# Qüestiió

## Quaderns d'Estadística i Investigació Operativa

Any 1993, volum 17, núm. 1

Entitats patrocinadores:

Universitat de Barcelona
Universitat Politècnica de Catalunya
Institut d'Estadística de Catalunya

Generalitat de Catalunya
**Institut d'Estadística de Catalunya**

Sumari

QÜESTIIÓ, Vol 17, 1 pp. 3-32, 1993

# ESTIMACIÓN DE LA FUNCIÓN DE DENSIDAD CON OBSERVACIONES OBTENIDAS EN INSTANTES ALEATORIOS

JOSÉ A. VILAR y JUAN M. VILAR*

Departamento de Matemáticas

Universidad de La Coruña

*Sea $X(t)$ un proceso estacionario en tiempo continuo con función de densidad marginal univariante $f(x)$. A partir de un conjunto de n observaciones; $X(\tau_1), X(\tau_2), \ldots, X(\tau_n)$ recogidas en instantes muestrales, $\tau_i$, espaciados irregularmente o aleatorios, se estudia la estimación no paramétrica de $f(x)$, utilizando un estimador recursivo tipo núcleo.*

*Asumiendo condiciones débiles de dependencia ($\alpha-$mixing) se obtiene la expresión del sesgo y varianza del estimador definido, así como propiedades de normalidad asintótica.*

**Estimation of the pobability density from random sampling.**

**Key words:** estimación recursiva de la densidad, muestreo aleatorio, procesos mixing de parámetro continuo.

**Clasificación A.M.S. (1991):** 62G07.

---

*Juan Manuel Vilar Fernández. Facultad de Informática. Campus da Zapateira s/n. LA CORUÑA, 15071.

–Article rebut el març de 1992.

–Acceptat el desembre de 1992.

## 1. INTRODUCCIÓN. DEFINICIONES

En los estudios relativos a la estimación de curvas notables (densidad, distribución, regresión...) asociadas a un proceso estacionario real en tiempo continuo, $X(t)$, a partir de una muestra $X(\tau_1), X(\tau_2), \ldots, X(\tau_n)$, generalmente se supone que los instantes muestrales se han tomado equiespaciados. En muchos casos, este planteamiento no es correcto, ya que por distintas causas: instrumentos de medición sujetos a un margen de error, imposibilidad física de muestreo periódico, diseño del experimento, etc., los instantes de observación son aleatorios, lo que va a influir en los resultados de la estimación, pues éstos se ven afectados por la posible dependencia de los elementos muestrales que, en general, es función de la distancia temporal que hay entre ellos.

En este trabajo se estudian propiedades asintóticas de la estimación no paramétrica, recursiva, tipo núcleo de la función de densidad asociada a un proceso estacionario $X(t)$ en tiempo continuo a partir de un conjunto de observaciones tomadas en instantes aleatorios. Indicando la influencia de la aleatoriedad de las observaciones en el error cuadrático medio de la estimación. Se considerarán dos estructuras aleatorias concretas de recogida de datos que se exponen a continuación:

### 1.1. El Modelo

Sea el modelo estocástico (continuo-discreto) definido por el par $(X, T)$, donde,

* **La primera componente $\boldsymbol{X} = \{X(t): t \in \mathbb{R}\}$** es un proceso estocástico real en tiempo continuo, verificando:

1. Es estrictamente estacionario, con función de densidad $f(x)$ contínua y acotada.

2. Verifica la condición de dependencia "fuertemente mixing", ($\alpha-$mixing), esto es, si denotamos por $F_a^b$ la $\sigma-$álgebra generada por $\{X(t): a \leq t \leq b\}$, con $-\infty < a \leq b < +\infty$, se verifica para $t > 0$,

$$\text{Sup } \left\{|P(AB) - P(A)P(B)|: A \in F_{-\infty}^0,\ B \in F_t^{+\infty}\right\} = \alpha(t) \downarrow 0$$

Esta condición de dependencia fue introducida por Rosenblatt y es de las más débiles, verificándose en múltiples situaciones.

* **La segunda componente** $\boldsymbol{T} = \{\tau_k : k \in \mathbb{N}\}$ es una sucesión real, estrictamente creciente de instantes aleatorios, a la que dotaremos de dos posibles estructuras:

**ESTRUCTURA OI (Observaciones Irregulares)**: el proceso $\boldsymbol{T} = \{\tau_k : k \in \mathbb{N}\}$ es de la forma $\tau_k = k/\beta + Z_k$, con $k = 0, 1, 2, \ldots$, $\beta > 0$, siendo $Z_k$ variables aleatorias independientes e igualmente distribuidas, con función de densidad $g(x)$ simétrica y con soporte $[-1/2\beta,\ 1/2\beta]$.

La estructura OI es la combinación de un factor determinista $(k/\beta)$ y un factor aleatorio $(Z_k)$, siendo ésta la que lo diferencia del muestreo periódico.

**ESTRUCTURA PR (Procesos de Renovación)**: el proceso $\boldsymbol{T} = \{\tau_k : k \in \mathbb{N}\}$ es de la forma $\tau_k = \sum_{i=1}^{k} t_i$, con $k = 1, 2, 3, \ldots$ y $\tau_0 = 0$, siendo la sucesión de tiempos intermedios, $t_i$, variables aleatorias i.i.d., con distribución $G(t)$ absolutamente contínua definida en $[0, \infty)$ y densidad $g(t)$, verificando que $E(t_i) = \frac{1}{\beta} < \infty$. Se denotará $N_t$ al "número de observaciones realizadas en el intervalo $(0, t]$", siendo la "función de renovación" (renewal function): $H(t) = E(N_t)$, y la "densidad de renovación" (renewal density): $h(t) = \mathrm{d}H(t)/\,\mathrm{d}t$. Una interpretación de este modelo se basa en considerar que $\beta$ representa "a largo plazo" el número medio de observaciones realizadas en la unidad de tiempo. De particular interés es el caso en que $G(t)$ es exponencial, entonces PR es un proceso de Poisson de media $\beta$.

Estas estructuras **OI** y **PR** son bastante usuales y marcan una progresiva aleatorización en la recogida de los datos. Han sido utilizadas, entre otros, por Masry (1983) aunque autores como Stoyanov-Robinson (1991) y Blum-Boyles (1981) han utilizado otros modelos.

Además en este trabajo se supone que el proceso en tiempo contínuo, $X(t)$, y el proceso en tiempo discreto, $T$, son independientes.

### 1.2. Definición del estimador

La estimación no paramétrica de la densidad ha sido ampliamente estudiada en los últimos años, tanto en un contexto de datos independientes (Silverman, 1986) como de dependencia (Gyorfi y otros, 1989), por ser un conjunto de técnicas intuitivas y de fácil cálculo que permiten obtener buenos resultados bajo condiciones muy generales. Siendo el estimador más estudiado y utilizado el tipo

núcleo, introducido por Rosenblatt-Parzen y cuya definición es la siguiente:

$$(1) \qquad f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_n(x - X_i), \quad \text{con } K_n(u) = \frac{1}{h_n} K\left(\frac{u}{h_n}\right)$$

con $K(u)$ la función núcleo (generalmente una función de densidad) y $h_n$ el parámetro ventana que indica el nivel de suavización que se introduce en la estimación.

Este estimador ha sido estudiado por Masry (1983) con muestreo aleatorio, aunque bajo este supuesto consideramos de interés la utilización de estimadores recursivos, ya que al ir obteniendo secuencialmente nuevos datos, las estimaciones son más fáciles de actualizar ganando en tiempo de computación y ahorro de memoria. Por ello, en este trabajo, se estudia el siguiente estimador recursivo, tipo núcleo:

$$(2) \qquad \hat{f}_n(x) = \left(\sum_{i=1}^{n} h_i^{\eta}\right)^{-1} \left[\sum_{j=1}^{n} h_j^{\eta} K_j\left(x - X(\tau_j)\right)\right], \quad \eta \in \mathbb{R}$$

que verifica la siguiente relación de recursividad:

$$(3) \qquad \hat{f}_{n+1}(x) = H_{n+1}^{-1}(x)\left[H_n \hat{f}_n(x) + h_{n+1}^{\eta} K_{n+1}\left(x - X(\tau_{n+1})\right)\right], \text{ con } H_n = \left(\sum_{i=1}^{n} h_i^{\eta}\right)$$

El estimador definido, aunque muy general, es un caso particular del introducido por Deheuvels (1974) y estudiado por Wolverton-Wagner (1969) para $\eta = 0$ en el supuesto de independencia y por Masry (1986) para observaciones regulares dependientes. El parámetro $\eta$ influye, de forma inversa, en el sesgo y la varianza del estimador y normalmente se elige en el intervalo [0,1] que es donde se obtienen los mejores resultados (Wertz, 1985).

En el apartado 2, de este trabajo, se obtienen las expresiones del sesgo y varianza del estimador de la densidad definido en (2) obtenido a partir de una muestra $X(\tau_1), X(\tau_2), \ldots, X(\tau_n)$, teniendo los instantes muestrales una estructura OI o PR, e indicando la influencia de la aleatoriedad en la recogida muestral. Como consecuencia se obtiene la Media del Error Cuadrático Integrado, MECI, del estimador y su consistencia en media cuadrática. En el apartado 3 se dan las condiciones para obtener la normalidad asintótica del estimador. Las demostraciones de los resultados obtenidos pueden verse en el Apéndice final.

## 2. SESGO Y VARIANZA

En el resto del trabajo se supone que se verifican las siguientes hipótesis:

**H.1.** La función núcleo, $K(u)$, está acotada, tiene soporte compacto y es de orden $s$ de forma que:

$$\int K(u)\,u^j = C_j = 0 \text{ para } j = 1, \ldots, s-1; \int K(u) = C_0 = 1, \quad \int K(u)\,u^s = C_s < \infty$$

**H.2.** La sucesión de parámetros $h_n$ verifica:

*i)* $h_n \longrightarrow 0$ y $nh_n \longrightarrow \infty$ cuando $n \longrightarrow \infty$

*ii)* $$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{h_i}{h_n}\right)^{j+\eta} = \theta_{j+\eta} < \infty \text{ para } j \leq s+1,\ j \in \mathbb{N} \text{ y } \eta \in [0,1]$$

Estas hipótesis son poco restrictivas, siendo usual utilizar funciones núcleo de orden dos o superior. La primera parte de la hipótesis H1 es clásica en los estudios de estimación no paramétrica, siendo la segunda parte de tipo técnico, verificándose para la elección usual de $h_n = Cn^{-\alpha}$, $\alpha \in (0,1)$, ya que entonces $\theta_j = (1-\alpha j)^{-1}$, para $0 < \alpha j < 1$.

Bajo estas hipótesis se verifica el siguiente Lema que será útil en la demostración de los resultados obtenidos. Su demostración puede verse en Masry (1986).

**Lema 1**

Sea $g(x)$ una función de $L_1$, entonces:

*i)* $$\int_{-\infty}^{+\infty} K_n(x-u)g(u)\,\mathrm{d}u = g(x)$$

*ii)* $$\lim_{n\to\infty} \int_{-\infty}^{+\infty} h_n K_n(x-u)K_n(y-u)g(u)\,\mathrm{d}u = \delta_{x,y}\,g(x)D_k,$$

con $D_k = \int K^2(u)\,\mathrm{d}u$

para todo $x$ punto de continuidad de $g$.

A continuación se obtiene la expresión del sesgo del estimador definido en (1).

**Teorema 1**

Se verifican las hipótesis H1–H2 y además:

**H.3.** La función $f(x)$ es $(s+1)$ veces continuamente diferenciable con derivadas acotadas.

Entonces:

$$\text{Sesgo}\left(\hat{f}_n(x)\right) = \frac{f^{(s)}(x)}{s!} C_s H(\eta)(-h_n)^s + O\left(h_n^{s+1}\right) \tag{4}$$

siendo $H(\eta) = \theta_{\eta+s}/\theta_\eta$.

**Comentarios:**

El sesgo del estimador $\hat{f}_n$ no depende del tipo de instantes muestrales ni del mayor o menor grado de dependencia de los datos. Y sí depende de la función de densidad a estimar, $f(x)$, por medio de sus derivadas, de la función núcleo por la constante $C_s$, del parámetro $\eta \in [0,1]$, a través de la función $H(\eta)$, que para la elección del parámetro de suavización $h_n = Cn^{-\alpha}$, se obtiene que $H(\eta)$ es una función estrictamente creciente y, por tanto, el menor sesgo se obtiene para $\eta = 0$. Y, finalmente, el sesgo depende del parámetro de suvización, cuanto menor sea, menor es el sesgo pero mayor será la varianza como se verá a continuación.

Comparando el estimador recursivo $\hat{f}_n$ con el no recursivo $f_n$ (dado en (1)) se obtiene que $\text{Sesgo}\left(\hat{f}_n(x)\right) = H(\eta)\,\text{Sesgo}\left(f_n(x)\right)$, siendo la función $H(\eta)$ acotada inferiormente por 1, por tanto, el sesgo del estimador recursivo es siempre mayor que el del no recursivo (en la mayoría de las situaciones $H(\eta)$ está entre 3 y 4) aunque son del mismo orden ($h_n^s$ bajo las hipótesis del teorema).

A continuación se estudia la covarianza del estimador $\hat{f}_n(x)$ que sí depende de la estructura de los instantes aleatorios en que se obtienen los datos muestrales y de la dependencia entre éstos.

En primer lugar se considera la estructura de OBSERVACIONES IRREGULARES (OI), obteniéndose:

**Teorema 2**

Sea el modelo estocástico $(X, T = \text{OI})$ definido en el apartado 1, se verifican las hipótesis H1–H2′ (H2′ es igual que H2, con $j = 0$ y $\eta \in [-1,1] \subset \mathbb{R}$) y además la sucesión de coeficientes mixing $\alpha(t)$ es tal que:

**H.4.** $\int_0^\infty (\alpha(t))^q < \infty$, para algún $0 < q < 1$

Entonces

$$\left|\text{Cov}\left(\hat{f}_n(x), \hat{f}_n(y)\right)\right| \leq 20\beta \left(\frac{1}{nh_n^{1+q}}\right) (f(x)f(y))^{\frac{1-q}{2}} K_q^{1-q} V(\eta) \int_0^\infty (\alpha(t))^q \, \mathrm{d}t +$$

$$+ \left(\frac{1}{nh_n}\right) \delta_{x,y} V(\eta) f(x) D_k + o\left(\frac{1}{nh_n}\right) \tag{5}$$

siendo $K_q = \int K(u)^{2/(1-q)} \, \mathrm{d}u$, $V(\eta) = \theta_{2\eta-1}/\theta_\eta^2$, $D_k = \int K^2(u) \, \mathrm{d}u$

Si H4 se verificase para algún $0 < q < 1/2$ y además,

**H.5.** La densidad conjunta $f(x, y, t)$ asociada al vector aleatorio $(X(\tau), X(\tau + t))$ existe y está acotada sobre el plano XY uniformemente cuando $|t| > 1/\beta$.

Entonces:

$$\lim_{n\to\infty} nh_n \left|\text{Cov}\left(\hat{f}_n(x), \hat{f}_n(y)\right)\right| \leq V(\eta) \left(f(x)\delta_{x,y} + [f(x)f(y)]^{1/2}\right) D_k \tag{6}$$

(Esta expresión es válida para toda densidad $g(x)$ asociada a $Z_k$).

Finalmente, si además se tiene:

**H.6.** $p(x, y) := \int_0^\beta \gamma(t - \beta^{-1}) f(x, y, t) \, \mathrm{d}t \leq M < \infty$, siendo $\gamma(t)$ la función de densidad asociada a la variable $(Z_i - Z_j)$.

Entonces:

$$\lim_{n\to\infty} nh_n \left|\text{Cov}\left(\hat{f}_n(x), \hat{f}_n(y)\right)\right| = V(\eta) f(x) \delta_{x,y} D_k \tag{7}$$

**Comentarios:**

1. En la cota para la covarianza del estimador $\hat{f}_n$ (expresión (5)) puede verse la influencia del muestreo aleatorio, ya que ésta es directamete proporcional a $\beta$. Como $1/\beta$ representa el promedio de tiempo transcurrido entre dos instantes muestrales consecutivos $\tau_i$ y $\tau_{i+1}$, para grandes valores $\beta$ habrá una fuerte dependencia entre $X(\tau_i)$ y $X(\tau_{i+1})$, por lo que será mayor la covarianza del estimador.

2. El orden de la cota de la covarianza es $(nh_n^{1+q})^{-1}$, mayor que en el supuesto de independencia que es $(nh_n)^{-1}$. Se puede conseguir que la cota sea de este orden exigiendo hipótesis más severas, como se indica en las expresiones (6) y (7), o bien, bajo condiciones de dependencia más restrictivas, como la de ser uniformemente mixing ($\varphi-$mixing) con coeficientes, $\varphi(t)$, verificando $\int_0^{+\infty} \varphi(t)^{1/2}\, \mathrm{d}t < \infty$.

3. De las expresiones (5), (6) y (7) se deduce de forma directa una cota y expresiones asintóticas de la Varianza de $\hat{f}_n(x)$, en distintos supuestos con sólo hacer $x = y$.

