boldsymbol{\Delta}^{(\alpha)} = (\delta_{ij}{}^{\alpha})$ es euclídea, siendo $\alpha = (1/2)^k$, para todo entero $k \geq 1$. La dimensión de esta representación es en general $n - 1$.*

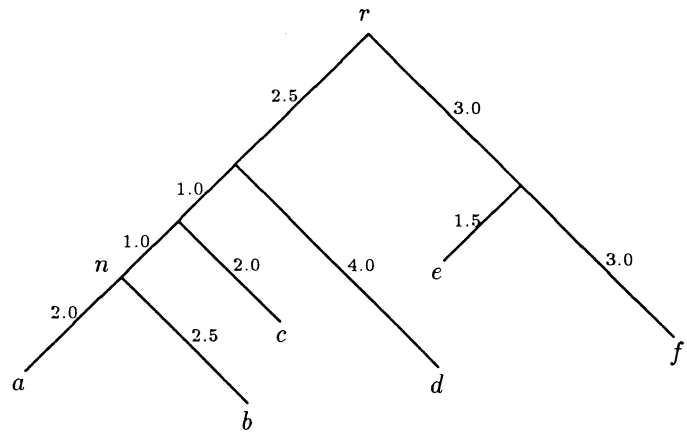

Figura 3.
Árbol aditivo representando la matriz cuadripolar Δ_c.

4.4. Representación de Robinson

La motivación proviene de la necesidad de dar a un conjunto U un orden compatible con la disimilaridad $\boldsymbol{\Delta}$. Por ejemplo, en la seriación (orden cronológico) de objetos arqueológicos, la disimilaridad debe ser menor entre objetos cercanos en el tiempo y mayor entre objetos alejados, es decir, existe una estructura unidimensional de los datos que se manifiesta dando a U un orden adecuado.

La matriz $\boldsymbol{\Delta}$ resultante de esta ordenación verifica la definición (2)(4.) de matriz de Robinson, equivalente a la propiedad de que sus elementos no decrecen cuando nos apartamos de la diagonal principal a lo largo de cualquier fila o columna. Éste fue el planteamiento original de Robinson [70].

Estas matrices aparecen tambien en el estudio de las *pirámides*, una generalización de las jerarquías introducida por Diday [28] y Fichet [33].

Una *pirámide* en U es una clase de conjuntos $P \subset \mathcal{P}(U)$ que verifica:

1. $U \in P$, y $\forall i \in U, \quad \{i\} \in P$.
2. La intersección de cualquier par p, $p' \in P$ puede ser $\emptyset$, o bien $p \cap p' \in P$.

3. Existe una ordenación de U compatible con P.

La última propiedad significa que si $p \in P$ contiene i_1, i_2, entonces todos los elementos comprendidos entre i_1 y i_2 tambien pertenecen a p.

En una pirámide los clusters pueden solaparse: es posible tener $p \subset p'$ y $p \subset p''$ estrictamente, siendo $p' \neq p''$. Sin embargo, cada $p \in P$ tiene un máximo de dos *predecesores inmediatos*, es decir, elementos de P que contienen estrictamente a p sin que exista otro elemento de P comprendido entre los dos. Esta propiedad permite dibujar un diagrama de una pirámide análogo a un dendrograma. Por ejemplo, la siguiente matriz sobre $U = \{a, b, c, d\}$

$$\Delta_R = \begin{pmatrix} 0 & 1 & 2 & 3 \\ & 0 & 1 & 2 \\ & & 0 & 1 \\ & & & 0 \end{pmatrix},$$

es de Robinson, y define la pirámide $P = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, b\}, \{b, c\}, \{c, d\}, \{a, b, c\}, \{b, c, d\}, U\}$. Su representación viene dada en la figura 4. Obsérvese que Δ_R no es ultramétrica.

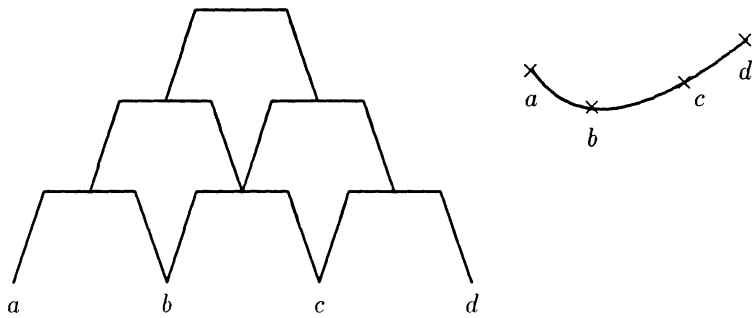

Figura 4.
Representación piramidal correspondiente a la matriz de Robinson Δ_R. A la derecha aparece una posible ordenación cronológica.

Una *pirámide indexada* (P, α) es una pirámide P, con un índice α tal que $\alpha(\{i\}) = 0$, para todos los $i \in P$, y $\alpha(p) \leq \alpha(p')$ si $p \subset p'$. Es *indexada en sentido amplio* si para dos elementos p, p' de P, la inclusión estricta $p \subset p'$, junto con la igualdad $\alpha(p) = \alpha(p')$, implican la existencia de p_1 y p_2 distintos de p tales que $p = p_1 \cap p_2$.