4. Bajo la hipótesis H5 se anula la influencia del valor de $\beta$ y de los coeficientes mixing, ello es debido a que esta hipótesis es rigurosa para grandes valores de $\beta$ puesto que $f(x, y, t)$ es divergente cuando $t \longrightarrow 0$. Pero, aún en este supuesto, $B_n(x, y) = V(\eta)[f(x)f(y)]^{1/2} D_k (nh_n)^{-1}$, que aparece en (6), recoge la influencia del muestreo aleatorio a través ahora de la densidad $g(x)$, obteniendo la misma expresión en el término principal (de orden $(nh_n)^{-1}$) que en el caso de independencia de las observaciones. De esta forma, si $x = y$, la cota para la varianza del estimador $\hat{f}_n(x)$ duplica en términos absolutos a la del mismo estimador cuando la distancia entre los tiempos de muestreo no están sujetos a ningún tipo de irregularidad aleatoria. Recogiéndose la influencia de la dependencia en términos secundarios de menor orden.

5. La hipótesis H6 finalmente consigue eliminar toda la influencia de la densidad $g(x)$ y permite igualar la cota alcanzada por este mismo estimador en el supuesto de independencia y observaciones periódicas.

6. La diferencia en las expresiones 5–6–7 entre utilizar el estimador recursivo, $\hat{f}_n(x)$, o el no recursivo, $f_n(x)$, viene dada por el factor $V(\eta)$, que en el último caso vale 1.

   La función $V(\eta) = \theta_{2\eta-1}/\theta_\eta^2$, con $\eta \in [0, 1]$, para la elección clásica del parámetro de suavización $h_n = Cn^{-\alpha}$, tiene la forma: $\mathrm{V}(\eta) = \dfrac{1 - 2\eta\alpha + \eta^2\alpha^2}{1 - 2\eta\alpha + \alpha}$

que es una función decreciente y acotada superiormente por 1. Por tanto, cuanto mayor es $\eta$, mayor es el sesgo pero menor la varianza del estimador recursivo, siendo para todo $\eta \in [0,1]$ menor la varianza del estimador recursivo que la del no recursivo.

Es conocido que una de las medidas de bondad de ajuste más utilizadas en la estimación no paramétrica de curvas de probabilidad es el Error Cuadrático Medio Integrado (MECI) (Hardle-Marron, 1986) que para la función de densidad viene definido como sigue:

$$\begin{aligned}\text{MECI}\left(\hat{f}_n\right) &= E\left\{\int \left(\hat{f}_n(x) - f(x)\right)^2 W(x)\, \mathrm{d}x\right\} = \\ &= E\left\{\int \left[\text{Sesgo}\left(\hat{f}_n(x)\right)\right]^2 W(x)\, \mathrm{d}x\right\} + \\ &+ E\left\{\int \text{Varianza}\left(\hat{f}_n(x)\right) W(x)\, \mathrm{d}x\right\}\end{aligned}$$

siendo $W(x)$ una función peso.

De esta expresión y de los teoremas 1 y 2 se sigue de forma inmediata el siguiente resultado:

**Corolario 1**

"Bajo las hipótesis H1–H5 se obtiene que:

$$\begin{aligned}\text{MECI}\left(\hat{f}_n\right) &= \frac{h_n^{2s} C_s^2 H(\eta)^2}{s!^2} \int \left(f^{(s)}(x)\right)^2 W(x)\, \mathrm{d}x + \\ &+ \frac{2}{nh_n} K_2 V(\eta) \int f(x) W(x)\, \mathrm{d}x + o\left(h_n^{2s} + \frac{1}{nh_n}\right)\text{"}\end{aligned} \tag{8}$$

De la expresión (8) se deduce que el elemento de mayor influencia en el MECI es el parámetro de suavización, que actúa de balanza entre el sesgo y la varianza. Un efecto análogo tiene el parámetro $\eta$, pero con mucho menor peso, y como puede verse en Wertz (1985) el valor $\eta = 0$ minimiza el MECI, lo que hace que sea el más utilizado. En este caso el MECI del estimador recursivo es mayor que el del no recursivo, aunque no en mucha proporción, por lo que es recomendable la utilización del estimador recursivo en la situación en estudio.

Para la elección $h_n = Cn^{-\alpha}$ se obtiene que el $\alpha$ que minimiza el MECI es $\alpha = 1/(1+2s)$, siendo el MECI $= O\left(n^{-2s/(1+2s)}\right)$, resultado igual al obtenido para el caso de datos independientes (Silverman, 1986).

Sobre la covarianza del estimador $\hat{f}_n(x)$ cuando se considera la estructura de PROCESOS DE RENOVACIÓN (PR), se ha obtenido:

**Teorema 3**

Sea el modelo estocástico $(X, T = \text{PR})$ definido en el apartado 1, donde se ha llamado $h(t)$ a la "densidad renewal" que se supone acotada en $[0, \infty)$. Si se verifican las hipótesis H1–H2′ y H4, se obtiene:

$$\left|\text{Cov}\left(\hat{f}_n(x), \hat{f}_n(y)\right)\right| \leq 20\left(\frac{1}{nh_n^{1+q}}\right)(f(x)f(y))^{\frac{1-q}{2}} K_q^{1-q} V(\eta) \int_0^\infty (\alpha(t))^q \, h(t) \, dt$$

$$+\left(\frac{1}{nh_n}\right)\delta_{x,y} V(\eta) f(x) K_2 + o\left(\frac{1}{nh_n}\right) \tag{9}$$

Si además se verifica:

**H.7.** $\int_0^\infty (1+t)\,(\alpha(t))^q < \infty$, para algún $0 < q < 1$ (más estricta que H4)

**H.8.** $p(u, v, s) := \int_0^\infty f(u, v, t+s) g(t) \, dt \leq D < \infty$, para $u, v \in \mathbb{R}, s \geq 0$

Entonces

$$\lim_{n\to\infty} nh_n \left|\text{Cov}\left(\hat{f}_n(x), \hat{f}_n(y)\right)\right| = V(\eta) f(x) \delta_{x,y} D_k \tag{10}$$

**Comentarios:**

1. Se observa que en la cota de la covarianza obtenida en (9) influye el muestreo aleatorio por la densidad $h(t)$, asociada al modelo PR, y la dependencia de las observaciones ($\alpha$–mixing) por el factor $\int \alpha(t)^q h(t) \, dt$. Además en el supuesto de un proceso de Poisson se obtiene que $h(t) = \beta$, para todo $t$, y, por tanto, la cota dada en (9) es igual a la obtenida en (5) bajo la hipótesis de observaciones irregulares.

2. Nuevamente se ha obtenido una cota de la covarianza de orden $\left(nh_n^{1+q}\right)^{-1}$, aunque bajo hipótesis adicionales H7–H8 se obtiene que la expresión de la covarianza tiene el mismo término principal que en el caso de independencia, que es del orden $(nh_n)^{-1}$. Influyendo el tipo de muestreo y la dependencia en términos secundarios de menor orden.

3. Al igual que en el teorema 2, la diferencia entre utilizar el estimador recursivo $\hat{f}_n(x)$ y el no recursivo $f_n(x)$ viene dada por el factor $V(\eta)$, siendo válidos los comentarios allí realizados.

4. Haciendo $x = y$ de las expresiones (9) y (10) se deduce la cota y forma de la varianza de $\hat{f}_n(x)$ cuando el muestreo es del tipo PR, lo que permite deducir el siguiente corolario:

**Corolario 2**

"Bajo las hipótesis de los Teoremas 1 y 3 se sigue que el MECI de $\hat{f}_n$ es el dado en (8)".

## 3. NORMALIDAD ASINTÓTICA

En este apartado se estudia el problema de encontrar sucesiones de números reales $\{a_n\}$, con $a_n \uparrow \infty$ cuando $n \uparrow \infty$, tales que $\{a(\hat{f}_n(x) - f(x))\}$ tenga distribución asintótica, que en nuestro caso será gaussiana. Esto, intuitivamente, parece posible tras haber comprobado que bajo hipótesis relativamente suaves se obtiene la convergencia en media cuadrática de $\hat{f}_n(x)$ a $f(x)$. Además, este tipo de resultados son interesantes ya que permiten calcular intervalos de confianza asintóticos en torno a $f(x)$ utilizando solamente la información que proporciona la muestra. El resultado que se ha obtenido es el siguiente:

**Teorema 4**

"Si se verifican las hipótesis del Teorema 2 o del Teorema 3 y las hipótesis adicionales:

**H.9.** Para algún $\gamma, 0 < \gamma < 1, nh_n^{3-2\gamma+4\eta} \longrightarrow \infty$

**H.10.** Existe una sucesión de enteros $\{q_n\} \uparrow \infty$ tal que:

*i)* $q_n = o\left(nh_n^{3-2\gamma+4\eta}\right)^{1/2}$

*ii)* $\left(\frac{n}{h_n}\right)^{1/2} \sum_{k=q_n, k\in\mathbb{Z}} (\alpha(k/\beta))^{1-\gamma} \longrightarrow 0$

Entonces:

$$(nh_n)^{1/2}\left\{\hat{f}_n(x) - \mathrm{E}\,\hat{f}_n(x)\right\} \xrightarrow{d} \mathrm{N}(0, \sigma_f) \tag{11}$$

siendo $\sigma_f^2 = \mathrm{V}(\eta) f(x) D_k$ (expresión obtenida en (7) y (10) para $x = y$)."

Teniendo en cuenta que $(nh_n)^{1/2}\left\{\hat{f}_n(x) - f(x)\right\}$ se puede descomponer como la suma de: $(nh_n)^{1/2}\left\{\hat{f}_n(x) - \mathrm{E}\,\hat{f}_n(x)\right\} + (nh_n)^{1/2}\left\{\mathrm{E}\,\hat{f}_n(x) - f(x)\right\}$, y, por tanto de los teoremas 1 y 4 se deduce de forma inmediata el siguiente corolario:

**Corolario 3**

"Si se verifican las hipótesis de los teoremas 1 y 4. Y la hipótesis:

**H.11.** $nh_n^{1+2s} \downarrow 0$

Entonces:

$$(nh_n)^{1/2}\left\{\hat{f}_n(x) - f(x)\right\} \xrightarrow{d} \mathrm{N}(0, \sigma_f) \tag{12}$$

**Comentarios:**

1. En la demostración del Teorema 4 se utiliza el denominado "método Bernstein" (ver Peligrad M., 1985) que consiste en descomponer la suma de las variables aleatorias que definen el estimador en sumas de grandes bloques separados por bloques más pequeños, probándose a continuación que la aportación de los bloques pequeños es asintóticamente nula, mientras que los grandes bloques tienden a ser independientes lo que permite aplicar el Teorema Central del Límite de Lindenberg-Feller para variables aleatorias independientes. Este procedimiento ha sido ampliamente utilizado entre otros autores por Rosenblatt (1970), Robinson (1983) ó Masry (1986).

2. Las hipótesis H9 y H10 se pueden debilitar a costa de exigir condiciones de dependencia más restrictivas (imponiendo velocidades de convergencia de tipo exponencial a los coeficientes $\alpha$−mixing), lo que además permite demostrar de forma más sencilla el teorema 4, utilizando un teorema central de Bradley (1981) para disposiciones triangulares de variables aleatorias fuertemente mixing.

## APÉNDICE. (Demostración de los Teoremas)

### Demostración del TEOREMA 1

Basta tener en cuenta que $\mathrm{E}\left(\hat{f}_n(x)\right) = H_n^{-1}\sum_{j=1}^{n} h_j^{\eta}\left[\int_{-\infty}^{+\infty}\frac{1}{h_j}K\left(\frac{x-u}{h_j}\right)f(u)\,\mathrm{d}u\right]$ haciendo el cambio de variable $u = x - vh_j$ y un desarrollo de Taylor de orden $s$ en $f(u)$ se obtiene que:

$$\mathrm{E}\left(\hat{f}_n(x)\right) = H_n^{-1}\sum_{j=1}^{n} h_j^{\eta}\left[\sum_{r=0}^{s}\frac{f^{(r)}(x)}{r!}C_r(-h_j)^r\right] + O\left(h_j^{s+1}\right)$$

De la aplicación del Lema de Toeplitz, el lema 1 y las hipótesis del teorema se sigue la conclusión de éste.

■

Las demostraciones de los Teoremas 2, 3 y 4 se han desarrollado siguiendo razonamientos análogos a los realizados por Masry (1986), quién, para datos muestrales observados de forma regular y en un contexto de dependencia, estudia el propio estimador $\hat{f}_n(x)$ para $\eta = 0$ y otro estimador dado por:

$$\tilde{f}_n(x) = \frac{1}{nh_n^{1/2}}\left[\sum_{j=1}^{n} h_j^{1/2}K_j\left(x - X_j\right)\right] .$$

Acerca de éste último nótese que el estimador definido en (2), $\hat{f}_n(x)$, con $\eta = \frac{1}{2}$ es asintóticamente equivalente a una versión reescalada de $\tilde{f}_n(x)$. En efecto, por el apartado (ii) de la hipótesis H2 relativa a la sucesión de parámetros de suavización $h_n$, haciendo $\eta = 1/2$ y $j = 0$, se tiene la existencia de:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_i}{h_n}\right)^{1/2} = \theta_{1/2} < \infty$$

Por tanto y en virtud del Lema de Toeplitz:

$$\hat{f}_n(x) \cong \theta_{1/2}^{-1}\tilde{f}_n(x)$$

En consecuencia la varianza asintótica de $\tilde{f}_n(x)$ es del orden de $\theta_{1/2}^2$ por la varianza asintótica de $\hat{f}_n(x)$, con $\eta = 1/2$, y, en este sentido, el estimador definido en (2), $\hat{f}_n(x)$, generaliza también el estudiado por Masry $\left(\tilde{f}_n(x)\right)$.

A continuación se exponen, de forma esquemática, las citadas demostraciones, en las que será útil el siguiente resultado:

**Lema 2 (Doe, 1973)**

"Sea $\{X(t): t \in \mathbb{R}\}$ un proceso fuertemente mixing y sean $\xi$ y $\eta$ dos variables aleatorias medibles respecto a las $\sigma-$álgebras $F_{-\infty}^{0}$ y $F_{t}^{+\infty}$, respectivamente, con $t \geq 0$ arbitrario. Si para algún $\mu > 0$ se verifica: $\mathrm{E}\,|\xi|^{2+\mu} < \infty$ y $\mathrm{E}\,|\eta|^{2+\mu} < \infty$, entonces:

$$|\mathrm{Cov}\,(\xi, \eta)| \leq 10\,(\alpha(t))^{\mu/(2+\mu)}\left(\mathrm{E}\,|\xi|^{2+\mu}\,\mathrm{E}\,|\eta|^{2+\mu}\right)^{1/(2+\mu)}\text{,,}$$

**Demostración del TEOREMA 2**

$$\mathrm{Cov}\left(\hat{f}_n(x), \hat{f}_n(y)\right) = T_{n,0}(x,y) + \mathcal{R}_n(x,y) \tag{13}$$

siendo

$$\mathcal{R}_n(x,y) = \sum_{j=1}^{n-1}\left(T_{n,j}(x,y) + T_{n,-j}(x,y)\right) \tag{14}$$

$$T_{n,j}(x,y) = \left(\sum_{k=1}^{n} h_k^{\eta}\right)^{-2} \sum_{i=1}^{n-|j|} h_{i+|j|}^{\eta} h_i^{\eta}\,\mathrm{Cov}\left\{K_{i+|j|}\left(x - X\left(\tau_{|j|}\right)\right) K_i\left(y - X(0)\right)\right\}$$

en virtud de la estacionariedad de $X(t)$.

De la hipótesis H2' y el Lema 1 se sigue que:

$$T_{n,0}(x,y) = \frac{1}{nh_n} V(\eta) f(x) \delta_{x,y} D_k + \frac{1}{n} \tag{15}$$

Dado que $T = OI$, $\tau_j = \dfrac{j}{\beta} + z_j$, con $\{z_j\}$ v. a. independientes e idénticamente distribuidas según la densidad común $g$ de soporte en $\left[-\dfrac{1}{2\beta}, \dfrac{1}{2\beta}\right]$. Por tanto, para $i > j$, $\tau_i - \tau_j = \dfrac{i-j}{\beta} + (z_i - z_j) \Rightarrow$ la densidad asociada a $\tau_i - \tau_j$, para $i > j$ es $\gamma\left(t - \dfrac{i-j}{\beta}\right)$, siendo:

$$\gamma(t) = (g * g)(t) = \int_{-1/\beta}^{1/\beta} g(t-v) g(v)\,\mathrm{d}v$$

Nótese que las condiciones impuestas a $g$ conducen a que $\gamma$ sea simétrica y tenga soporte $[-1/\beta, 1/\beta]$.

Entonces aproximando $T_{n,j}(x,y)$, para $j \neq 0$, por su valor esperado respecto a la v. a. $\tau_j$ se obtiene que:

$$T_{n,j}(x,y) = \left(\sum_{k=1}^{n} h_k^{\eta}\right)^{-2} \sum_{i=1}^{n-j} h_{i+j}^{\eta} h_i^{\eta} \int_{-1/\beta}^{1/\beta} \operatorname{Cov}\left\{K_{i+j}\left(x - X\left(t + \frac{j}{\beta}\right)\right),\right.$$

$$\left. K_i\left(y - X(0)\right)\right\} \gamma(t)\, \mathrm{d}t \tag{16}$$

Sean $\xi = K_{i+j}\left(x - X\left(t + \frac{j}{\beta}\right)\right)$, $\eta = K_i\left(y - X(0)\right)$ y $\mu = \dfrac{2q}{1-q}$.