El siguiente resultado generaliza la biyección entre ultramétricas y jerarquías indexadas: Si Δ es de Robinson (salvo permutaciones), entonces U se puede representar mediante una pirámide indexada en sentido amplio y recíprocamente (Diday [29]).

En general, las disimilaridades de Robinson no son cuadripolares ni euclídeas. La relación con la propiedad cuadripolar requiere la siguiente definición: Una disimilaridad $\Delta = (\delta_{ij})$ es *Robinson fuerte* si es de Robinson y para todas las cuaternas ordenadas $i \leq j \leq k \leq l \in U$ se verifica que

$$\begin{array}{rcl} \delta_{ij} = \delta_{ik} & \Longrightarrow & \delta_{hj} = \delta_{hk}, \quad \text{si} \quad h \leq i, \\ \delta_{jl} = \delta_{kl} & \Longrightarrow & \delta_{jm} = \delta_{km}, \quad \text{si} \quad m \geq l. \end{array}$$

La figura 5 visualiza esta propiedad, que puede interpretarse diciendo que j y k aparecen simultáneamente en el tiempo.

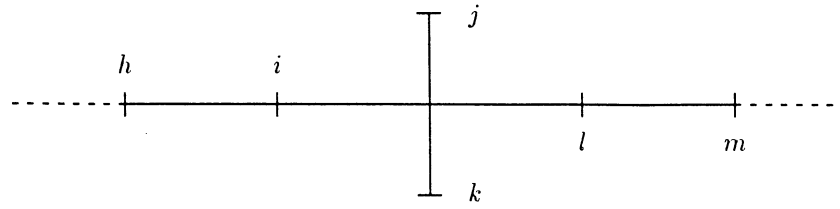

Figura 5.
Ordenación cronológica definida por una matriz Robinson fuerte.

El siguiente resultado describe las matrices de Robinson que pueden ser representadas mediante un árbol aditivo: Si Δ es de Robinson y cuadripolar, entonces es Robinson fuerte (Critchley[12]).

Finalmente, la relación entre las distintas clases de disimilaridades que permiten representar un conjunto finito U, es la siguiente:

$$\begin{array}{ccccc} & & \text{Euclídea} & & \\ & & \Uparrow & & \\ \text{Ultramétrica} & \Longrightarrow & \text{Aditiva} & \Longrightarrow & \text{Métrica} \\ & & \Downarrow & & \\ & & \text{Robinson} & & \end{array}$$

entendiendo por *disimilaridad métrica* aquella que cumple la desigualdad triangular.

5. PREDICCIÓN BASADA EN DISTANCIAS

Sea Y una variable dependiente de un conjunto Ξ de variables, posiblemente de tipo mixto, es decir, conteniendo variables continuas, binarias, y cualitativas. Supongamos que la observación de Ξ sobre un conjunto U de n indivíduos permite obtener una matriz de datos, a partir de la cual construimos una matriz $n \times n$ de distancias Δ. El esquema de la predicción basada en distancias es:

$$\left.\begin{array}{lll} U & \xrightarrow{\Xi} & \Delta \longrightarrow X \\ U & \xrightarrow{Y} & y \\ \{n+1\} & \xrightarrow{\Xi} & \xi_{n+1} \end{array}\right\} \quad y_{n+1} = f(X, y, \xi_{n+1})$$

es decir, la predicción y_{n+1} de Y para un nuevo individuo $\{n+1\}$ es función de la matriz X de coordenadas principales (obtenida de Δ según el teorema 4.1), del vector y de observaciones de Y sobre U, y de las observaciones ξ_{n+1} de Ξ sobre $\{n+1\}$. La formulación general de este problema ha sido presentada por Cuadras [17].

La principal ventaja de estos métodos de predicción reside en que, al depender solamente de distancias entre observaciones, no precisan hipótesis sobre distribuciones de probabilidad. Para variables mixtas, por ejemplo, resulta más natural construir una distancia que postular un modelo probabilístico apropiado.

Vamos a considerar tres tipos de problemas:

1. Predecir una variable contínua Y como una función de regresión de un conjunto Ξ de variables de tipo mixto.

2. Predecir Y cuando la relación con Ξ es no lineal.

3. Predecir Y, discreta con g estados, como un problema de clasificación, siendo Ξ un conjunto mixto de variables.

5.1. Predicción con variables mixtas

Utilizamos el modelo de regresión

$$y = \mu \mathbf{1}_n + X_k \cdot \beta_k + e, \tag{8}$$

donde X_k es una matriz $n \times k$, resultante de elegir $k \leq n-1$ columnas de X según un criterio conveniente, y β_k es un vector $k \times 1$ de parámetros. Este modelo ha

sido estudiado por Cuadras y Arenas [21], probando que:

$$(9) \qquad \widehat{\mu} = \overline{y}, \quad \widehat{\beta}_k = {\Lambda_k}^{-1} \cdot {X_k}' \cdot y, \quad \widehat{y_{n+1}} = \overline{y} + {x_k}' \cdot {\Lambda_k}^{-1} \cdot {X_k}' \cdot y.$$