Se tiene que:

$$\mathrm{E}\,|\xi|^{2+\mu} = \frac{1}{h_{i+j}}\, q_{i+j}(x) \quad \text{y} \quad \mathrm{E}\,|\xi|^{2+\mu} = \frac{1}{h_i}\, q_i(y)$$

siendo

$$q_i(z) = \int \frac{1}{h_i} \left|K\left(\frac{z-u}{h_i}\right)\right|^{2+\mu} f(u)\, \mathrm{d}u \xrightarrow[i \longrightarrow \infty]{} f(z) \int |K(u)|^{2+\mu}\, \mathrm{d}u \tag{17}$$

en virtud del lema 1 y de la hipótesis de continuidad establecida sobre $f$.

Por aplicación del Lema de Doe en (16):

$$|T_{n,j}(x,y)| \leq 10\left(\alpha\left(t + \frac{j}{\beta}\right)\right)^{q} \left(\sum_{j=1}^{n} h_j^{\eta}\right)^{-2} \sum_{i=1}^{n-j} h_{i+j}^{\eta} h_i^{\eta} \left(h_{i+j} h_i\right)^{-\frac{1+q}{2}} \cdot$$

$$\cdot \left(q_{i+j}(x) q_i(y)\right)^{\frac{1-q}{2}} \leq$$

$$\leq 10\left(\alpha\left(t + \frac{j}{\beta}\right)\right)^{q} \left(\sum_{j=1}^{n} h_j^{\eta}\right)^{-2} \sum_{i=1}^{n} h_i^{2\eta-1} h_i^{-q} \left(q_i(x) q_i(y)\right)^{\frac{1-q}{2}} \tag{18}$$

Supliendo (18) en (14) se tiene, en virtud del Lema de Toeplitz y de (17)

$$|\mathcal{R}_n(x,y)| \leq 20 V(\eta) \frac{1}{n h_n^{1+q}} \left(f(x) f(y)\right)^{\frac{1-q}{2}} K_q^{1-q} \sum_{j=1}^{\infty} \left(\alpha\left(\frac{j}{\beta}\right)\right)^{q} \tag{19}$$

Llevando ahora las cotas obtenidas en (15) y (19) a (13) obtendremos por H4 la cota de Cov $\left(\hat{f}_n(x), \hat{f}_n(y)\right)$ dada en (5).

Para obtener las expresiones asintóticas de la covarianza de $\hat{f}_n(x)$ dadas en (6) y (7), se considera una sucesión de enteros positivos $c_n$ tal que $c_n \uparrow \infty$, $h_n c_n \downarrow 0$ y $h_n^{2q} c_n \uparrow \infty$ para algún $0 < q < 1/2$. Se considera entonces la descomposición del término $\mathcal{R}_n(x,y)$ :

$$(20)\quad \mathcal{R}_n(x,y) = (T_1 + T_{-1}) + \sum_{j=2}^{c_n} (T_j + T_{-j}) + \sum_{j=c_n+1}^{n-1} (T_j + T_{-j}) = \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3$$

Se desarrolla cada uno de estos tres sumandos y se obtiene:

$$T_1(x,y) = \left(\sum_{i=1}^{n} h_i^{\eta}\right)^{-2} \sum_{j=1}^{n-1} h_{j+1}^{\eta} h_j^{\eta} \int_{-1/\beta}^{1/\beta} \mathrm{E}\left\{K_{j+1}\left(x - X\left(t + \frac{1}{\beta}\right)\right) \cdot \right.$$

$$\left. \cdot\, K_j\left(y - X(0)\right)\right\} \gamma(t)\,\mathrm{d}t - \left(\sum_{i=1}^{n} h_i^{\eta}\right)^{-2} \sum_{j=1}^{n-1} h_{j+1}^{\eta} h_j^{\eta} \mathrm{E}\left\{K_j\left(y - X(0)\right)\right\}$$

$$(21)\qquad \int_{-1/\beta}^{1/\beta} \mathrm{E}\left\{K_{j+1}\left(x - X\left(t + \frac{1}{\beta}\right)\right)\right\} \gamma(t)\,\mathrm{d}t = A_n + B_n$$

Del Lema 1 y el Lema de Toeplitz se sigue que $B_n = O(1/n)$. Y de la desigualdad de Cauchy-Schwartz y la hipótesis H5 se sigue:

$$A_n \leq \left(\sum_{i=1}^{n} h_i^{\eta}\right)^{-2} \sum_{j=1}^{n-1} h_{j+1}^{\eta} h_j^{\eta} \left\{\int_0^{1/\beta} \gamma\left(t - \frac{1}{\beta}\right)\,\mathrm{d}t\right\} \frac{1}{h_j^{1/2} h_{j+1}^{1/2}}$$

$$\left\{\int K_{j+1}^2(x-u) f(u) h_{j+1}\,\mathrm{d}u \int K_j^2(y-v) f(v) h_j\,\mathrm{d}v\right\}^{1/2} \quad \Rightarrow$$

$A_n \leq \dfrac{1}{nh_n} V(\eta) \left(f(x)f(y)\right)^{1/2} \dfrac{1}{2} D_k$, y por simetría se obtiene:

$$(22)\qquad \mathcal{S}_1 \leq \frac{1}{nh_n} V(\eta) \left(f(x)f(y)\right)^{1/2} D_k$$

A continuación se prueba que los sumandos $\mathcal{S}_2$ y $\mathcal{S}_3$ tienen cotas de orden inferior a $(nh_n)^{-1}$.

$$\mathcal{S}_2 = 2\sum_{j=2}^{c_n} T_j = 2\left(\sum_{i=1}^{n} h_i^{\eta}\right)^{-2} \sum_{j=2}^{c_n} \sum_{i=1}^{n-j} h_{i+j}^{\eta} h_i^{\eta}$$

$$\int_{-1/\beta}^{1/\beta} \left[ \int\int K_{i+j}(x-u) K_i(y-v) \left( f\left(u, v, t+\frac{j}{\beta}\right) - f(u) f(v) \right) \mathrm{d}u \, \mathrm{d}v \right] \gamma(t) \, \mathrm{d}t$$

Utilizando H5 y el hecho de ser $f$ acotada se obtiene que:

$$\begin{aligned} \mathcal{S}_2 &\leq \text{Cte.} \left( \sum_{i=1}^{n} h_i^{\eta} \right)^{-2} \sum_{j=1}^{c_n} \sum_{i=1}^{n} h_i^{2\eta} \int_{-1/\beta}^{1/\beta} \gamma(t) \, \mathrm{d}t = \\ &= \text{Cte.}\, c_n \left( \sum_{i=1}^{n} h_i^{\eta} \right)^{-2} \sum_{i=1}^{n} h_i^{2\eta} = \\ &= \text{Cte.}\, c_n \frac{\theta_{2\eta}}{\theta_{\eta}^2} \frac{c_n}{n} = O(c_n/n) = o(nh_n)^{-1} \end{aligned} \tag{23}$$

Para acotar el sumando $\mathcal{S}_3$ se utiliza el Lema 2 (Doe), con $\mu = \dfrac{4q}{1-2q}$

$$\begin{aligned} \mathcal{S}_3 &\leq \left( \sum_{j=1}^{n} h_j^{\eta} \right)^{-2} \sum_{i=1}^{n} h_i^{2\eta-1} h_i^{-2q} \left( q_i(x) q_i(y) \right)^{\frac{1-2q}{2}} 10 \sum_{j=c_n+1}^{n-1} \left( \alpha\left( \frac{j-1}{\beta} \right) \right)^{2q} \leq \\ &\leq \frac{1}{nh_n^{2q+1}} V(\eta) \left( f(x) f(y) \right)^{\frac{1-2q}{2}} K_q^{1-2q} \sum_{j=c_n}^{\infty} \left( \alpha\left( \frac{j}{\beta} \right) \right)^{2q} \leq \\ &\leq \frac{1}{nh_n^{2q+1}} V(\eta) \left( f(x) f(y) \right)^{\frac{1-2q}{2}} K_q^{1-2q} \frac{1}{c_n} \beta \int_0^{\infty} \left( \alpha(t) \right)^{2q} \mathrm{d}t \leq \\ (24) &\leq O\left( \frac{1}{nc_n h_n^{1+2q}} \right) = O\left( \frac{1}{nh_n} \frac{1}{h_n^{2q} c_n} \right) = o\frac{1}{nh_n} \end{aligned}$$

La última desigualdad se obtiene utilizando H4. Ahora sustituyendo (22), (23) y (24) en (20) se sigue que: $\mathcal{R}_n(x,y) \leq \dfrac{1}{nh_n} V(\eta) \left( f(x) f(y) \right)^{1/2} D_k$, de esto y la expresión (15) se obtiene el resultado (6) del Teorema.

Si se utiliza la hipótesis H6 para acotar el término $A_n$ se sigue de manera inmediata que $A_n = o\dfrac{1}{nh_n}$, de donde se obtiene la expresión (7).

■

La prueba del Teorema 3 sigue la misma línea que la del Teorema 2. La única modificación importante se debe a la nueva estructura de $T$ (que es ahora la de un proceso de renovación) y afecta a la distribución de probabilidad de la variable $\tau_i - \tau_j$, $i > j$. En este caso y debido a que la sucesión de tiempos intermedios $\{t_i\}$ es independiente e idénticamente distribuida según una distribución $G(t)$ sobre $[0, \infty)$; la distribución de

$$\tau_{i+j} - \tau_i = \sum_{\ell=i+1}^{i+j} t_\ell = \sum_{\ell=1}^{j} t_\ell$$

es la $j-$ésima convolución de $G$ consigo mismo. Si se denota a ésta por $G_j(t)$, al aproximar $T_{n,j}(x, y)$ por su valor esperado respecto a $\tau_j$, con $j > 0$, tal y como se hizo en (16), se obtiene:

$$\mathrm{T_{n,j}(x,y)} = \left(\sum_{\mathrm{k=1}}^{\mathrm{n}} \mathrm{h_k^\eta}\right)^{-2} \sum_{\mathrm{i=1}}^{\mathrm{n-j}} \mathrm{h_{i+j}^\eta h_i^\eta} \int_0^\infty \mathrm{Cov}\ \{\mathrm{K_{i+j}\,(x - X(t), K_i\,(y - X(O))\}\ dG_j(t)}$$

**Demostración del TEOREMA 4**

Se supone que el modelo estocástico $(X, T)$ tiene una estructura $OI$, siendo análoga la demostración para el supuesto de que la estructura fuese PR.

Se desea probar que:

$$\sqrt{nh_n}\frac{\hat{f}_n(x) - \mathrm{E}\,\hat{f}_n(x)}{\sigma_f} \xrightarrow{d} \mathrm{N}(0,1), \quad \text{con } \sigma_f = V(\eta)f(x)D_k \tag{25}$$

Sean

$$Y_j = h_j^\eta\,(K_j\,(x - X(\tau_j)) - \mathrm{E}\,K_j\,(x - X(\tau_j))) \tag{26}$$

$$\mathcal{S}_n = \sum_{j=1}^{n} Y_j = \sum_{j=1}^{n} h_j^\eta\,(K_j\,(x - X(\tau_j)) - \mathrm{E}\,K_j\,(x - X(\tau_j))) \tag{27}$$

Entonces, en virtud de la definición de $\hat{f}_n$ en (2) se tiene que:

$$\mathcal{S}_n = \left(\sum_{j=1}^{n} h_j^\eta\right)\left(\hat{f}_n(x) - \mathrm{E}\,\hat{f}_n(x)\right) \quad \text{con} \quad \mathrm{E}[\mathcal{S}_n] = 0 \tag{28}$$

y dado que el teorema actual verifica las hipótesis del teorema 2 se tiene por (10):

$$\lim_{n\to\infty} nh_n \operatorname{Var}\left(\hat{f}_n(x)\right) = \sigma_f \tag{29}$$

De (27) y (28) se deduce que probar (25) es equivalente a probar:

$$\frac{\mathcal{S}_n}{\sigma(\mathcal{S}_n)} \xrightarrow{d} \mathrm{N}(0,1), \tag{30}$$

Pero de (26), (29) y la hipótesis H2, ii) se tiene también que:

$$\frac{nh_n}{n^2 h_n^{2\eta}}\sigma^2(\mathcal{S}_n) = \left(\frac{1}{n}\sum_{k=1}^{n}\frac{h_k^\eta}{h_n^\eta}\right)^2 \left(nh_n \operatorname{Var}\left(\hat{f}(x)\right)\right) \longrightarrow \theta_\eta^2 \sigma_f^2$$

Por tanto bastará demostrar que:

$$\left(nh_n^{2\eta-1}\right)^{-(1/2)} \mathcal{S}_n \xrightarrow{d} \mathrm{N}\left(0, \theta_\eta \sigma_f\right) \tag{31}$$

Para ello se siguen los siguientes pasos:

PASO 1: DESCOMPOSICIÓN EN BLOQUES DE $\mathcal{S}_n$

Por las hipótesis H9 y H10 existe una sucesión $\{r_n\} \subset \mathbb{N}$, $\{r_n\} \uparrow \infty$ tal que

$$r_n q_n = o\left(nh_n^{3-2\gamma+4\eta}\right)^{1/2} \tag{32}$$

y

$$r_n\left(nh_n^{-1}\right)^{1/2} \sum_{k=q_n}^{\infty} \alpha\left[k/\beta\right]^{1-\gamma} \longrightarrow 0 \tag{33}$$

A partir de $\{r_n\}$ se define una nueva sucesión $\{p_n\} \subset \mathbb{N}$ en la forma:

$$p_n = \left[\frac{(nh_n)^{1/2}}{r_n}\right] \tag{34}$$

donde $[a]$ denota la función parte entera de $a$.

De H9, H10, (32), (33) y (34) se deducen las siguientes relaciones

$$\frac{q_n}{p_n h_n^{2\eta-\gamma+1}} \simeq \frac{q_n r_n}{\left(nh_n^{3-2\gamma+4\eta}\right)^{1/2}} \longrightarrow 0 \quad \left(\Rightarrow \frac{q_n}{p_n} \longrightarrow 0 \quad \text{y} \quad \frac{q_n}{p_n h_n^{2\eta}} \longrightarrow 0\right) \tag{35}$$

$$\frac{p_n}{nh_n^{2\eta-\gamma+1}} \simeq \frac{(nh_n)^{1/2}}{r_n nh_n^{2\eta-\gamma+1}} = \frac{h_n}{r_n\left(nh_n^{3-2\gamma+4\eta}\right)^{1/2}} \longrightarrow 0 \quad \left(\Rightarrow \frac{p_n}{n} \longrightarrow 0\right) \tag{36}$$

$$\frac{1}{h_n^{2\eta-\gamma+1}}\sum_{i=q_n}^{\infty}\alpha\left(\frac{i}{\beta}\right)^{1-\gamma}=\left((nh_n^{-1})^{1/2}\sum_{i=q_n}^{\infty}\alpha\left(\frac{i}{\beta}\right)^{1-\gamma}\right)\frac{h_n}{(nh_n^{3-2\gamma+4\eta})^{1/2}}\longrightarrow 0$$

(37)

$$\frac{n}{p_n}\alpha\left(\frac{q_n}{\beta}\right)\leq\frac{n}{p_n}\sum_{i=q_n}^{\infty}\alpha\left(\frac{i}{\beta}\right)^{1-\gamma}\simeq r_n(nh_n^{-1})^{1/2}\sum_{i=q_n}^{\infty}\alpha\left(\frac{i}{\beta}\right)^{1-\gamma}\longrightarrow 0$$

(38)

A continuación sea, para cada $n\in\mathbb{N}$, la partición de $\mathcal{S}_n$ en $2k_n+1$ subconjuntos donde $k_n=[n/(p_n+q_n)]$ realizada en la siguiente forma:

$$\mathcal{S}_n=\mathcal{S}_n^1+\mathcal{S}_n^2+\mathcal{S}_n^3=\sum_{j=0}^{k-1}\beta_j+\sum_{j=0}^{k-1}\pi_j+\theta_k \tag{39}$$

con

$$\beta_j=\sum_{i=1}^{p}Y_{m_j+i},\ \pi_j=\sum_{i=p+1}^{p+q}Y_{m_j+i},\ \theta_k=\sum_{i=m_k+1}^{n}Y_i \tag{40}$$

siendo $m_j=j(p+q)$, $j=0,\ldots,k$; de tal forma que cada $\beta_j$ constituye un bloque "grande" sumando $p_n$ variables, cada $\pi_j$ uno "pequeño" sumando $q_n$ variables y $\theta_k$ un bloque residual.

PASO 2: CUANDO $n\longrightarrow\infty$ SE PRUEBA QUE:

$$\frac{1}{nh_n^{2\eta-1}}\,\mathrm{E}\left(|\mathcal{S}_n^2|\right)^2\longrightarrow 0\quad\text{y}\quad\frac{1}{nh_n^{2\eta-1}}\,\mathrm{E}\left(|\mathcal{S}_n^3|\right)^2\longrightarrow 0$$

Primeramente nótese que por ser $K$ acotado y verificar H1 también $\int|K(x)|^{2/\gamma}\,\mathrm{d}x<\infty$ para $0<\gamma<1$, y por el Lema 1, apdo. i:

$$\mathrm{E}\left\{\left|k_i(x-X(\tau_i))\right|^{2/\gamma}\right\}=h_i^{-(2/\gamma)+1}\int\left|h_i^{-1}K\left(\frac{x-u}{h_i}\right)\right|^{2/\gamma}f(u)\,\mathrm{d}u=$$
$$=h_i^{-(2/\gamma)+1}\Phi_i(x)$$

con

$$\Phi_i(x)\ \xrightarrow{i}\ f(x)\int|K(u)|^{2/\gamma}\,\mathrm{d}u \tag{41}$$

De forma análoga

$$\text{(42)} \qquad \operatorname{Var} Y_i \leq h_i^{2\eta} \operatorname{E} K_i^2\big(x - X(\tau_i)\big) = h_i^{2\eta-1}\Lambda_i(x)$$

siendo $\Lambda_i(x) := h_i \operatorname{E} K_i^2\big(x - X(\tau_i)\big)$, que en virtud del apartado ii del Lema 1 converge a $f(x)D_k$ cuando $i$ tiende a $\infty$ dado que $f$ es, por hipótesis, contínua.