Λ_k es la matriz diagonal $k \times k$ con los k valores propios de B (ver teorema (4.1)) que corresponden a los vectores seleccionados en X_k, y x_k se obtiene como

$$x_k = \frac{1}{2}{\Lambda_k}^{-1} \cdot {X_k}' \cdot (b - d),$$

donde $b = (b_{11}, \ldots, b_{nn})'$ es el vector columna cuyos elementos son los de la diagonal de B, y $d = (\delta^2_{11}, \ldots, \delta^2_{nn})'$ es el vector columna cuyos elementos son los cuadrados de las n distancias del nuevo individuo $\{n+1\}$ a los de U.

5.2. Predicción no lineal

Supongamos que

$$Y = f(\Xi_1, \ldots, \Xi_p) + e,$$

es decir, Y es una función de regresión no lineal de un conjunto $\Xi = (\Xi_1, \ldots, \Xi_p)$ de p variables, que suponemos contínuas. Sean $(\xi_{i1}, \ldots, \xi_{ip})$ y $(\xi_{j1}, \ldots, \xi_{jp})$ observaciones sobre un par (i, j) de elementos de U. Cuadras [21] prueba que adoptando la distancia δ_{ij} definida por

$$\delta_{ij} = \sqrt{\sum_{h=1}^{p} |\xi_{ih} - \xi_{jh}|},$$

y aplicando el modelo (8), se consigue una buena predicción de Y sin necesidad de conocer f. Una justificación de esta propiedad predictiva del modelo ha sido recientemente encontrada por Cuadras y Fortiana [23] en términos de polinomios de Tchebychev.

5.3. Análisis discriminante

Si Y tiene g estados que corresponden a las poblaciones $\pi_1, \ldots, \pi_g$, y se dispone de una muestra global $U = U_1 \cup U_2 \cup \ldots \cup U_g$ de tamaño n, donde cada U_k es un conjunto de n_k individuos de π_k, predecir Y para un individuo $\{n+1\}$ equivale a clasificarlo en una de las g subpoblaciones. Cuadras [17] estudia una regla de clasificación que parte de las g funciones discriminantes

$$(10) \qquad f_k(\{n+1\}) = \frac{1}{n_k}\sum_{i=1}^{n_k} {\delta_i}^2(k) - \frac{1}{{n_k}^2}\sum_{i=1}^{n_k}\sum_{j=i}^{n_k} {\delta_{ij}}^2(k),$$

donde $\Delta(k) = (\delta_{ij}(k))$ es la matriz de distancias de U_k, y $\delta_i(k)$, $(i = 1, \ldots, n_k)$ las distancias de $\{n+1\}$ a los n_k individuos de esta submuestra. La regla de clasificación es:

[DB] Asignar $\{n+1\}$ a π_i si $f_i(\{n+1\}) = \min\{f_1(\{n+1\}), \ldots, f_g(\{n+1\})\}$.

Este método de discriminación goza de buenas propiedades:

- Coincide con el discriminador lineal clásico cuando δ_{ij} es la distancia de Mahalanobis.
- La estimación de la probabilidad de clasificación errónea es fácilmente calculable.
- En caso de conocerse las probabilidades de asignación *a priori*, éstas se pueden incorporar al modelo.
- Puede ser aplicado correctamente a discriminación con variables mixtas.

Numerosos ejemplos de aplicación de **[DB]**, con datos reales y simulados, han sido estudiados por Cuadras [20].

5.4. Predicción en el caso poblacional

Las fórmulas (9) y (10) se refieren a muestras finitas de una cierta variable y a una o varias matrices de distancias. ¿Pueden generalizarse al caso de variables aleatorias cualesquiera?

La versión poblacional de (10) es simple. Si Ξ es un vector aleatorio con densidad de probabilidad $p_k(\xi)$ respecto a una cierta medida λ en la subpoblación π_k, la función discriminante ligada a una distancia $\delta(\cdot, \cdot)$ es

$$f_k(\xi_0) = H_{k0} - \frac{1}{2} H_k, \quad k = 1, \ldots, g, \tag{11}$$

siendo ξ_0 el individuo a clasificar, $H_{k0} = \int \delta^2(\xi_0, \xi)\, p_k(\xi)\, d\lambda(\xi)$ el valor esperado de $\delta^2(\xi_0, \xi)$ en π_i, y $H_k = \int \delta^2(\xi, \eta)\, p_k(\xi)\, p_k(\eta)\, d\lambda(\xi)\, d\lambda(\eta)$ el valor esperado de $\delta^2(\xi, \eta)$ en $\pi_k \times \pi_k$, donde ξ, η se suponen independientes. La regla de discriminación sigue siendo **[DB]**.

Propiedades destacables de esta regla de discriminación basada en (11) son las siguientes:

1) Si Ξ en π_k es $N(\mu_k, \Sigma)$, y δ^2 es la distancia de Mahalanobis, entonces **[DB]** es equivalente al discriminador lineal. Una sencilla modificación de δ^2 nos proporcionaría el discriminador cuadrático si las matrices de covarianzas son diferentes.

2) En el caso de una variable discreta genérica con m estados y probabilidades $(p_{k1}, \ldots, p_{km})$ para la población π_k, si adoptamos (cfr. [56]) la distancia

$$\delta^2(\xi_1, \xi_2) = (1 - \delta_{rs})(p_{kr}^{-1} + p_{ks}^{-1}),$$

cuando se han presentado los estados r y s para ξ_1 y ξ_2, respectivamente, entonces la regla se reduce a asignar ξ_0 a aquélla π_k tal que la probabilidad p_{kr} es máxima.

3) Si Ξ_1 y Ξ_2 son vectores independientes con distancias asociadas δ_1, δ_2, y tomamos para $\Xi = (\Xi_1, \Xi_2)$ la distancia

$$\delta^2(\cdot, \cdot) = \delta_1^2(\cdot, \cdot) + \delta_2^2(\cdot, \cdot),$$

entonces $f_k(\xi) = f_k(\xi_1) + f_k(\xi_2)$.

4) Supongamos que se conocen las probabilidades a priori de observar $\pi_1, \ldots, \pi_g$, es decir,

$$q_k = P(\pi_k), \quad k = 1, \ldots, g, \quad \sum_{k=1}^{g} q_k = 1.$$

Entonces se puede probar que la función discriminante es

$$f_k(\xi_0) = H_{k0} - \frac{1}{2} H_k + (q_k^{-1} - 1) \quad k = 1, \ldots, g. \tag{12}$$

Obsérvese que una probabilidad alta q_k para π_k proporcionará un valor bajo en (12), luego tenderemos a asignar ξ_0 a π_k.

Como es bien sabido, la regla óptima de clasificación es la regla de Bayes basada en

$$B_{kl}(\xi_0) = V_{kl}(\xi_0) + \log q_k - \log q_l,$$

donde la función discriminante $V_{kl}(\xi_0) = \log p_k(\xi_0) - \log p_l(\xi_0)$ da lugar a la regla de máxima verosimilitud (que coincide con la regla de Bayes si las probabilidades a priori son iguales).

En general, la regla basada en (12) es distinta. Sin embargo, en los casos multinomial y normal multivariante, puede probarse que B_{kl}, V_{kl}, y la regla basada en distancias (12), proporcionan los mismos resultados, o resultados bastante similares (véase Cuadras [18]).

Consideramos finalmente la extensión contínua de (8) al caso de la regresión de una variable Y sobre un vector aleatorio X. La mejor solución, si fuera conocida la distribución conjunta de (Y, X), es la curva de regresión de la media de Y sobre X. El modelo (8) requiere obtener las coordenadas principales a

partir de una matriz B de orden $n \times n$, luego parece que al pasar de una muestra a la población, (es decir, haciendo $n \to \infty$), nos vayamos a encontrar con un problema insuperable. No obstante, una extensión continua es posible cuando X es una variable uniforme $(0,1)$ y se utiliza la distancia

$$\delta(u,v) = \sqrt{|u-v|} \quad u,v \in (0,1).$$

Entonces, las coordenadas principales se asocian al sistema numerable de variables centradas e incorrelacionadas

$$\left\{-(\sqrt{2}/j\,\pi)\,\cos(j\,\pi\,X)\right\}_{j\in\mathbb{N}}, \tag{13}$$

cumpliéndose formalmente las propiedades del Teorema 4.1. La generalización del modelo de predicción equivale entonces a una regresión múltiple sobre un subconjunto finito de (13). Para más detalles, véase Cuadras y Fortiana [23].

6. LECTURAS ADICIONALES

Con esta exposición hemos tratado de proporcionar una visión general de las aplicaciones a la Estadística del concepto de distancia. La importancia que este tema posee se demuestra por la reciente celebración del congreso internacional DISTANCIA'92, organizado por el "European Network of Mathematical Structures for Dissimilarity Analysis" (Rennes, 22–26 Junio, 1992). Las actas del congreso [47], son una extensa recopilación de contribuciones en aspectos teóricos, metodológicos y aplicados.

Monografías recientes de interés para el lector que desee ampliar información son: [13], una visión general de las distancias en Estadística, con una declaración de perspectivas futuras, [10], una amplia exposición de la metodología basada en distancias aplicada a series temporales y a procesos estocásticos, y el libro de U. Jensen [45], dedicado en su totalidad al estudio de la distancia de Rao, con aplicaciones a la Econometría.

7. REFERENCIAS

[1] **Arrow, K.J.** (1951). *Social Choice and Individual Values.* Wiley.

[2] **C. Atkinson** and **A. F. S. Mitchell** (1981). "Rao's distance measure". *Sankhyā*, **43A**, 345–365.

[3] **J. P. Benzécri** (1965). "Problèmes et méthodes de la Taxinomie". Publ. Inst. Statistique. Univ. de Paris.