Entonces:

$$\text{(43)} \qquad \operatorname{E}\left(\left|\mathcal{S}_n^2\right|\right)^2 \leq \sum_{j=0}^{k-1} \operatorname{Var} \pi_j + 2 \sum_{\substack{i=0 \\ i\neq j}}^{k-1} \sum_{j=0}^{k-1} \left|\operatorname{Cov}\left\{\pi_i, \pi_j\right\}\right| = A + B$$

donde

$$\text{(44)} \qquad \operatorname{Var} \pi_j = \sum_{i=1+p}^{p+q} \operatorname{Var} Y_{m_j+i} + \sum_{\substack{r=p+1 \\ r<\ell}}^{p+q} \sum_{\ell=p+1}^{p+q} \left|\operatorname{Cov}\left\{Y_{m_j+r}, Y_{m_j+\ell}\right\}\right|$$

Si $2\eta - 1 > 0$, entonces por (42) y el carácter decreciente de $h_n$ se tiene:

$$\text{(45)} \qquad \frac{1}{h_n^{2\eta-1}} \sum_{i=1+p}^{p+q} \operatorname{Var} Y_{m_j+i} \leq \sum_{i=1+p}^{p+q} \left(\frac{h_1}{h_n}\right)^{2\eta-1} \Lambda_{m_j+i}(x) \leq \frac{q_n}{h_n^{2\eta-1}} \operatorname{Cte.}_1(x)$$

Si $2\eta - 1 < 0$, entonces $\left(\dfrac{h_i}{h_n}\right)^{2\eta-1} < 1$, y la cota sería $q_n \operatorname{Cte.}_2(x)$.

En virtud del Lema de Doe, si $\gamma$ es el del enunciado y denotando por $\|Y_i\|$ a $\|Y_i\| = \left(\operatorname{E}\left\{|Y_i|^{2/\gamma}\right\}\right)^{\gamma/2}$, se tiene:

$$\frac{1}{h_n^{2\eta-1}} \sum_{\substack{r=p+1 \\ r<\ell}}^{p+q} \sum_{\ell=p+1}^{p+q} \left|\operatorname{Cov}\left\{Y_{m_j+r}, Y_{m_j+\ell}\right\}\right| \leq$$

$$\leq \frac{10}{h_n^{2\eta-1}} \sum_{\substack{r=p+1 \\ r<\ell}}^{p+q} \sum_{\ell=p+1}^{p+q} \left\|Y_{m_j+r}\right\| \left\|Y_{m_j+\ell}\right\| \left[\alpha\left(\tau_s - \tau_\ell\right)\right]^{1-\gamma}$$

Pero por (41) y ser $h_n$ decreciente:

$$\|Y_i\| \leq h_i^{\eta} h_1^{-1+\gamma/2} \Phi_i^{\gamma/2}(x) \leq h_1^{\eta} h_n^{-1+\gamma/2} \Phi_i^{\gamma/2}(x) \Rightarrow$$
$$\text{Máx.}\left\{\|Y_i\|; \quad 1 \leq i \leq n\right\} \leq h_n^{-1+\gamma/2} \operatorname{Cte}_3(x)$$

Así pues:

$$\frac{1}{h_n^{2\eta-1}} \sum_{\substack{r=p+1 \\ r<\ell}}^{p+q} \sum_{\ell=p+1}^{p+q} \left|\text{Cov}\left\{Y_{m_j+r}, Y_{m_j+\ell}\right\}\right| \leq$$

$$\leq \text{Cte}_4(x) h_n^{\gamma-1-2\eta} \sum_{\substack{r=p+1 \\ r<\ell}}^{p+q} \sum_{\ell=p+1}^{p+q} \left[\alpha\left(\tau_s - \tau_\ell\right)\right]^{1-\gamma} \leq$$

$$\leq \text{Cte}_4(x) h_n^{\gamma-1-2\eta} \sum_{r=1}^{q-1} (q_n - r) \left[\alpha\left(\frac{r-1}{\beta}\right)\right]^{1-\gamma} \leq$$

$$\leq \text{Cte}_4(x) h_n^{\gamma-1-2\eta} q_n \sum_{r=1}^{\infty} \left[\alpha\left(\frac{r}{\beta}\right)\right]^{1-\gamma}$$

donde se ha de notar que el último factor es necesariamente finito por (38).

Tasladando esta cota y la obtenida en (45) a (44):

$$(46) \quad \frac{1}{nh_n^{2\eta-1}} B = \frac{1}{nh_n^{2\eta-1}} \sum_{j=0}^{k-1} \text{Var}\, \pi_j \leq \text{Cte}_1(x) \frac{k_n q_n}{nh_n^{2\eta-1}} + \text{Cte}_5(x) \frac{k_n q_n}{nh_n^{2\eta-\gamma-1}}$$

Y la convergencia de (46) a cero se sigue de (35) y (36).

De forma similar se procede a acotar $\frac{1}{nh_n^{2\eta-1}} B$. En efecto:

$$B \leq 2 \sum_{\substack{i=0 \\ i>j}}^{k-1} \sum_{j=0}^{k-1} \sum_{s=p+1}^{p+q} \sum_{t=p+1}^{p+q} \left|\text{Cov}\left\{Y_{m_i+s}, Y_{m_j+t}\right\}\right|$$

Como $i > j$, los índices $(m_i + s)$ y $(m_j + t)$ difieren cuando menos en $p$ unidades, por tanto, volviendo a aplicar el Lema de Doe y dado que $\frac{q_n}{p_n} \longrightarrow 0$:

$$(47) \quad \frac{1}{nh_n^{2\eta-1}} B \leq 4 \sum_{s=1}^{n-p} \sum_{t=s+p}^{n} \left|\text{Cov}\left\{Y_s, Y_t\right\}\right| \leq \text{Cte}_6(x) h_n^{\gamma-1-2\eta} \sum_{i=q_n}^{\infty} \alpha\left(\frac{i}{\beta}\right)^{1-\gamma}$$

Y (47) converge a cero por (37), lo que unido a la convergencia a cero de (46) y sustituido en (43) prueba que: $\frac{1}{nh_n^{2\eta-1}} \text{E}\left(\left|\mathcal{S}_n^2\right|\right)^2 \longrightarrow 0$.

Finalmente:

$$\frac{1}{nh_n^{2\eta-1}} \mathrm{E}\left(\left|\mathcal{S}_n^3\right|\right)^2 = \frac{1}{nh_n^{2\eta-1}} \left( \sum_{i=m_k+1}^{n} \mathrm{Var}\, Y_i + 2 \sum_{\substack{\ell=m+1 \\ \ell<s}}^{n} \sum_{s=m_k+1}^{n} \mathrm{Cov}\left\{Y_\ell, y_s\right\} \right)$$

Idénticos argumentos a los establecidos para acotar $A$ y $B$ permiten obtener:

$$\frac{1}{nh_n^{2\eta-1}} \mathrm{E}\left(\left|\mathcal{S}_n^3\right|\right)^2 \leq \frac{\mathrm{Cte}_7(x)}{n} \left\{ \left|\Delta_{k_n}^3\right| + \frac{\left|\Delta_{k_n}^3\right|}{h_n^{2\eta+1-\gamma}} \right\}$$

siendo

$$\begin{aligned} \left|\Delta_{k_n}^3\right| &= n - k_n(p_n + q_n) = n - \left[\frac{n}{p_n + q_n}\right](p_n + q_n) \leq \\ &\leq n - \left(\frac{n}{p_n + q_n} - 1\right)(p_n + q_n) = (p_n + q_n) \end{aligned}$$

por tanto:

$$\frac{1}{nh_n^{2\eta-1}} \mathrm{E}\left(\left|\mathcal{S}_n^3\right|\right)^2 \leq \frac{\mathrm{Cte}_8(x)}{n} \left\{ 1 + \frac{1}{h_n^{2\eta+1-\gamma}} \right\} (p_n + q_n) \to 0 \quad \text{cuando } n \to \infty$$

en virtud de (36).

PASO 3: CUANDO $n \longrightarrow \infty$ SE TIENE QUE:

$$\left| \mathrm{E}\left(e^{iu\mathcal{S}_n^1}\right) - \prod_{j=0}^{k_n-1} \left(\mathrm{E}\, e^{iu\beta_j}\right) \right| \longrightarrow 0 \tag{48}$$

Para demostrar (48), y equivalentemente la independencia asintótica de los bloques grandes, se utiliza el siguiente resultado, que es una extensión del Teorema de Volkonskii-Rozanov (1959): "Sea $V_j = F_j(B_{j_G})$ $j = 1, \ldots, N$, y $|F_j(x)| \leq 1$. Y sea $H$ la $\sigma-$álgebra generada por el proceso de los instantes de muestreo $\{\tau_k : k = 1, 2, \ldots\}$. Y sea $X$ un proceso fuertemente mixing con coeficientes $\alpha(t)$, entonces se verifica:

$$\left| \mathrm{E}\left(\prod_{i=1}^{N} V_i\right) - \left(\prod_{i=1}^{N} \mathrm{E}\, V_i\right) \right| = \left| \mathrm{E}\left\{ \mathrm{E}\left(\prod_{i=1}^{N} V_i | H\right) - \left(\prod_{i=1}^{N} \mathrm{E}\left(V_i | H\right)\right) \right\} \right| \leq$$

$$(49) \leq 4(N-1)\alpha(q_n/\beta)"$$

Haciendo $F_j(x) = e^{-itx}$ en el Lema anterior se obtiene:

$$\left| \mathrm{E}\left(e^{iu\mathcal{S}_n^1}\right) - \prod_{j=0}^{k_n-1} \left(\mathrm{E}\, e^{iu\beta_j}\right) \right| \leq 4(k_n+1)\alpha\left(\frac{q_n}{\beta}\right) \simeq 4\frac{n}{p_n+q_n}\alpha\left(\frac{q_n}{\beta}\right) \simeq$$

$$\simeq 4\frac{n}{p_n}\alpha\left(\frac{q_n}{\beta}\right)$$

Y ahora (38) permite probar (48).

PASO 4: SE PRUEBA:

$$\frac{1}{nh_n^{2\eta-1}} \sum_{j=0}^{k-1} \mathrm{E}\,\beta_j^2 \longrightarrow \theta_\eta^2 \sigma_f^2 \tag{50}$$

Procediendo como en (44) para la $\mathrm{Var}\,\pi_j$ se demuestra que el segundo sumando de:

$$\mathrm{E}\,\beta_j^2 = \sum_{r=1}^{p} \mathrm{Var}\, Y_{m_j+r} + 2 \sum_{\substack{i=1 \\ i<\ell}}^{p} \sum_{\ell=1}^{p} \left|\mathrm{Cov}\,\left\{Y_{m_j+i}, Y_{m_j+\ell}\right\}\right|$$

Por (26), (29) y el Lema 1 (ii):

$$\frac{1}{nh_n^{2\eta-1}} \mathrm{E}\,\left|\mathcal{S}_n^1\right|^2 \longrightarrow \theta_\eta^2 \sigma_f^2$$

de ambos hechos y del paso 2 de la demostración se deduce (50).

PASO 5: $\mathcal{S}_n^1$ VERIFICA LA HIPÓTESIS CLÁSICA DE LINDEBERG-FELLER

Según (48) y (50), los sumandos $\beta_j$ en $\mathcal{S}_n^1$ son asintóticamente independientes y tales que la suma normalizada de sus varianzas es $\theta_\eta^2\sigma_f^2$. Por ello la condición de Lindeberg-Feller para la normalidad asintótica de $\mathcal{S}_n^1$ toma la forma: para $\epsilon > 0$,

$$g_n(\epsilon) = \frac{1}{nh_n^{2\eta-1}\sigma_f^2\theta_\eta^2} \sum_{j=0}^{k-1} \mathrm{E}\left(\beta_j^2 I_{\left\{|\beta_j| \geq \epsilon\sigma_f\theta_\eta\left(nh_n^{2\eta-1}\right)^{1/2}\right\}}\right) \longrightarrow 0 \text{ si } n \to \infty \tag{51}$$

Por el Lema 1, $\mathrm{E}\left(K_i(x - X(\tau_i)\right)$ es convergente y dado que $h_i^\eta$ decrece a cero y $K$ es acotado por hipótesis:

$$|Y_i| \leq \text{Cte.}\left(h_i^{1-\eta} + h_i^{\eta}\right) \leq \text{Cte.}\, h_i^{1-\eta} \leq \text{Cte.}\, h_n^{1-\eta} \quad \forall i \leq n \Rightarrow$$

$$\Rightarrow \text{Máx}\left\{|\beta_j|, \quad 0 \leq j \leq k_n - 1\right\} \leq \text{Cte.}\, p_n h_n^{1-\eta} \tag{52}$$

Por tanto en (51) se tiene:

$$g_n(\epsilon) = \frac{k_n p_n^2 h_n^{2\eta-2}}{n h_n^{2\eta-1} \sigma_f^2 \theta_\eta^2} \max_{0 \leq j \leq k_n - 1} P\left\{|\beta_j| \geq \epsilon \sigma_f \theta_\eta \left(n h_n^{2\eta-1}\right)^{1/2}\right\} =$$

$$g_n(\epsilon) = \text{Cte} \cdot k \left(\frac{p_n}{(n h_n)^{1/2}}\right) \max_{0 \leq j \leq k_n - 1} P\left\{|\beta_j| \geq \epsilon \sigma_f \theta_\eta \left(n h_n^{2\eta-1}\right)^{1/2}\right\}$$

Ahora bien, de (52)

$$\frac{\max_{0 \leq j \leq k_n - 1} |\beta_j|}{\sigma_f \theta_\eta \left(n h_n^{2\eta-1}\right)^{1/2}} \leq \text{Cte.}\, \frac{p_n h_n^{\eta-1}}{\left(n h_n^{2\eta-1}\right)^{1/2}} = \text{Cte.} \frac{p_n}{(n h_n)^{1/2}} = \frac{1}{r_n} \longrightarrow 0$$

Y, en consecuencia, el $\max_{0 \leq j \leq k_n - 1} P\left\{|\beta_j| \geq \epsilon \sigma_f \theta_\eta \left(n h_n^{2\eta-1}\right)^{1/2}\right\} = 0$ a partir de un $n$ suficientemente grande, deduciéndose entonces (51).

Resumiendo:

$$\left.\begin{array}{l} \text{Por (50) y (51), } \left(n h_n^{2\eta-1}\right)^{-(1/2)} \mathcal{S}_n^1 \longrightarrow \text{N}\left(0, \theta_\eta \sigma_f\right) \\ \text{y por el paso 2, } \left(n h_n^{2\eta-1}\right)^{-(1/2)} \left(\mathcal{S}_n^2 + \mathcal{S}_n^3\right) \longrightarrow 0 \text{ en probabilidad} \end{array}\right\} \Rightarrow$$

$$\left(n h_n^{2\eta-1}\right)^{-(1/2)} \mathcal{S}_n \xrightarrow{d} \text{N}(0, \theta_\eta \sigma_f)$$

■

## 4. BIBLIOGRAFÍA


[1] **Blum-Boyles** (1981). "Random sampling from a continous parameter stochastic process". *Analytical Methods in Probability Theory, Lecture Notes in Mathematics,* **86**, Springer-Verlag.

[2] **Bradley, R.** (1981). "Central Limit Theorems under Weak Dependence". *Journal of Multivariate Analysis,* **11**, 1–16.

[3] **Deheuvels** (1974). "Conditions nécessaires et suffisantes de convergence ponctuelle presque súre". *C.R. Acad. Sci. Paris Ser. A,* **278**, 1217–20.

[4] **Doe** (1973). "A note on empirical processes for strong mixing processes". *Ann. Prob.,* **1**, 870–875.

[5] **Gyorfi-Hardle-Sarda-Vieu** (1989). "Nonparametric curve estimation from time series". *Lecture Notes in Statistics,* **60**.

[6] **Marron, J. & Hardle, W.** (1986). "Random aproximations to some measures of accuracy in nonparametric curve estimation". *Journal of Multivariate Analysis,* **20**, 91–113.

[7] **Masry, E.** (1983). "Probability Density Estimation from Sampled Data". *IEEE, vol. IT-29,* **5**, 696–709.

[8] **Masry, E.** (1986). "Recursive Probability Density Estimation for Weakly Dependent Stationary Processes". *IEEE, vol. IT-32,* **2**, 254–267.

[9] **Peligrad, M.** (1985). "Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (A Survey)". En *Dependence in Probability and Statistics.* E. Eberlein, M. Taqqu ed. Birkhauser.

[10] **Rosenblatt** (1970). "Density estimates and Markov Sequences". En *Nonparametric Techniques in Statistical Inference.* Cambridge University Press, 199–210.

[11] **Robinson** (1983). "Nonparametric estimators for time series". *J. Time Series Anal.,* **40**, 185–207.

[12] **Silverman, B.W.** (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.

[13] **Stoyanov-Robinson** (1991). "Semiparametric and nonparametric inference from irregular observations on continuous time stochastic processes". *Nonparametric Functional Estimation and related Topics,* 553–558, Kluwer Academic Publishers.