[4] **G. Brossier** and **G. Le Calve** (1985). "Analyse des dissimilarités sous l'éclairage $\sqrt{D}$. Application a la récherche d'arbres additifs optimaux". *INRIA, 4th. Int. Symp. Data Analysis and Information*, Tome 1, pp. 17–26.

[5] **P. Buneman** (1971). *The recovery of trees from measures of dissimilarity*, en [39, pp. 387–395].

[6] **J. Burbea** and **C. R. Rao** (1982). "Entropy differential metric, distance and divergence measures in probability spaces: a unified approach". *J. Multivariate Anal.*, **12**, 575–596.

[7] **J. Burbea** and **J. M. Oller** (1988). "The information metric for univariate linear elliptic models". *Statistics & Decisions*, **6**, 209–221.

[8] **J. Burbea** and **J. M. Oller** (1989). "On Rao distance asymptotic distribution". Univ. de Barcelona Math. *Preprint Series* **67**.

[9] **M. Calvo** and **J. M. Oller** (1990). "A Distance between Multivariate Normal Distributions based in an Embedding into the Siegel Group". *J. Multivariate Anal.*, **35** 223–242.

[10] **M. Corduas** (1992). "Misure di distanza tra serie storiche e modelli parametriche". *Quaderni dell'Instituto Economico Finanziario*, N.º **3**, Universita degli Studi di Napoli.

[11] **F. Critchley** and **W. Heiser** (1988). "Hierarchical trees can be perfectly scaled in one dimension". *Journal of Classification*, **5**, 5–20.

[12] **F. Critchley** (1989). "On exchangeability–based equivalence relations induced by strongly Robinson and, in particular, by quadripolar Robinson dissimilarity matrices". Dept. of Statistics, Univ. of Warwick, *Tech. Report* **152**.

[13] **F. Critchley, P. Marriott** and **M. Salmon**. "Distances in Statistics". *Proceedings XXXVI Riunione Scientifica*, Societa Italiana di Statistica, Roma: CISU, pp. 39–60.

[14] **C. M. Cuadras** (1974). "Análisis discriminante de funciones paramétricas estimables". *Trab. Estad. Inv. Oper.*, **25** 3–31.

[15] **C. M. Cuadras** (1991). *Métodos de Análisis Multivariante* EUNIBAR, Barcelona (1981). 2^a edición, PPU, Barcelona.

[16] **C. M. Cuadras** (1988). "Distancias estadísticas". *Estadística Española*, **30**, 295–378.

[17] **C. M. Cuadras**. "Distance Analysis in discrimination and classification using both continuous and categorical variables", en [31, pp. 459–473].

[18] **C. M. Cuadras** (1991). "A distance based approach to Discriminant Analysis and its properties". Univ. de Barcelona Math. *Preprint Series* **90**.

[19] **C. M. Cuadras** (1992). "Probability distributions with given multivariate marginals and given dependence structure". *J. of Multivariate Analysis,* **42,** 51–66.

[20] **C. M. Cuadras** (1992). "Some examples of distance based discrimination". *Biometrical Letters,* **29(1),** 3–20.

[21] **C. M. Cuadras** and **C. Arenas** (1990). "A distance based regression model for prediction with mixed data". *Commun. Statist. -Theory Meth.,* **19,** 2261–2279.

[22] **C. M. Cuadras** and **F. Carmona** (1983). "Dimensionalitat euclidiana en distàncies ultramètriques". *Qüestió,* **7,** 353–358.

[23] **C. M. Cuadras** and **J. Fortiana** (1993). "Continuous Metric Scaling and Prediction", en: *Multivariate Analysis: Future Directions 2.* (C.M. Cuadras and C.R. Rao, eds.). Elsevier, Amsterdam. (in press).

[24] **C. M. Cuadras** and **J. M. Oller** (1987). "Eigenanalysis and metric multidimensional scaling on hierarchical structures". *Qüestió,* **11,** 37–58.

[25] **C. M. Cuadras, J. M. Oller, A. Arcas** and **M. Ríos** (1985). "Métodos geométricos de la Estadística". *Qüestió,* **9,** 219–250.

[26] **J. De Leeuw, W. Heiser, J. Meulman** and **F. Critchley, (eds.)**. *Multidimensional Data Analysis* DSWO Press, Leiden.

[27] **L. Devroye** and **L. Gyrfi** (1985). *Nonparametric Density Estimation: The L_1 View* John Wiley & Sons, New York.

[28] **E. Diday** (1984). "Une répresentation visuelle des classes empiétantes: les pyramides". *Rapport de Recherche INRIA*, No. **291**.

[29] **E. Diday** (1986). "Orders and overlapping clusters in pyramids", en [26, pp. 201–234].

[30] **W. R. Dillon** and **M. Goldstein** (1978). "On the performance of some multinomial classification rules". *J. Am. Stat. Assoc.,* **73,** 305–313.

[31] **Y. Dodge (ed.)** (1989). *Statistical Data Analysis and Inference* North–Holland, Amsterdam.

[32] **J. R. Ferrandiz** (1985). "Bayesian inference on Mahalanobis distance: An alternative Approach to Bayesian model testing", en *Bayesian Statistics 2*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (Eds.), Elsevier Science Publishers B. V. (North–Holland), Amsterdam, pp. 645–654.

[33] **B. Fichet** (1984). "Sur une extension de la notion de hierarchie et son équivalence avec certaines matrices de Robinson". *Journées de Statistique, Montpellier.*

[34] **W. González Manteiga** (1988). "Una perspectiva general con nuevos resultados de la aplicación de la estimación no paramétrica a la regresión lineal". *Estadística Española,* **30**, 141–179.

[35] **J. C. Gower** (1966). "Some distance properties of latent root and vector methods in Multivariate Analysis". *Biometrika,* **53**, 315–328.

[36] **J. C. Gower** (1971). "A general coefficient of similarity and some of its properties". *Biometrics,* **27**, 857–874.

[37] **J. C. Gower** and **C. F. Banfield** (1975). "Goodness–of–fit criteria for hierarchical classification and their empirical distributions in relation with the external variables", en *Proc. 8th. Inter. Biometric Conference*, 347–361.

[38] **J. A. Hartigan** (1967). "Representation of similarity matrices by trees". *J. Am. Stat. Assoc.,* **62**, 1140–1158.

[39] **F. R. Hodson, D. G. Kendall** and **P. Tautu (eds.)** (1971). *Mathematics in the Archaeological and Historical Sciences.* Edinburgh University Press.

[40] **E. W. Holman** (1972). "The relation between hierarchical and Euclidean models for psychological distances". *Psychometrika,* **37**, 417–423.

[41] **H. Hotelling** (1931). "The generalization of Student's ratio". *Annals of Math. Stat.,* **2**, 360–378.

[42] **W. J. Huster, R. Brookmeyer** and **S. G. Self** (1989). "Modelling paired survival data with covariates". *Biometrika,* **45**, 145–156.

[43] **C. J. Jardine, N. Jardine** and **R. Sibson** (1967). "The structure and construction of taxonomic hierarchies". *Math. Biosci.,* **1**, 173–179.

[44] **N. Jardine** and **R. Sibson** (1968). "The construction of hierarchic and nonhierarchic classifications". *Comput. J.,* **11**, 177–184.

[45] **U. Jensen** (1993). *Derivation, calculation and economical application of Rao distance (in german).* Josef Eul Verlag, Köln.

[46] **S. C. Johnson** (1967). "Hierarchical clustering schemes". *Psychometrika,* **32**, 241–254.

[47] **S. Joly** and **G. Le Calve (eds.)** (1992). *Distancia'92.* Université de Rennes.

[48] **J. T. Kent** (1982). "Robust properties of likelihood ratio tests". *Biometrika,* **69**, 19–27.

[49] **M. Knowles** and **D. Siegmund** (1989). "On Hotelling's approach to testing for a nonlinear parameter in regression". *Int. Statist. Rev.,* **57**, 205–220.

[50] **W. J. Krzanowski** (1983). "Distance between populations using mixed continuous and categorical variables". *Biometrika,* **79**, 235–243.

[51] **W. J. Krzanowski** (1987). "A comparison between two distance–based discriminant principles". *J. of Classification,* **4**, 73–84.

[52] **P. C. Mahalanobis** (1936). "On the generalized distance in Statistics". *Proc. Nat. Inst. Sci. India,* **2**, 49–55.

[53] **K.V. Mardia, J. T. Kent** and **J. M. Bibby** (1979). *Multivariate Analysis.* Academic Press.

[54] **K. Matusita** (1955). "Decision rules based on the distance for problems of fit, two samples and estimation". *Ann. Math. Stat.,* **26,** 631–640.

[55] **K. Matusita** (1964). "Distance and decision rule". *Ann. Inst. Stat. Math.,* **16,** 305–315.

[56] **A. Miñarro** and **J. M. Oller** (1992). "Some remarks on the individuals–score distance and its applications to Statistical Inference". *Qüestió,* **16,** 43–57.

[57] **A. F. S. Mitchell** (1992). "Estimative and predictive distances". *Test,* **1,** 105–121.

[58] **D. F. Morrison** (1976). *Multivariate Statistical Methods, 2nd edition.* McGraw–Hill, New York.

[59] **J. Neyman** and **E. S. Pearson** (1928). "On the use and interpretation of certain test criteria for purposes of statistical inference". *Biometrika,* **20A,** 175–240, 263–294.

[60] **F. Oliva, C. Bolance, L. Diaz** and **R. Serrano** (1993). "Aplicació de l'Anàlisi Multivariant a un estudi sobre les llengües europees". *Qüestió,* **17(1),** 139–161.

[61] **J. M. Oller** and **C. M. Cuadras** (1982). "Defined distances for some probability distributions", en *Proc. 2nd World Conf. Math. at the Serv. of Man,* pp. 563–565.

[62] **J. M. Oller** and **C. M. Cuadras** (1987). "Sobre ciertas condiciones que deben verificar las distancias en espacios probabilísticos", en *Actas XV reunión SEIO,* pp. 503–509.