[14] **Volkonskii-Rozanov** (1959). "Some limit theorems for random functions I". *Theory Prob. Appl.,* **4**, 178–197.

[15] **Wertz** (1985). "Sequential and Recursive Estimators of the Probability Density". *Statistics, 16,* **2**, 277–295.

[16] **Wolverton-Wagner** (1969). "Recursive estimates of probability densit". *IEEE Trans. Systems Sci. Cybernet,* **5**, 246–247.

# ENGLISH SUMMARY:

## ESTIMATION OF THE PROBABILITY DENSITY FROM RANDOM SAMPLING

José A. Vilar and Juan M. Vilar

## 1. INTRODUCTION. DEFINITIONS

Data are not always collected at equally spaced time intervals and there are various causes by which data may be irregularly recorded over time: small random deviations, the own random experimental design...

In this paper we study the asymptotic behaviour of a recursive nonparametric estimate of the probability density function associated with a stationary continuous-time process $X$, where the sampling instants are assumed to be random.

### 1.1. The model

Let the continuous-discrete stochastic model be defined as the pair $(X, T)$ where:

1. The first component $X = \{X(t);\, t \in \mathbb{R}\}$ is a continuous time process, which is assumed to be strictly stationary asn strongly mixing ($\alpha-$mixing), with marginal density function $f(x)$ continuous and bounded.

2. The second component, $T = \{\tau_k;\, k \in \mathbb{N}\}$, consists of a strictly increasing sequence of random times that we confine to two specific structures:

- **"IO" STRUCTURE (Irregularly Observations)** where:

$$\tau_k = \frac{k}{\beta} + z_k \qquad k = 0, \pm 1, \ldots \quad \beta > 0$$

being $z_k's$ i.i.d. random variables with a symmetric density function $g(t)$ supported on $[\frac{-1}{2\beta}, \frac{1}{2\beta}]$.

- **"RP STRUCTURE (Renewal Process)** where:

$$\tau_k = \sum_{i=1}^{k} \quad k = 1, 2, \ldots \quad \text{and} \quad \tau_0 \equiv 0$$

where $\{t_i\}_{i=1}^{+\infty}$ is a sequence of i.i.d. random variables with a common absolutely continuous distribution $G(t)$ on $[0, \infty)$, satisfying: $\mathrm{E}[t_i] = \beta^{-1}$.

Both structures turn to a progressive randomness of the sampling data: the IO structure is the mixture of a periodic sampling and a random deviation, and the RP structure is completely random.

Throughout this paper we also suppose that $X$ and $T$ are independent.

### 1.2. Definition of the estimate

The nonparametric estimation of an unknown probability density function $f(x)$ has received much attention in recent years (see Silverman (1986) for independent observations and Gyorfy e.a. (1989) for dependent observations).

Masry (1983) has studied this same stochastic model for the nonrecursive type kernel estimator of the density function defined by (1):

$$f_n(x) = \frac{1}{n}\sum_{j=1}^{n} K_n(x - X_j), \quad K_n(u) = \frac{1}{h_n} K\left(\frac{u}{h_n}\right)$$

being $K(u)$ the Kernel function and $h_n$ a bandwidth sequence.

But the recursive estimator has a clear advantage over nonrecursive estimator as they can be updated with each additional observation. Thus we will consider the recursive type Kernel estimate defined by (2):

$$\hat{f}_n(x) = H_n^{-1} \sum_{j=1}^{n} K_j(x - X(\tau_j)), \quad \text{with } \eta \in \mathbb{R} \quad \text{and } H_n = \left(\sum_{j=1}^{n} h_j^{\eta}\right)$$

This estimator, which one was introduced by Deheuvels (1974), can be computed recursively by (3). From equally spaced data, Wolverton-Wagner (1969) have used (2) with $\eta = 0$ for independent observations, and this same estimator has been considered by Masry (1986) under the assumption of mixing conditions.

## 2. BIAS AND VARIANCE

Under weak regularity conditions on $K$ and $\{h_n\}$, the theorem 1 related to bias of $\hat{f}_n$ was established.

We remark that the asymptotic bias expression (4) obtained in theorem 1 depends on Kernel function $K$, on the underlying probability density $f$ and their derivatives, and on parameter $\eta$. Specifically, large values of $\eta$ result in estimations with a large bias. But the bias of the estimation does not depend neither on the particular sampling scheme used nor on the dependence structure type.

Finally, comparing this expression to that in nonrecursive case it is seen that the recursive estimate has a larger bias than a nonrecursive one.

In theorems 2 and 3 we establish bounds for the asymptotic covariance of $\hat{f}_n(x)$ when $T \equiv$ IO and $T \equiv$ RP respectively.

If $T \equiv$ IO, expression (5) is obtained under a weak constraint on the mixing coefficients $\{\alpha_t\}$ (see (H4), and it is interesting to remark some observations about this result:

- The sampling scheme (according to IO structure) contributes to an increasing of the covariance, which is proportional to $\beta$.
- The convergence rate of the bound is $O\left(nh_n^{1+q}\right)^{-1}$ compared to $O\left(nh_n\right)^{-1}$ for regular observations.
- But, under the considerably more restrictive assumptions H4 with $0 < q < \frac{1}{2}$, and H5, the new bound for the asymptotic covariance (see(6)) is independent of $\beta$ and the corresponding rate of convergence is now $O\left(nh_n\right)^{-1}$.
- Finally, under condition H6, the contribution of the sampling scheme is asymptotically negligible.
- These results are extensions to the recursive case of results by Masry (1983) for nonrecursive case. In this sense, it is interesting to note that nonrecursive estimate has a larger variance than a recursive one (see remark 6).

The performance of the estimate $\hat{f}_n(x)$ when $T \equiv$ RP is very similar such as it is shown in teorem 3. In this case the influence of random sampling is present through the renewal density function $h(t)$, which is supposed to be bounded (see (9)). Analogous conclusions to those in theorem 2 are now made for theorem 3.

Combining the results of theorem 1 on the bias of the estimator $\hat{f}_n(x)$ and of theorem 2 (for IO structure) or theorem 3 (for RP structure) on the variance, the quadratic-mean consistency and the corresponding rate of $\hat{f}_n(x)$ is immediately obtained in corollaries 1 and 2 respectively.

## 3. ASYMPTOTIC NORMALITY

In this section, asymptotic normality of the recursive estimate defined by (2) is established.

In first place we write:

$$(nh_n)^{\frac{1}{2}} \left(\hat{f}_n(x) - f(x)\right) = (nh_n)^{\frac{1}{2}} \left(\hat{f}_n(x) - \mathrm{E}\,\hat{f}_n(x)\right) + (nh_n)^{\frac{1}{2}} \left(\mathrm{E}\,\hat{f}_n(x) - f(x)\right)$$

Following Pelligrad (1985) *(Bernstein method)* the asymptotic normality of $\hat{f}_n(x)$ centered at $\mathrm{E}\,\hat{f}_n(x)$ is stated in theorem 4 under H9 and H10.

Finally the sketch of the proofs is given in the last section of the paper.

QÜESTIIÓ, Vol 17, 1 pp. 33-38, 1993

# A SIMPLE STATISTIC TO TEST GENERALIZED PALINDROMIC SYMMETRY MODEL IN A 4 × 4 CONTINGENCY TABLE

SADAO TOMIZAWA*

Science University of Tokyo

*For a 4 × 4 contingency table, this note gives a simple statistic to test the goodness-of-fit of the generalized palindromic symetry (GPS) model considered by McCullagh (1978). Also an asymptotic confidence interval for a parameter of interest in the GPS model is given. Two sets of unaided vision data are used as example.*

**Key words:** Confidence interval; Cumulative odds ratio; Delta method; $z$—statistic.

## 1. INTRODUCTION

For the $R \times R$ square contingency table, let $p_{ij}$ denote the probability that an observation will fall in the cell in row $i$ and column $j (i = 1, \ldots, R; j = 1, \ldots, R)$. The generalized palindromic symmetry (GPS) model defined in McCullagh (1978) is given by

*Department of Information Sciences, Faculty of Sciences & Technology, Science University of Tokyo, Noda City, Chiba 278 Japan.
-Article rebut el juliol de 1992.
-Acceptat el novembre de 1992.

$$\begin{cases} F_{ij} &= \exp\{\Delta_i/2\} \cdot \dfrac{\alpha_i}{\alpha_{j-1}}\phi_{ij} \qquad (1 \le i < j \le R), \\ G_{ji} &= \exp\{-\Delta_i/2\} \cdot \dfrac{\alpha_{j-1}}{\alpha_i}\phi_{ji} \quad (1 \le i < j \le R), \\ p_{ii} &= \phi_{ii} \qquad\qquad\qquad\qquad (1 \le i \le R), \end{cases}$$

where $\phi_{ij} = \phi_{ji}$, $\alpha_1 = 1$ and where

$$F_{ij} = \sum_{s=1}^{i}\sum_{t=j}^{R} p_{st} \quad \text{and} \quad G_{ji} = \sum_{s=j}^{R}\sum_{t=1}^{i} p_{st}.$$

Note that the GPS model is defined when $R \ge 4$. A special case of the GPS model obtained by putting $\Delta_1 = \Delta_2 = \cdots = \Delta_{R-1}$ is the original palindromic symmetry model considered in McCullagh (1978). And a further special case of the GPS model obtained by putting $\alpha_1 = \alpha_2 = \cdots = \alpha_{R-1}$ and $\Delta_1 = \Delta_2 = \cdots = \Delta_{R-1}$ is the conditional symmetry model (see McCullagh 1978 and Tomizawa 1989b).

Tomizawa (1989a, b) pointed out that the GPS model can be expressed as

$$\Theta^U_{ij,st} = \Theta^L_{st,ij} \qquad (1 \le i < j < s < t \le R), \tag{1.1}$$

where $\Theta^U_{ij} = F_{is}F_{jt}/(F_{js}F_{it})$ and $\Theta^L_{st,ij} = G_{si}G_{tj}/(G_{sj}G_{ti})$, being the odds ratios based on the $\{F_{ij}\}$ and $\{G_{ji}\}$. From (1.1), the GPS model states that for $1 \le i < j < s < t \le R$, if the odds that an observation is in row $j$ or below rather than in row $i$ or below is $\theta$ times higher when the observation is in column $t$ or above rather than when it is in column $s$ or above, then the odds that the observation is in column $j$ or below rather than in column $i$ or below is identically $\theta$ times higher when the observation is in row $t$ or above rather than when it is in row $s$ or above.

By the way, for testing the goodness-of-fit of the GPS model, the values of likelihood ratio chi-squared statistics $G^2$ and Pearson's chi-squared statistics $\chi^2$ *could not* be calculated *easily* (even when $R = 4$) though the calculation would be possible, because *(i)* the model is not a multiplicative form, and *(ii)* the maximum likelihood estimates (MLEs) of expected frequencies cannot be written as a closed-form expression of the observations (see McCullagh 1978).

The purpose of this note is to give a simple statistic to test the goodness-of-fit of the GPS model when $R = 4$.

## 2. A SIMPLE TEST STATISTIC

Consider a 4×4 contingency table (i.e., $R = 4$). Then, from (1.1), the GPS model can be simply expressed as

$$\psi = 0 \qquad (\text{or } e^{\psi} = 1),$$

where

$$\psi = \log \Theta^{U}_{12,34} - \log \Theta^{L}_{34,12}.$$

Let $n_{ij}$ denote the observed frequency in the cell $(i, j)$ of the 4×4 table $(i = 1, 2, 3, 4; j = 1, 2, 3, 4)$. Assuming that the $\{n_{ij}\}$ result from full multinomial sampling, we shall give a simple statistic to test the GPS model (i.e., $\psi = 0$), using the *delta method* of which descriptions are given by Bishop *et al.* (1975, Sec. 14.6) and Agresti (1984, p. 185, Appendix C). The sample version of $\psi$, i.e., $\hat{\psi}$, is given by $\psi$ with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$, where $\hat{p}_{ij} = n_{ij}/n$ and $n = \Sigma\Sigma n_{ij}$. Using the delta method, $\sqrt{n}(\hat{\psi} - \psi)$ has asymptotically (as $n \to \infty$) a normal distribution with mean zero and variance

$$\begin{aligned} \sigma^2 \;=\; & \frac{1}{F_{14}} - \frac{1}{F_{13}} - \frac{1}{F_{23}} - \frac{1}{F_{24}} + \frac{2F_{14}}{F_{13}F_{24}} + \\ & + \frac{1}{G_{41}} - \frac{1}{G_{31}} - \frac{1}{G_{32}} - \frac{1}{G_{42}} + \frac{2G_{41}}{G_{31}G_{42}}. \end{aligned}$$

Let $\hat{\sigma}^2$ denote $\sigma^2$ with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$. When the GPS model holds true, $z = \sqrt{n}\hat{\psi}/\hat{\sigma}$ has asymptotically (as $n \to \infty$) the standard normal distribution; thus $z^2$ has asymptotically (as $n \to \infty$) a chi-squared distribution with one degree of freedom (df). In addition, the term $\hat{\sigma}/\sqrt{n}$ is an estimated approximate standard error for $\hat{\psi}$, and $\hat{\psi} \pm z_{p/2}\hat{\sigma}/\sqrt{n}$ is an approximate $100(1-p)$ percent confidence interval of $\psi$, where $z_{p/2}$ is the percentage point from the standard normal distribution corresponding to a two-tail probability equal to $p$.

## 3. EXAMPLES

*Example* 1: Table 1 is constructed from the data on unaided distance vision of 7477 women aged 30-39 employed in Royal Ordnance factories in Britain from 1943 to 1946. These data have been analyzed by many statisticians using various

statistical models and methods; see, for example, Stuart (1955), Bishop *et al.* (1975, p. 284), McCullagh (1978), Goodman (1979), Tomizawa (1989a, b).

When the GPS model is applied to the data in Table 1, the value of $z^2$ (given in Section 2) is 5.71 ($P < 0.025$) with 1 df. Therefore this value is significant at 5% level. By the way, by calculating the MLEs of expected frequencies using the $G^2 = 6.18$ (from Tomizawa 1989a) and $\chi^2 = 6.15$ with 1 df. We can see now that the values of these three statistics are close.

**Table 1**

Unaided distance vision of British women; from Stuart (1955).

| Right eye | Left eye grade | | | | |
|---|---|---|---|---|---|
| grade | Highest | Second | Third | Lowest | Total |
| Highest | 1520 | 266 | 124 | 66 | 1976 |
| Second | 234 | 1512 | 432 | 78 | 2256 |
| Third | 117 | 362 | 1772 | 205 | 2456 |
| Lowest | 36 | 82 | 179 | 492 | 789 |
| Total | 1907 | 2222 | 2507 | 841 | 7477 |

Next, for these data, the estimated value of $\psi$ is $\hat{\psi} = -0.350$. The estimated approximate standard error of $\hat{\psi}$ is 0.146, and an approximate 95% confidence interval for $\psi$ is $(-0.636, -0.063)$. [The corresponding confidence interval for $e^{\psi} = \Theta^U_{12,34}/\Theta^L_{34,12}$ is $(e^{-0.636}, e^{-0.063})$, or $(0.529, 0.939)$.] Since this interval for $\psi$ does not contain the value zero, this would indicate that $\Theta^U_{12,34}$ is not equal (rather than equal) to $\Theta^L_{34,12}$.

*Example* 2: Table 2 is constructed from the data on unaided distance vision of 4746 students aged 18 to about 25 including women of about 10% in Faculty of Science and Technology, Science University of Tokyo in Japan examined in April 1982. These data have been analyzed earlier by Tomizawa (1984, 1989a).

When the GPS model is applied to the data in Table 2, the value of $z^2$ is 1.45 ($P > 0.2$) with 1 d.f. Also the values of $G^2$ and $\chi^2$ are both 1.47 ($G^2$ value is taken directly from Tomizawa 1989a). Therefore the values of three statistics are quite close. (See Tomizawa 1989a for the interpretation obtained under the GPS model).

**Table 2**

Unaided distance vision of students in Japan; from Tomizawa (1984).

| Right eye grade | Left eye grade | | | | |
|---|---|---|---|---|---|
| | Highest | Second | Third | Lowest | Total |
| Highest | 1291 | 130 | 40 | 22 | 1483 |
| Second | 149 | 221 | 114 | 23 | 507 |
| Third | 64 | 124 | 660 | 185 | 1033 |
| Lowest | 20 | 25 | 249 | 1429 | 1723 |
| Total | 1524 | 500 | 1063 | 1659 | 4746 |

Next, for these data, the estimated value of $\psi$ is $\hat{\psi} = -0.241$. The estimated approximate standard error of $\hat{\psi}$ is 0.200, and an approximate 95% confidence interval for $\psi$ is $(-0.633, 0.151)$. [The corresponding confidence interval for $e^{\psi}$ is $(e^{-0.633}, e^{0.151})$, or $(0.531, 1.163)$.] Since this interval for $\psi$ contains the value zero, this would indicate that $\Theta^U_{12,34}$ is equal to $\Theta^L_{34,12}$, or even if $\Theta^U_{12,34}$ is not equal to $\Theta^L_{34,12}$, the degree of non-equality between two odds ratios is slight.

## 4. REMARK

As seen in Sections 2 and 3, the $z^2$–statistic (or $z$–statistic) has the advantage such that it can be easily calculated without the use of computer. Therefore, if one wants to check whether the GPS model fits well or poorly for a 4×4 table data, we recommend using the $z^2$–statistic (or $z$–statistic) given in this note.