[63] **J. M. Oller** (1989). "Some geometrical aspects of Data Analysis and Statistics", en [31, pp. 41–58].

[64] **W. C. Parr**. *Minimum distance method,* en Encyclopedia of Statistical Sciences. J. Wiley, N. York.

[65] **B. L. S. Prakasa Rao** (1983). *Non parametric functional estimation.* Academic Press, New York.

[66] **C. R. Rao** (1945). "Information and the accuracy attainable in the estimation of statistical parameters". *Bull. Calcutta Math. Soc.,* **37,** 81–91.

[67] **C. R. Rao** (1947). "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation". *Proc. Camb. Phil. Soc.,* **44,** 50–57.

[68] **C. R. Rao** (1973). *Linear Statistical Inference and its Applications, 2nd edition.* John Wiley & Sons, New York.

[69] **M. Ríos** and **C. M. Cuadras** (1986). "Distancia entre Modelos lineales Normales". *Qüestió,* **10,** 83–92.

[70] **W. S. Robinson** (1951). "A method for chronologically ordering archaeological deposits". *Am. Antiq.*, **16**, 293–301.
[71] **M. M. Royall** (1986). "Model robust confidence intervals using maximum likelihood estimators". *International Statistical Review*, **54**, 221–226.
[72] **S. Sattah** and **A. Tversky** (1977). "Additive similarity trees". *Psychometrika*, **42**, 319–345.
[73] **G. A. F. Seber** (1984). *Multivariate Observations.* John Wiley & Sons.
[74] **W. S. Torgerson** (1958). *Theory and methods of scaling.* John Wiley & Sons.
[75] **H. Weyl** (1939). "On the volume of tubes". *Am. J. of Math.*, **61**, 461–472.
[76] **J. Wolfowitz** (1957). "The Minimum Distance Method". *Ann. Math. Stat.*, **28**, 75–88.

ENGLISH SUMMARY:

APPLYING DISTANCES IN STATISTICS

C.M. Cuadras and J. Fortiana

1. INTRODUCTION

Since its beginning, modern Statistics has depended on Probability Theory, Analysis, Measure Theory and Algebra. But also Geometry, especially the study of properties related to distances, has been of a great importance.

Early use can be traced back to K. Pearson's Chi–square and Student's t tests, where the measure of divergence between *expected* and *observed* are variants of the Mahalanobis distance $(x-y)' \cdot \Sigma^{-1} \cdot (x-y)$ where $x,y \in \mathbb{R}^p$, and Σ is a covariance matrix.

A non–trivial example, due to Hotelling [41] and Weyl [75], is commented, showing a specific hypothesis on a nonlinear regression model in which a geodesic distance should be used instead of the likelihood ratio.

The present paper summarizes the application of distances to Statistics in:

- Point estimation
- Testing hypotheses
- Geometric representation of sets
- Prediction models

2. POINT ESTIMATION

2.1. Linear Models

The neatest and most elegant use of distance arises in the normal linear model $y = X \cdot \beta + e$, where the estimation of regression parameters and variances can be expressed in terms of linear projections and norms. The F–test of $H_0 : \Psi = \Psi_0$, where Ψ is an estimable parametric function, can be expressed in terms of a Mahalanobis–like distance between $\widehat{\Psi}$ and Ψ_0.

2.2. Kullback–Leibler Divergence

This divergence between probability densities p, q with respect to a measure μ, defined as $K(p,q) = \int p \log(p/q)\, d\mu$, plays an important role in estimation. Given a model $\Gamma = \{p(x,\theta),\ \theta \in \Theta\}$, the maximum likelihood estimation $\widehat{\theta}$ of θ verifies that the divergence K between $p(x,\widehat{\theta})$ and $p(x,\theta_0)$ is a minimum, where θ_0 is the true parameter. It is shown that this property holds even when the true density q does not belong to the model, by defining the *true parameter* in this case as that of the density in Γ nearest to q with respect to K.

2.3. The Minimum Distance Method

This method of estimation, proposed by J. Wolfowitz [76], is a useful tool in nonparametric estimation of densities, distributions, regression curves, etc. Given the model $\Gamma = \{F(x,\theta),\ \theta \in \Theta\}$, (where the F's are now distribution functions), the estimation $\widehat{\theta}$ for θ is obtained by minimizing the distance $\delta(G_n, F_\theta)$ between the empirical distribution G_n and distributions in Γ. Kolmogorov and Cramér–von Mises statistics are used as the measure δ of the distance.

3. TEST OF HYPOTHESES

3.1. Mahalanobis Distance

This distance is fundamental for multinormal inference. It appears in the Student's t and Hotelling's T^2 tests on means, in testing general linear hypotheses on a multivariate linear model $Y = X \cdot B + E$, in comparing two linear models, etc. It is shown that the Neyman–Pearson [59] likelihood ratio test Λ is asymptotically equivalent to the Rao [68] criterion based on efficient scores, i.e.

$$-2 \log \Lambda \stackrel{a}{=} V_{\theta^*}{}' \cdot \mathcal{F}_{\theta^*}{}^{-1} \cdot V_{\theta^*},$$

where the right hand of the above expression can again be interpreted as a Mahalanobis distance.

3.2. Matusita Distance

Defined as $\delta^2(F_1, F_2) = \int \left\{ \sqrt{f_1(x)} - \sqrt{f_2(x)} \right\}^2 dx$, it allows us to compare two distribution functions, to make inferences on means and covariances, to test independence of subsets of variables, etc. In the multinormal case, Matusita and other related distances are functions of the Mahalanobis distance.

3.3. Rao Distance

A statistical model is understood as a Riemannian manifold structure, with metric represented by the Fisher information matrix in appropriate coordinates. The Rao distance between two elements in the model is the length of a geodesic joining them. This distance has been computed for many parametric families and has an asymptotic distribution allowing us to make inferences. Furthermore, when the model is univariate elliptic, the test based on this distance is equivalent to the F test [7]. Recently, Mitchell [57] has studied the estimation of Rao distances.

4. REPRESENTING A FINITE SET

Many interesting applications of the distance concept in Statistics appear through geometrical representations of a finite set U, which can be classified

as Euclidean (plotting along ordinations axes), Ultrametric (by a dendrogram), Quadripolar (by an additive tree) and Robinson (by a pyramid).

4.1. Euclidean Representation

Given an $n \times n$ distance matrix Δ defined on U, Theorem 4.1 gives the condition for Δ to be Euclidean and provide an explicit set of optimal coordinates, called Principal Coordinates, allowing U to be represented optimally in reduced dimension. The Euclidean property holds when B is semidefinite positive. When it is not, then we need to introduce imaginary coordinates (Theorem 4.2).

4.2. Ultrametric Representation

An ultrametric distance on U is equivalent to an indexed hierarchy (C, α) of subsets of U. These structures, or their graphical counterpart, dendrograms (Fig. 1), are the basis of Numerical Taxonomy. Propositions 1 (Holman's theorem) and 2 give Euclidean properties of an ultrametric on U. Proposition 3 [24] explains this dimensionality under the perspective of Theorem 4.1, while Proposition 4 (Critchley [11]), shows that it is possible to find a (rather special) one–dimensional representation of a dendrogram.

4.3. Quadripolar Representation

A finite set with a quadripolar distance can be represented by an additive tree (a connected graph with no cycles, where the metric is defined by the length of the axes). One motivation for this representation is the study of evolutionary trees: in this case (Fig. 3), extremes of the tree are contemporary species while the other nodes correspond to common ancestors. A quadripolar distance is in general non Euclidean, but (Proposition 5) the square root transformation yields a Euclidean distance.

4.4. Robinson Representation

Now the motivation is the seriation of archaeological objects, where dimensionality is dominated by time. For a distance matrix Δ to have the Robinson property, distances must increase when moving away from the diagonal along rows or columns. There is a bijection between Robinson distances and pyramids, a kind of graph (Fig. 4) which generalizes dendrograms.

5. DISTANCE BASED PREDICTION

Distances can be used to predict a response variable Y, given a set Ξ of explanatory variables. We present the following cases:

1. Y continuous, Ξ mixed variables.
2. Y continuous, Ξ continuous, nonlinear relationship.
3. Y discrete with g states, Ξ mixed (Discriminant analysis).

5.1. Prediction with Mixed Variables

Based on model (8), where X_k is a suitable subset of columns of X, the principal coordinate solution obtained from $\boldsymbol{\Delta}$ (Theorem 4.1). The distance matrix $\boldsymbol{\Delta}$ has been found by defining dissimilarities between observations on the basis of the mixed set Ξ of variables. A good choice is Gower's coefficient [36]. This method, proposed by Cuadras and Arenas [17, 21], generalizes classical regression and reduces to it when the Euclidean distance is used.

5.2. Nonlinear Prediction

Model (8) also performs well for prediction when Y is related to Ξ by a nonlinear function. It is only necessary to use distance $\delta_{ij}^2 = \sum_{h=1}^{p} |\xi_{ih} - \xi_{jh}|$. Cuadras and Fortiana [23] prove the equivalence of this model to an orthogonal polynomial regression for one–dimensional Ξ.

5.3. Discriminant Analysis

Following the same idea, given g populations with g distance matrices $\boldsymbol{\Delta}_k$, $k = 1, \ldots, g$, (10) gives a discriminant function and **[DB]** provides an allocation rule. This distance–based method has good properties [20].

5.4. Prediction when populations are known

The population version of (10) is given in (11), where the discriminant functions depend on the expected value of the squared distances between observations. The allocation rule is still **[DB]**. This rule reduces to the linear discriminant when the Mahalanobis distance is used, is equivalent to the ML rule for

multinomial data, is additive and provides results similar to those based on the Bayes rule when prior probabilities are known.

The population version of the regression model (8) is obtained [23] by finding a continuous version of Principal Coordinate Analysis with respect to distance $d(u,v) = \sqrt{|u-v|}$ $u, v \in (0,1)$ for a uniform $(0,1)$ distribution. This solution can be used in prediction and generalized for any continuous random variable.