## 5. REFERENCES

[1] **Agresti, A.** (1984). *Analysis of Ordinal Categorical Data.* John Wiley & Sons, New York.

[2] **Bishop, Y.M.M., Fienberg, S.E.** and **Holland, P.W.** (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass: MIT Press.

[3] **Goodman, L.A.** (1979). "Multiplicative models for square contingency tables with ordered categories". *Biometrika,* **66**, 413–418.

[4] **McCullagh, P.** (1978). "A class of parametric models for the analysis of square contingency tables with ordered categories". *Biometrika,* **65**, 413–418.

[5] **Stuart, A.** (1955). "A test for homogeneity of the marginal distributions in a two-way classification". *Biometrika,* **42**, 412–416.

[6] **Tomizawa, S.** (1984). "Three kinds of decompositions for the conditional symmetry model in a square contingency table". *Journal of the Japan Statistical Society,* **14**, 35–42.

[7] **Tomizawa, S.** (1989a). "Analogy between generalized palindromic symmetry and quasi-odds symmetry models for square contingency tables with ordered categories". *Bulletin of the Biometric Society of Japan,* **10**, 1–9.

[8] **Tomizawa, S.** (1989b). "Decompositions for conditional symmetry model into palindromic symmetry and modified marginal homogeneity models." *The Australian Journal of Statistics,* **31**, 287–296.

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 $\Sigma$ 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 $\theta$ 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 $\Lambda$ 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 $\theta$ no es identificable cuando $\beta = 0$, no es factible aplicar la teoría asintótica sobre la distribución de $\Lambda$, 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 $\beta$ 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 $\Psi$. Nótese que el numerador de $F$ es una distancia tipo Mahalanobis entre $\widehat{\Psi}$ y $\Psi_0$.

### 2.2. Divergencia de Kullback–Leibler

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

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

juega un importante papel en el llamado problema de *la especificación* en inferencia estadística. Supongamos, para concretar, que $\mu$ 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 $\theta_0$ 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 $\theta$ 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 $\theta$. 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 $\theta$? 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* $\theta_0$ del parámetro $\theta$ se puede definir como aquel $\theta_0$ tal que $p(x,\theta_0) \in \Gamma$ es la densidad más próxima a $q$ de acuerdo con la divergencia (2). $\theta_0$ 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 $\theta_0$. Veamos ahora qué ocurre con el estimador ML $\widehat{\theta}$ obtenido considerando el modelo $\Gamma$. 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 $\theta_0$, $Z(x,\theta) = Z(x,\theta_0) + (\theta - \theta_0)\,(\partial Z/\partial\theta)_{\theta_0} + \cdots$, 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 $\theta_0$ 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 $\theta$ 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 $E_q\left(Z(x,\theta_0)\right) = 0$, 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 $\theta_0$, el valor del parámetro más próximo a q respecto la divergencia de Kullback–Leibler.

Ventajas y aplicaciones de la estimación ML del *verdadero valor* $\theta_0$ 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 $\theta$ 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 $\theta$ y tienen propiedades de robustez en el caso de desviaciones locales del modelo. Incluso, si $G \notin \Gamma$, tomando $\delta_C$ 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 $\Gamma$ (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 $\theta$ 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 $\Sigma$ 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 $\mu_0$ 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 $\mu_1$, $\mu_2$, $\Sigma$ ([53, 58]), es decir, a la $T^2$ de Hotelling, que es también proporcional a la estimación de la distancia de Mahalanobis entre $\mu_1$ y $\mu_2$.

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 $\Sigma$. Entonces, el contraste de hipótesis $H_0 : \Psi = \Psi_0$, donde $\Psi_0$ 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 ${\chi^2}_{q-r}$, 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 ${\chi^2}_q$.

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 $\theta$ dentro de $\Theta_0$.

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 $\theta_0$ 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 $\mu$, 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 $\epsilon$ 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 $\Theta$ 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 $\delta_R^2$.

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 $(\beta, \sigma)$, 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, $\Sigma_0$ 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 $\delta_R$ es la longitud de una curva contenida en la subvariedad de dimensión finita $\Theta \subset \mathcal{E}$, mientras que $\delta_M$ 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* $\delta_{ij}$ *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.* $\boldsymbol{\Delta}$ *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* $\boldsymbol{\Delta} = (\delta_{ij})$ *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.*

$\boldsymbol{\Delta}$ *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 $\Lambda$ 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 $Y$:

| | | $\lambda_1$ | $\lambda_2$ | $\cdots$ | $\lambda_p$ | 0 | $\mu_1$ | $\mu_2$ | $\cdots$ | $\mu_q$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 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* $(C, \alpha)$ 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 $(C, \alpha)$ 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.

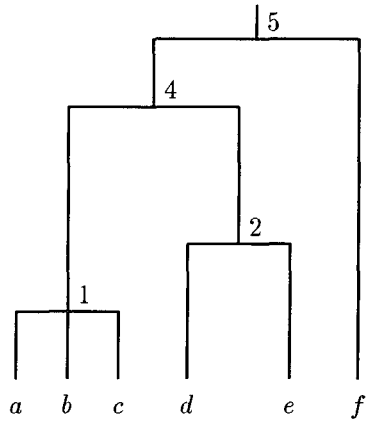



**Figura 1.**
Dendrograma representando la matriz ultramétrica $\boldsymbol{\Delta}$.

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 $\mu_0$ 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.

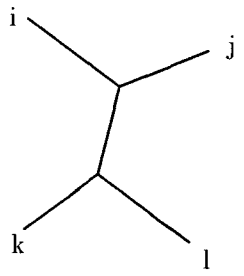



**Figura 2.**
Grafo de cuatro puntos verificando la desigualdad aditiva.

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 $\psi : U \to \mathbb{R}$, 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 $n - 1$.*

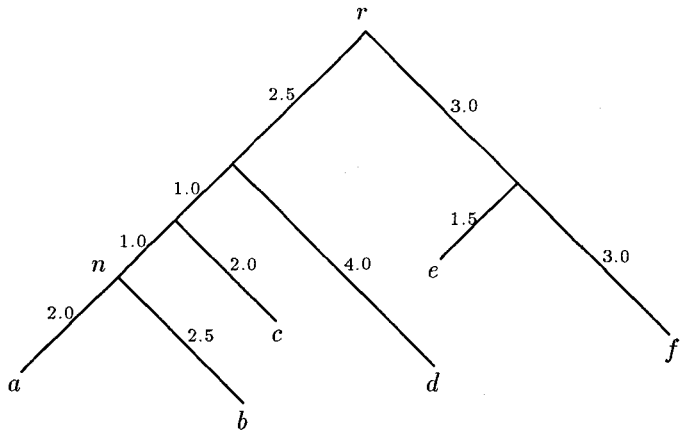



**Figura 3.**
Árbol aditivo representando la matriz cuadripolar $\Delta_c$.

### 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 $\Delta_R$ no es ultramétrica.

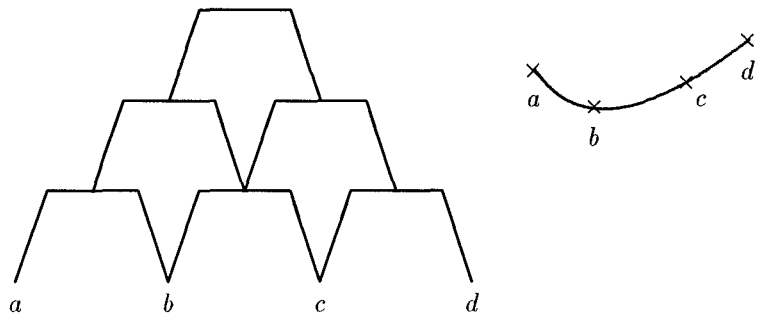



**Figura 4.**
Representación piramidal correspondiente a la matriz de Robinson $\Delta_R$. A la derecha aparece una posible ordenación cronológica.

Una *pirámide indexada* $(P, \alpha)$ es una pirámide $P$, con un índice $\alpha$ 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 $\boldsymbol{\Delta}$ 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 $\boldsymbol{\Delta} = (\delta_{ij})$ 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.

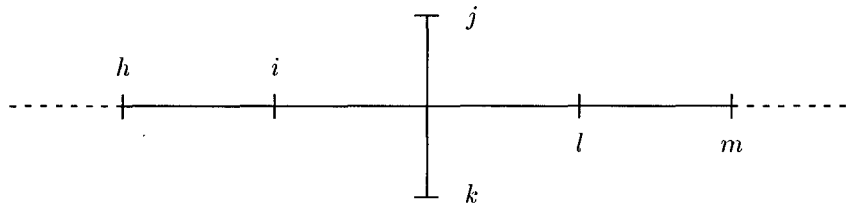



**Figura 5.**
Ordenación cronológica definida por una matriz Robinson fuerte.

El siguiente resultado describe las matrices de Robinson que pueden ser representadas mediante un árbol aditivo: Si $\boldsymbol{\Delta}$ 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 $\Xi$ de variables, posiblemente de tipo mixto, es decir, conteniendo variables continuas, binarias, y cualitativas. Supongamos que la observación de $\Xi$ 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 $\boldsymbol{\Delta}$. El esquema de la predicción basada en distancias es:

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

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 $\boldsymbol{\Delta}$ según el teorema 4.1), del vector $y$ de observaciones de $Y$ sobre $U$, y de las observaciones $\xi_{n+1}$ de $\Xi$ 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 $\Xi$ de variables de tipo mixto.

2. Predecir $Y$ cuando la relación con $\Xi$ es no lineal.

3. Predecir $Y$, discreta con $g$ estados, como un problema de clasificación, siendo $\Xi$ 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 $\beta_k$ es un vector $k \times 1$ de parámetros. Este modelo ha

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

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

$\Lambda_k$ 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 $\delta_{ij}$ 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 $\pi_k$, 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 $\boldsymbol{\Delta}(k) = (\delta_{ij}(k))$ 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 $\pi_i$ 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 $\delta_{ij}$ 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 $\Xi$ es un vector aleatorio con densidad de probabilidad $p_k(\xi)$ respecto a una cierta medida $\lambda$ en la subpoblación $\pi_k$, 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 $\xi_0$ 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 $\pi_i$, 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 $\xi$, $\eta$ 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 $\Xi$ en $\pi_k$ es $N(\mu_k, \Sigma)$, y $\delta^2$ es la distancia de Mahalanobis, entonces **[DB]** es equivalente al discriminador lineal. Una sencilla modificación de $\delta^2$ 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 $\pi_k$, 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 $\xi_1$ y $\xi_2$, respectivamente, entonces la regla se reduce a asignar $\xi_0$ a aquella $\pi_k$ tal que la probabilidad $p_{kr}$ es máxima.

3) Si $\Xi_1$ y $\Xi_2$ son vectores independientes con distancias asociadas $\delta_1$, $\delta_2$, 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 $\pi_k$ proporcionará un valor bajo en (12), luego tenderemos a asignar $\xi_0$ a $\pi_k$.

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). 2ª 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 $\Sigma$ 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 $\Psi$ is an estimable parametric function, can be expressed in terms of a Mahalanobis–like distance between $\widehat{\Psi}$ and $\Psi_0$.

### 2.2. Kullback–Leibler Divergence

This divergence between probability densities $p$, $q$ with respect to a measure $\mu$, 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 $\theta$ verifies that the divergence $K$ between $p(x,\widehat{\theta})$ and $p(x,\theta_0)$ is a minimum, where $\theta_0$ 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 $\Gamma$ 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 $\theta$ is obtained by minimizing the distance $\delta(G_n, F_\theta)$ between the empirical distribution $G_n$ and distributions in $\Gamma$. Kolmogorov and Cramér–von Mises statistics are used as the measure $\delta$ 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 $\Lambda$ 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 $\boldsymbol{\Delta}$ defined on $U$, Theorem 4.1 gives the condition for $\boldsymbol{\Delta}$ 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 $(C, \alpha)$ 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 $\boldsymbol{\Delta}$ 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 $\Xi$ of explanatory variables. We present the following cases:

1. $Y$ continuous, $\Xi$ mixed variables.
2. $Y$ continuous, $\Xi$ continuous, nonlinear relationship.
3. $Y$ discrete with $g$ states, $\Xi$ 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 $\Xi$ 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 $\Xi$ 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 $\Xi$.

### 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.

QÜESTIIÓ, Vol 17, 1 pp. 75-101, 1993

# MÉTODOS PARA LA COMPROBACIÓN DE LA INTEGRIDAD EN BASE DE DATOS DEDUCTIVAS

LAURA MOTA HERRANZ y MATILDE CELMA GIMÉNEZ*

Universidad Politécnica de Valencia

*La comprobación de la integridad es un problema clásico en bases de datos; los primeros métodos fueron propuestos para simplificar la comprobación de restricciones estáticas en bases de datos relacionales extendiéndose posteriormente a las bases de datos deductivas. Estos métodos se basan en la idea común de evaluar instancias de las restricciones, obtenidas a partir de actualizaciones inducidas por la transacción, y se diferencian entre sí en la estrategia seguida para la instanciación y evaluación de las restricciones.*

*En este trabajo se presenta una clasificación de los métodos más importantes propuestos en la literatura haciendo un análisis de los mismos.*

**Methods for Integrity Checking in Deductive Databases.**

**Key words:** Bases de Datos Deductivas, Restricciones de Integridad.

---

*Dept. Sistemas Informáticos y Computación. Universidad Politécnica de Valencia. email mati@dsic.upv.es.

-Article rebut el juny de 1991.

-Acceptat el desembre de 1992.

## 1. INTRODUCCIÓN

Las Bases de Datos Deductivas (BDD) extienden la capacidad expresiva de las Bases de Datos Relacionales (BDR) incorporando reglas que permiten derivar información a partir de la explícitamente almacenada.

El esquema de una BDD consiste en un conjunto de esquemas de relación de la forma: $R(A_1{:}\, D_1, \ldots, A_n{:}\, D_n)$ donde $R$ es el nombre de la relación, $A_1, \ldots, A_n$ son identificadores de atributos y $D_1, \ldots, D_n$ los dominios asociados; y un conjunto de Restricciones de Integridad (RI). En este esquema se distinguen dos tipos de relaciones: básicas cuyas tuplas se almacenan explícitamente y derivadas definidas a partir de reglas deductivas. Una BDR es un caso particular de BDD sin relaciones derivadas. Un estado de BDD consiste en un conjunto de hechos (tuplas de las relaciones básicas) más un conjunto de reglas deductivas.

Desde la perspectiva de la lógica, una BDD puede formalizarse de la forma siguiente (Reiter (1984), Lloyd (1987b)):

- el esquema de la BDD se representa por un par ($R$, RI) donde:

  $R$ es un lenguaje relacional definido a partir de los esquemas de relación y RI es el conjunto de restricciones de integridad (fórmulas cerradas de $R$)

- un estado $D$ se representa por una teoría de primer orden definida en $R$:

$$D = \{A{:}\, A \text{ es un átomo base (hecho)}\}$$

$$\cup$$

$$\{A \leftarrow W{:}\, A \text{ es un átomo y } W \text{ una fórmula bien formada (fbf)}\}.$$

Como es sabido, las restricciones de integridad son propiedades que la base de datos debe satisfacer en cualquier instante. La evolución en el tiempo de una BDD puede describirse por una secuencia de estados donde dado un estado, $D$, su sucesor, $D'$, se obtiene aplicando a $D$ una transacción $T$ (conjunto de inserciones y/o borrados de hechos y/o reglas). Dependiendo del número de estados implicados en la propiedad existen dos tipos de restricciones de integridad: restricciones estáticas que dependen sólo del estado actual de la base de datos y restricciones dinámicas que dependen de dos o más estados. En el presente trabajo sólo se tratarán las restricciones estáticas.

La comprobación de la integridad es un problema clásico en bases de datos; los primeros métodos fueron propuestos para la comprobación de restricciones estáticas en bases de datos relacionales extendiéndose posteriormente a las bases de datos deductivas. La forma más sencilla de comprobar las restricciones estáticas es evaluar cada una de ellas después de la transacción, sin embargo,

esta aproximación puede ser muy costosa en bases de datos voluminosas, ya que no explota el hecho de que la base de datos era íntegra (satisfacía todas las restricciones) antes de la transacción. Basándose en esta hipótesis, los métodos propuestos **simplifican la comprobación de la integridad,** evitando comprobar instancias de las restricciones que se satisfacían antes de la transacción y que además no se ven afectadas de ésta. Para que esta simplificación sea correcta las fórmulas que representan las restricciones de integridad deben ser independientes del dominio. Intuitivamente, una fórmula es independiente del dominio si su evaluación sólo depende de las extensiones de los predicados que aparecen en ella (Topor(1987)).

La estructura del trabajo es la siguiente: en el apartado 2 se presenta el método para la comprobación simplificada de la integridad en BDR de Nicolas que contiene las ideas básicas utilizadas en los métodos para BDD que se presentan en el apartado 3; en el apartado 4 se hace un análisis de estos métodos.

## 2. COMPROBACIÓN DE INTEGRIDAD EN BDR: MÉTODO DE NICOLAS (NICOLAS (1982))

El método de Nicolas consiste en dada una base de datos íntegra y una transacción $T$ (conjunto de inserciones o borradas de tuplas) obtener *instancias simplificadas de las restricciones de integridad relevantes para la transacción* que será suficiente comprobar en $D'$ para asegurar su integridad.

### Restricciones de integridad relevantes para $T$

Si $W$ es una restricción de integridad de rango restringido (independiente del dominio) podemos afirmar que:

> "***W* será relevante respecto a la inserción (borrado) de la tupla** $(e_1, e_2, \ldots, e_n)$ en $R$ si y sólo si $R(e_1, e_2, \ldots, e_n)$ es unificable con un átomo que ocurre negativamente (resp. positivamente) en $W$"

### Instancias de las restricciones de integridad

Sea:

- $W$ una restricción de integridad relevante respecto a la inserción (resp. borrado) de $(e_1, e_2, \ldots, e_n)$

- $\phi$ el unificador más general (mgu) que unifica $R(e_1, e_2, \ldots, e_n)$ con un átomo que ocurre negativamente (resp. positivamente) en $W$
- $\theta$ la restricción de $\phi$ a aquellas variables cuantificadas universalmente no precedidas de un cuantificador existencial

entonces, definimos una **instancia de $W$** generada por la inserción (resp. borrado) de $(e_1, e_2, \ldots, e_n)$ como la fórmula $W\phi$.

**Simplificación de la comprobación de la integridad**

Si $\theta$ es el conjunto de sustituciones obtenidas de la forma anterior considerando todas las operaciones de la transacción $T$ y $W_s$ es la fórmula resultante de aplicar a $W\theta_1 \wedge W\theta_2 \wedge \cdots \wedge W\theta_n$ $(\forall\theta_i \in \Theta)$ las siguientes reglas de simplificación:

- sustituir cada ocurrencia de una tupla insertada (resp. borrada) por el valor cierto (resp. falso)
- aplicar reglas de absorción

entonces se cumple: ***D' satisface W si y sólo si D' satisface $W_S$.***

Si $\Theta$ es el conjunto vacío, la restricción $W$ no se ve afectada por la transacción es decir, ésta sigue satisfaciéndose en $D'$.

Si una sustitución $\theta \in \Theta$ coincide con la sustitución identidad, la restricción $W$ no se puede simplificar y debe ser evaluada en $D'$ en su forma original.

El método de Nicolas que se ha presentado utiliza comp($D$) como teoría de primer orden que representa el estado actual $D$, donde comp($D$) es la compleción de $D$ definida en Clark (1978). En esta semántica, si $(R, \text{RI})$ es el esquema de una BDR, se dice que el estado $D$ satisface la restricción $W$ ($W \in$ RI) si y sólo si $W$ es consecuencia lógica de comp($D$) (Reiter (1984)).

**Ejemplo:**

Sea:

$D = \{p(1,1), p(2,2), q(1,1,1), q(1,2,2), q(2,1,1)\}$ un estado íntegro de base de datos,

$\text{RI} = \{W = \forall x \forall y (p(x,y) \rightarrow \exists z\, q(z,x,y))\}$ el conjunto de restricciones de integridad y

$T =$ {insertar $p(3,3)$, borrar $q(1,1,1)$} una transacción.

El método detecta que $W$ es relevante respecto a ambas operaciones:

| | | | |
|---|---|---|---|
| insertar $p(3,3)$: | $\theta_1$ | $=$ | $\{x/3, y/3\}$ |
| | $W\theta_1$ | $=$ | $p(3,3) \rightarrow \exists z\, q(z,3,3)$ |
| borrar $q(1,1,1)$: | $\theta_2$ | $=$ | $\{x/1, y/1\}$ |
| | $W\theta_2$ | $=$ | $p(1,1) \rightarrow \exists z\, q(z,1,1)$ |

Simplificando las fórmulas anteriores obtenemos:

$$W_S = \exists z\, q(z,3,3) \wedge p(1,1) \rightarrow \exists z\, q(z,1,1)$$

$$D' \text{ viola } W_S \Longrightarrow D' \text{ viola } W$$

## 3. COMPROBACIÓN DE INTEGRIDAD EN BASES DE DATOS DEDUCTIVAS

### 3.1. Introducción

Siguiendo las ideas de Nicolas para el caso relacional, los métodos para la comprobación simplificada de la integridad propuestos para BDD se basan en la idea común de evaluar instancias de las restricciones, obtenidas a partir de actualizaciones inducidas por la transacción, y se diferencian entre sí en la estrategia seguida para la instanciación y evaluación de las restricciones. La existencia de reglas deductivas introduce un nuevo problema ya que las actualizaciones inducidas por la transacción no son sólo las explícitamente requeridas por ésta, sino también las inducidas por estas reglas.

Independientemente de la estrategia seguida, los métodos propuestos simplifican la comprobación de la integridad en dos fases: en una primer fase, de generación, se obtiene instancias simplificadas de las restricciones de integridad a partir de las actualizaciones de la transacción; en una segunda fase, de evaluación, estas instancias se evalúan en el nuevo estado según el concepto de satisfacción asumido. De acuerdo a esto, los métodos objeto de estudio pueden clasificarse en dos grupos:

*a) Métodos con fase de generación potencial (sin acceso a los hechos).*

En estos métodos en la fase de generación se obtiene, a partir de las operaciones de la transacción y de las reglas deductivas, un conjunto de actualizaciones

potenciales con las cuales se instancian y simplifican las restricciones de integridad siguiendo las ideas de Nicolas. (Las actualizaciones potenciales son átomos parcialmente instanciados tales que cualquier actualización real es una instancia de uno de ellos).

En este grupo se incluyen, entre otros, los métodos presentados por Lloyd (1987a), Bry (1988) y Asirelli (1988).

**Ejemplo:**

Sea:

- $D$ la siguiente Base de Datos Deductiva:

$$\begin{array}{lll} p(x,y) & \leftarrow & t(x), r(y) \\ r(a) & & \\ t(a) & & \\ s(a,a) & & \end{array}$$

- RI $= \{\forall x \forall y \ (s(x,y) \rightarrow p(x,y))\}$ una restricción de integridad
- $T = \{\text{borrar } (t(a))\}$ la transacción

En este ejemplo, el conjunto de actualizaciones potenciales es:

$$\{\neg t(a), \neg p(a,y)\}.$$

Siguiendo las ideas de Nicolas se obtiene la instancia de la restricción:

$$(\forall y(s(a,y) \rightarrow p(a,y))$$

que evaluada en el nuevo estado nos permite concluir que la transacción viola la restricción de integridad.

*b) Métodos sin fase de generación potencial (con acceso a los hechos).*

En estos métodos, en la fase de generación se accede a la base de datos explícita.

En este grupo se incluyen, entre otros, los métodos presentados en Decker (1986), Sadri (1987), Olive (1991) y Das (1989).

Algunos de estos métodos representan las restricciones de integridad en forma negada, para ello cada restricción $W_i$ es sustituida por la fórmula $\leftarrow inc_i$, donde el predicado de inconsistencia $inc_i$ se define por la regla $inc_i \leftarrow \neg W_i$ que debe ser

incluida en la base de datos. La comprobación de la integridad en los métodos que utilizan esta representación se reduce a determinar las inserciones de átomos de inconsistencia generadas por la transacción.

**Ejemplo:**

Sea

- $D$ la siguiente Base de Datos Deductiva:

$$\begin{array}{lll} p(x,y) & \leftarrow & t(x), r(y) \\ r(a) & & \\ t(a) & & \\ s(a,a) & & \\ inc & \leftarrow & s(x,y),\ \neg p(x,y) \end{array}$$

- RI $= \{\forall x \forall y (s(x,y) \rightarrow p(x,y))\}$ una restricción de integridad cuya forma negada es: $\leftarrow inc$ donde el predicado $inc$ viene definido por la última regla deductiva.
- $T = \{$ borrar $(t(a))\}$ la transacción.

Para determinar si se ha insertado algún átomo de inconsistencia se pueden derivar actualizaciones inducidas a partir de la operación de la transacción:

$$\begin{array}{ll} \neg t(a) & \\ \Big| & p(x) \leftarrow r(x), t(x) \\ & \alpha = x/a \\ \neg p(a) & \\ \Big| & inc \leftarrow s(x,y),\ \neg p(x) \\ & \alpha = x/a \\ & \beta = y/b \\ \Big| & \\ inc & \end{array}$$

En este ejemplo, la transacción viola la integridad de la Base de Datos al provocar la inserción de un átomo de inconsistencia.

A continuación se presentan los métodos más importantes de cada grupo indicando para cada uno de ellos:

- concepto de satisfacción
- representación y propiedades de las restricciones

- tipo y propiedades de la base de datos
- características del método.

### 3.2. Concepto de satisfacción

En base de datos deductivas existen tres definiciones del concepto de satisfacción. Sea $D$ un estado de base de datos y $W$ una restricción de integridad:

*a)* punto de vista de la *demostración:*

$$\boldsymbol{D} \textbf{ satisface } \boldsymbol{W} \textbf{ si y sólo si } \textbf{comp}(\boldsymbol{D}) \models \boldsymbol{W}$$

*b)* punto de vista de la *consistencia:*

$$\boldsymbol{D} \textbf{ satisface } \boldsymbol{W} \textbf{ si y sólo si } \textbf{comp}(\boldsymbol{D}) \cup \boldsymbol{W} \textbf{ es consistente}$$

*c)* punto de vista *canónico:*

$$\boldsymbol{D} \textbf{ satisface } \boldsymbol{W} \textbf{ si y sólo si } \boldsymbol{W} \textbf{ es cierta en el modelo estándar de } \boldsymbol{D}$$

donde comp($D$) es la complección de $D$ definida en Clark (1978) y el modelo estándar de $D$ es el modelo minimal definido en Apt(1988).

### 3.3. Métodos con fase de generación potencial

#### 3.3.1. *Método de Lloyd, Sonemberg y Topor (Lloyd (1987a))*

El método de Lloyd *et al.* utiliza como concepto de satisfacción, el punto de vista de la demostración: $D$ satisface $W$ si y sólo si comp$(D) \models W$.

Este método trabaja con dos conjuntos de actualizaciones potenciales (inserciones y borrados) que pueden ser calculados sin acceder a la base de datos explícita.

**Actualizaciones potenciales**

Sea $T$ una transacción y $D$ y $D'$ estados consecutivos relacionados por $T$ tales que $D \subseteq D'$ entonces, se define $\mathbf{pos}_{D,D'}$ (conjunto de inserciones potenciales)

y $\mathbf{neg}_{D,D'}$ (conjunto de borrados potenciales) inductivamente de la forma siguiente:

$\text{pos}^0_{D,D'} = \{A: A \leftarrow W \in D' \backslash D\}$ (inserciones potenciales explícitas)

$\text{pos}^{n+1}_{D,D'} = \{A\theta: A \leftarrow W \in D, B$ ocurre positivamente en $W, C \in \text{pos}^n_{D,D'}$ y $\theta = \text{mgu}(B, C)\}$

$\cup$

$\{A\theta: A \leftarrow W \in D, B$ ocurre negativamente en $W, C \in \text{neg}^n_{D,D'}$ y $\theta = \text{mgu}(B, C)\}$

(inserciones potenciales inducidas)

$\text{neg}^0_{D,D'} = \{\ \}$ (borrados potenciales explícitos)

$\text{neg}^{n+1}_{D,D'} = \{A\theta: \leftarrow W \in D, B$ ocurre positivamente en $W, C \in \text{neg}^n_{D,D'}$ y $\theta = \text{mgu}(B, C)\}$

$\cup$

$\{A\theta: A \leftarrow W \in D, B$ ocurre negativamente en $W, C \in \text{pos}^n_{D,D'}$ y $\theta = \text{mgu}(B, C)\}$

(borrados potenciales inducidos)

$\text{pos}_{D,D'} = \bigcup_{n>0} \text{pos}^n_{D,D'}$

$\text{neg}_{D,D'} = \bigcup_{n>0} \text{neg}^n_{D,D'}$

Según la anterior definición, la obtención de los conjuntos $\text{pos}_{D,D'}$ y $\text{neg}_{D,D'}$ podría significar el cálculo de infinitos conjuntos $\text{pos}^n_{D,D'}$ y $\text{neg}^n_{D,D'}$. En la práctica, el cálculo puede realizarse en un número finito de pasos si se utiliza alguna regla de parada del tipo siguiente: en lugar de calcular los conjuntos $\text{pos}^n_{D,D'}$ y $\text{neg}^n_{D,D'}$ calcular los conjuntos $P^n$ y $N^n$ definidos de la siguiente forma:

$P^n(N^n) = \{A: A \in \text{pos}^n_{D,D'}(\text{neg}^n_{D,D'})$ y no existe $A' \in \text{pos}^k_{D,D'}(\text{neg}^k_{D,D'})(0 \leq k \leq n)$ tal que $A$ es una instancia de $A'\}$

La computación finaliza cuando para un valor de $n$, los conjuntos $P^n$ y $N^n$ son vacíos.

El conjunto $\text{pos}_{D,D'}$ (resp. $\text{neg}_{D,D'}$) se caracteriza porque cualquier inserción real (resp. borrado real) es una instancia de alguno de sus elementos.

### Teorema de Simplificación

Sea:

- $(R, \text{RI})$ el esquema de una base de datos deductiva, donde:

  $R$ es un lenguaje relacional tipado y
  $\text{RI} = \{W: \forall x_1 x_2 \ldots x_n W'\}$ el conjunto de restricciones de integridad en forma prenexa normal.

- $T$ una transacción que no modifica el lenguaje $R$ y que no es contradictoria (no exige la inserción y el borrado del mismo hecho o regla):
$$D - T_{\text{del}} \rightarrow D'' - T_{\text{ins}} \rightarrow D'$$
$T_{\text{del}}$: borrados de $T$
$T_{\text{ins}}$: inserciones de $T$.

- $D$ y $D'$ son estratificadas.

- $\Theta = \{\theta : \theta$ es la restricción a $x_1, x_2, \ldots, x_n$ de un mgu de un átomo que ocurre negativamente en $W$ y un átomo de $\text{pos}_{D'',D'}$ o de un átomo que ocurre positivamente en $W$ y un átomo de $\text{neg}_{D'',D'}\}$.

- $\Psi = \{\varphi : \varphi$ es la restricción a $x_1, x_2, \ldots, x_n$ de un mgu de un átomo que ocurre positivamente en $W$ y un átomo de $\text{pos}_{D'',D}$ o de un átomo que ocurre negativamente en $W$ y un átomo de $\text{neg}_{D'',D}\}$.

Se cumple:

a) $D'$ satisface $W$ si y sólo si $D'$ satisface $\forall(W'\phi)$ para todo $\phi \in \Psi \cup \Theta$.

b) Si $D' \cup \{\leftarrow \forall(W'\phi)\}$ tiene una refutación-SLDNF para todo $\phi \in \Psi \cup \Theta$ entonces $D'$ satisface $W$.

c) Si $D' \cup \{\leftarrow \forall(W'\phi)\}$ tiene un árbol-SLDNF fallado finitamente para algún $\phi \in \Psi \cup \Theta$ entonces $D'$ viola $W$.

Si $\Psi \cup \Theta$ es el conjunto vacío $W$ no se ve afectada por la transacción.

Si $\epsilon \in \Psi \cup \Theta$ $W$ no admite simplificación.

### 3.3.2. *Método de Bry, Decker y Manthey (Bry 1988)*

El método de Bry *et al.* utiliza como concepto de satisfacción, el punto de vista canónico: $D$ satisface $W$ si y sólo si $W$ es cierto en el modelo estándar de $D$.

Este método considera sólo operaciones simples de hechos: inserciones de hechos no explícitos en $D$ y borrados de hechos explícitos en $D$. En Bry (1987) esta propuesta se extiende considerando transacciones generales.

Sea $L$ un literal y $U$ un literal base que representa una operación (un literal positivo representa la inserción de un hecho y un literal negativo el borrado) entonces, se definen los siguientes conceptos:

### Actualizaciones potenciales

Un átomo $A$ (resp. $\neg A$) no necesariamente base *depende directamente de* $L$ si y sólo si:

- existe una regla deductiva, $A' \leftarrow B$ tal que $B$ contiene un literal $L'$ unificable con $L$ (resp. con el complementario de $L$)
- $A = A'\theta$, donde $\theta = \text{mgu}(L, L')$ (resp. mgu (complementario de $L, L'$)).

$A$ *depende* de $L$ si y sólo si $A$ depende directamente de $L$ o de un literal que depende de $L$. Todo literal que depende de $U$ es una *actualización potencial* inducida por $U$.

El concepto de actualización potencial definido de esta forma coincide con el de Lloyd, aunque Lloyd diferencia dos conjuntos de átomos *pos* (inserciones potenciales) y *neg* (borrados potenciales) en lugar del conjunto único de literales (actualizaciones potenciales) del método de Bry.

### Actualizaciones reales

Un átomo base $A$ (resp. $\neg A$) *es inducido directamente por* $L$ sobre $D'$ si y sólo si:

- existe una regla deductiva, $A' \leftarrow B$ tal que $B$ contiene un literal $L'$ unificable con $L$ (resp. con el complementario de $L$) con mgu $\theta$.
- $A = (A'\theta)\varphi$, donde $\varphi$ es una respuesta obtenida al evaluar $(B/L')\theta$ en $D'(B/L'$ representa $B$ sin $L'$ o cierto si $B = L')$
- $A$ (resp. $\neg A$) se evalúa a falso (resp. cierto) en $D$ (resp. en $D'$)

Un literal es *inducido* por $L$ sobre $D'$ si y sólo si es directamente inducido por $L$ sobre $D'$ o por un literal inducido por $L$ sobre $D'$. Todo literal inducido por $U$ sobre $D'$ es una *actualización real* inducida por $U$.

Al igual que en el método de Lloyd, toda actualización real es una instancia de una actualización potencial.

TEOREMA DE SIMPLIFICACIÓN

Sea:

- $(R$, RI) el esquema de una base de datos deductiva:

  $R$ es un lenguaje relacional y

RI el conjunto de restricciones de integridad, fórmulas de cuantificadores restringidos (independiente del dominio), en forma normalizada. En una fórmula de cuantificadores restringidos las subfórmulas cuantificadas tienen la forma:

$$\exists x_1, x_2, \ldots, x_n \, (A_1 \wedge \cdots \wedge A_m \wedge Q)$$
$$\forall x_1, x_2, \ldots, x_n \, (\neg A_1 \vee \cdots \vee \neg A_m \vee Q)$$

donde cada $x_i (1 \leq 1 \leq n)$ aparece en algún $A_j (1 \leq i \leq m)$ y $Q$ es una fórmula bien formada de cuantificadores restringidos.

La forma normalizada utilizada en el método se obtiene:

- reduciendo al máximo el ámbito de los cuantificadores
- expresando las implicaciones y equivalencias con las conectivas lógicas $\wedge, \vee, \neg$
- llevando las negaciones sobre los átomos
- distribuyendo las disyunciones respecto a las conjunciones.

- $D = \{A: A$ es un átomo base$\} \cup \{A \leftarrow W: W$ es una conjunción de literales$\}$.
- $U$ un literal base que representa la operación.
- $D, D'$ son estratificadas.
- *delta* es un meta-predicado, tal que para todo literal totalmente instanciado $L$, delta$(U, L)$ es cierto si y sólo si $L$ es cierto en $D'$ y falso en $D$.
- *new* un meta-predicado tal que new$(U, F)$ es cierto si y sólo si $F$ es cierto en $D'$.
- RA $= \{\forall$ delta$(U, L\theta) \rightarrow$ new$(U, W_s)\}$ el conjunto de restricciones auxiliares asociadas a $W$ que se obtiene aplicando a la restricción el método de simplificación de Nicolas a partir del conjunto de actualizaciones potenciales inducidas por $U$ donde:
  - $L$ es una actualización potencial inducida por $U$ o bien por $U$.
  - $\theta$ es la restricción de un mgu entre un literal $L'$ de $W$ y el complementario de $L$, a las variables de $L'$ cuantificadas universalmente no precedidas de un cuantificador existencial y a las variables de $L$.
  - $W_s$ es la instancia simplificada de $W$ obtenida a partir de $W\theta$ aplicando las siguientes reglas de simplificación
    i) eliminando los cuantificadores sobre las variables instanciadas por $\theta$

ii) reemplazando cada literal de $W\theta$ unificable con el complementario de $L\theta$ por falso y aplicando las reglas de absorción.

Se cumple:

$D'$ satisface $W$ si y sólo si las restricciones auxiliares asociadas a $W$ se satisfacen en $D'$.

## 3.4. Métodos sin fase de generación potencial

### 3.4.1. *Método de Decker (Decker (1986))*

El método de Decker utiliza como concepto de satisfacción el punto de vista de la demostración: $D$ satisface $W$ si y sólo si $\text{comp}(D) \models W$.

Este método obtiene instancias simplificadas de las restricciones de integridad a partir de las actualizaciones reales inducidas por la transacción.

#### Actualizaciones reales

Los conjuntos de actualizaciones reales asociados a una cláusula $C$ de la forma $H \leftarrow B$ se definen de la forma:

$$D^C = \{H\theta: H\theta \text{ es base, } \text{comp}(D') \models B\theta \text{ y } \text{comp}(D) \not\models H\theta\}$$
$$D_C = \{H\theta: H\theta \text{ es base, } \text{comp}(D) \models B\theta \text{ y } \text{comp}(D') \not\models H\theta\}.$$

ALGORITMO DE SIMPLIFICACIÓN

Sea:

- $(R, \text{RI})$ el esquema de una base de datos deductiva, donde:

  $R$ es un lenguaje relacional y

  RI el conjunto de restricciones de integridad, fórmulas de rango restringido

- $D = \{A: A \text{ es un átomo base}\} \cup \{A \leftarrow W: W \text{ es una conjunción de literales}\}$

- $T$ una transacción formadas por operaciones de la forma:

  - insertar $(C)$
  - borrar$(C)$

  donde $C$ es una cláusula.

- $D$ y $D'$ son estratificadas.
- RA = {insertar $\underline{L}$ sólo si $W^{\underline{L}}$: $L$ es un literal positivo que ocurre negativamente en $W$}
  $\cup$
  {borrar $\underline{L}$ sólo si $W_{\underline{L}}$: $L$ es un literal positivo que ocurre positivamente en $W$}

  el conjunto de restricciones auxiliares asociado a $W$, donde:
  - $\underline{L}$ se obtiene a partir de $L$ renombrando las variables de $L$ cuantificadas existencialmente o cuantificadas universalmente precedidas de un cuantificador existencial
  - las formas simplificadas $W^{\underline{L}}(W_{\underline{L}})$ se obtienen reemplazando cada ocurrencia positiva de $\underline{L}$ en $W$ por el valor cierto (falso), cada ocurrencia negativa de $\underline{L}$ en $W$ por el valor falso (cierto) y aplicando las correspondientes reglas de absorción.

La comprobación simplificada de la integridad se realiza practicando el siguiente algoritmo $\alpha$ con argumento insertar($C$) (resp. borrar($C$)) para cada cláusula $C$ para la cual existe una operación insertar($C$) (resp. borrar($C$)) en la transacción.

ALGORITMO $\alpha$

PASO 1: para cada átomo $L^*$ de $D^C$ (resp. $D_C$) y para cada restricción auxiliar **insertar $L$ sólo si $F$** (resp. **borrar $L$ sólo si $F$**) tal que $L$ unifica con $L^*$ con mgu $\theta$ evaluar $F\theta$ en $D'$. Si $F\theta$ se evalúa a falso parar (la transacción viola la integridad).

PASO 2: para cada átomo $L^*$ de $D^C$ (resp. $D_C$) y para cada cláusula ocurre_positivo $(L, R)$ tal que $L$ unifica con $L^*$ con mgu $\mu$, llamar al algoritmo con argumento insertar($R\mu$) (resp.borrar ($R\mu$)).

PASO 3: para cada átomo $L^*$ de $D^C$ (resp. $D_C$) y para cada cláusula ocurre_negativo $(L, R)$ tal que $L$ unifica con $L^*$ con mgu $\varphi$, llamar al algoritmo con argumento borrar($R\mu$) (resp.insertar($R\varphi$)).

Cuando todas las operaciones de la transacción han sido procesados por el algoritmo sin que se detecte una violación de la integridad, se puede afirmar que la transacción no viola la integridad de $D'$.

### 3.4.2. *Método de Sadri y Kowalski (Sadri (1987))*

El método de Sadri *et al.* utiliza como concepto de satisfacción, el punto de vista de la consistencia: $D$ satisface $W$ si y sólo si comp($D$)$\cup\{W\}$ es consistente.

El método se caracteriza por utilizar una extensión del procedimiento SLDNF permitiendo entonces aplicar resolución a partir de las actualizaciones de la transacción.

**Actualizaciones de la transacción**

El conjunto de actualizaciones de la transacción se define como sigue:

$\{A$: insertar_hecho$(A) \in T\}$
$\cup$
$\{A \leftarrow L_1, L_2, \ldots, L_n$: insertar_regla$(A \leftarrow L_1, L_2, \ldots, L_n) \in T\}$
$\cup$
$\{\neg A$: borrar_hecho$(A) \in T$ y $\exists$ árbol-SLDNF fallado finitamente para $D' \cup \{\leftarrow A\}\}$
$\cup$
$\{\neg A\theta$: borrar_regla$(A \leftarrow L_1, L_2, \ldots, L_n) \in T$ y $\theta$ es una respuesta computada para $D \cup \{\leftarrow L_1, L_2, \ldots, L_n\}$ y $\exists$ árbol-SLDNF fallado finitamente para $D' \cup \{\leftarrow A\}\}$.

**Procedimiento SLDNF* (SLDNF extendido)**

Sea:

- $S = D \cup$ RI donde:
  - $D$ es un conjunto de cláusulas de la forma:

    $A$ o bien
    $A \leftarrow L_1, L_2, \ldots, L_n$

  - RI un conjunto de cláusulas de la forma:

    $$\leftarrow L_1, L_2, \ldots, L_n$$

    Si $L_i$ es positivo se denomina condición positiva, si es negativo condición negativa.

- $C_0$ una cláusula de $S$ o un átomo negado $(\neg A)$ tal que $S \cup \{\leftarrow A\}$ falla finitamente utilizando el SLDNF*.
- $R$ una regla de computación segura (nunca selecciona un literal negativo no base).

Una *derivación* vía $R$ para $S \cup \{C_0\}$ es una secuencia posiblemente infinita $C_0, C_1, C_2 \ldots$ tal que $C_i$ para $i > 0$ es una cláusula, y para todo $i \geq 0$ $C_{i+1}$ se obtiene a partir de $C_i$ de la siguiente forma:

a) Si $R$ selecciona de $C_i$ un literal $L$ que no es una condición negativa de $C_i$, entonces $C_{i+1}$ es el resolvente sobre $L$ de $C_i$ y alguna cláusula de $S$.

b) Si $R$ selecciona una condición negativa, $\neg A$, de $C_i$. Entonces $C_{i+1}$ es $C_i$ eliminando el literal seleccionado, $\neg A$, si existe un árbol-SLDNF* fallado finitamente para $S \cup \{\leftarrow A\}$.

c) Si $C_i$ es $\neg A$ y en $S$ hay una cláusula $B \leftarrow \neg A', C$ de forma que $A$ y $A'$ unifican a través del mgu $\theta$ entonces $C_{i+1}$ es $(B \leftarrow C)\theta$.

El SLDNF* es **correcto** en el sentido siguiente:
"Si existe una refutación-SLDNF* para $S \cup \{C_0\}$ entonces comp($D$)$\cup$ RI es inconsistente".

TEOREMA DE SIMPLIFICACIÓN

Sea:

- ($R$, RI) el esquema de una base de datos relacional donde:

  $R$ es un lenguaje relacional y
  RI $= \{\leftarrow \text{inc}_i : \text{inc}_i \leftarrow \neg W_i,\ 1 \leq i \leq n\}$ el conjunto de restricciones de integridad en forma negada (el conjunto de cláusulas resultantes de aplicar el algoritmo de Lloyd (Lloyd (1987b)) a $\{\text{inc}_i \leftarrow \neg W_i : 1 \leq i \leq n\}$ deben ser de rango restringido y forma parte de cualquier estado de la base de datos).

- $D =$ $\{A : A$ es un átomo base$\}$ $\cup$
  $\{A \leftarrow W$: es de rango restringido y $W$ es una conjunción de literales$\}$.

- $T$ una transacción.

- $D$ y $D'$ son estratificadas.

Se cumple:

"Si existe una refutación-SLDNF* para $D'\cup$ RI $\cup\{C_0\}$ para alguna actualización $C_0$ de la transacción entonces $D'$ viola RI".

### 3.4.3. *Método de Olivé (Olivé 1991)*

El método de Olivé utiliza como concepto de satisfacción, el punto de vista de la demostración: $D$ satisface $W$ si y sólo si comp($D$) $\models W$.

El método se caracteriza por:

- extender el lenguaje de la base de datos con dos tipos de predicados:
  - i) predicados de *transición:* permiten simular la base de datos actualizada
  - ii) predicados de *eventos internos* (inserción y borrado): representan las actualizaciones generadas por la transacción
- ampliar la base de datos con las reglas deductivas que definen dichos predicados
- comprobación directa de las restricciones de integridad utilizando dichas reglas
- tratamiento uniforme para restricciones de integridad estáticas y de transición
- uso del procedimiento SLDNF para la comprobación de la integridad.

**Predicados de eventos internos: actualizaciones generadas por $T$**

Para cada predicado $P$, se introduce un predicado de transición $P'$, que representa el predicado $P$ en la base de datos actualizada $D'$ y dos predicados de eventos internos:

$\iota P$: predicado de evento interno de inserción
$\delta P$: predicado de evento interno de borrado

estos predicados representan las actualizaciones (inserciones y borrados) explícitas o inducidas generadas por una transacción $T$ sobre el predicado $P$. De la definición de actualización real inducida por una transacción podemos afirmar:

$$\forall x_1, x_2, \ldots, x_n(\iota P(x_1, x_2, \ldots, x_n) \leftrightarrow P'(x_1, x_2, \ldots, x_n) \wedge \neg P(x_1, x_2, \ldots, x_n))$$
$$\forall x_1, x_2, \ldots, x_n(\delta P(x_1, x_2, \ldots, x_n) \leftrightarrow P(x_1, x_2, \ldots, x_n) \wedge \neg P'(x_1, x_2, \ldots, x_n))$$

Si $P$ es un predicado base: $\iota P$, $\delta P$ serán predicados base cuyos hechos serán determinados directamente por las actualizaciones explícitas de la transacción $T$ sobre $P$.

Si $P$ es un predicado derivado: $\iota P, \delta P$ serán predicados derivados definidos por reglas deductivas que nos permitirán obtener las actualizaciones inducidas por la transacción $T$ sobre $P$.

**Reglas de eventos internos**

Para poder determinar las reglas deductivas que definen los predicados $\iota P$ y $\delta P$ para predicados derivados de la base de datos vamos a determinar previamente las reglas que definen el predicado $P'$.

Como $P'$ representa al predicado $P$ en la base de datos actualizada, se cumplirá:

$$\forall x_1, x_2, \ldots, x_n (P'(x_1, x_2, \ldots, x_n) \leftrightarrow (P(x_1, x_2, \ldots, x_n) \wedge \\ \wedge \neg\delta P(x_1, x_2, \ldots, x_n)) \vee \iota P(x_1, x_2, \ldots, x_n))$$

$$\forall x_1, x_2, \ldots, x_n (\neg P'(x_1, x_2, \ldots, x_n) \leftrightarrow (\neg P(x_1, x_2, \ldots, x_n) \wedge \\ \wedge \neg\iota P(x_1, x_2, \ldots, x_n)) \vee \delta P(x_1, x_2, \ldots, x_n))$$

Si el predicado derivado $P$ viene definido por $m$ reglas deductivas en $D$:

$$P \leftarrow P_i \quad 1 \leq i \leq m \quad \text{siendo} \quad P_i \leftrightarrow L_1 \wedge L_2 \wedge \cdots \wedge L_n$$

entonces $P'$ estará definido por $m$ reglas deductivas de la forma:

$$P' \leftarrow P'_i \quad 1 \leq i \leq m \quad \text{siendo} \quad P'_i \leftrightarrow L'_1 \wedge L_2 \wedge \cdots \wedge L'_n$$

reemplazando cada $L'_j$ de la siguiente forma:

si $L_j = Q_j(x_1, x_2, \ldots, x_n)$ (literal positivo) entonces:

$$L'_j = (Q_j(x_1, x_2, \ldots, x_n) \wedge \neg\delta Q_j(x_1, x_2, \ldots, x_n)) \vee \iota Q_j(x_1, x_2, \ldots, x_n)$$

si $L_j = \neg Q_j(x_1, x_2, \ldots, x_n)$ (literal negativo) entonces:

$$L'_j = (\neg Q_j(x_1, x_2, \ldots, x_n) \wedge \neg\iota Q_j(x_1, x_2, \ldots, x_n)) \vee \delta Q_j(x_1, x_2, \ldots, x_n)$$

con lo que se obtienen $m$ reglas (cláusulas no normales) que definen $P'$ en términos (conjunción de disyunciones) de predicados de base de datos y predicados de eventos internos. Para obtener un conjunto equivalente de cláusulas normales basta distribuir las conjunciones sobre las disyunciones.

Una vez obtenidas las reglas que definen $P'$, las reglas que definen $\iota P$ y $\delta P$ son:

$$\iota P(x_1, x_2, \ldots, x_n) \leftarrow P'(x_1, x_2, \ldots, x_n) \wedge \neg P(x_1, x_2, \ldots, x_n)$$
$$\delta P(x_1, x_2, \ldots, x_n) \leftarrow P'(x_1, x_2, \ldots, x_n) \wedge \neg P'(x_1, x_2, \ldots, x_n).$$

Las reglas de eventos internos pueden simplificarse después de aplicarles algunas transformaciones como se indica en Olivé (1991).

TEOREMA DE SIMPLIFICACIÓN

Sea:

- $(R, \mathrm{RI})$ el esquema de una base de datos deductiva donde:

  $R$ es un lenguaje relacional y
  $\mathrm{RI} = \{\leftarrow \mathrm{inc}_i : \mathrm{inc}_i \neg W_i,\ 1 \leq i \leq n\}$ el conjunto de restricciones de integridad en forma negada (el conjunto de cláusulas resultantes de aplicar el algoritmo de Lloyd a $\{\mathrm{inc}_i \leftarrow \neg W_i : 1 \leq i \leq n\}$ deben ser de rango restringido y forma parte de cualquier estado de la base de datos).

- $D = \{A : A$ es un átomo base$\} \cup$
  $\{A \leftarrow W$: es de rango restringido y $W$ es una conjunción de literales

- $T$ una transacción.

- $D$ y $D'$ son estratificadas.

Si $A(D)$ es la base de datos resultante de añadir a $D$ las reglas de transición y de eventos internos para cada predicado derivado y predicado de inconsistencia de $D$ entonces se cumple:

a) $D'$ viola la restricción $W_i$ si existe una refutación-SLDNF para $A(D) \cup T \cup \{\leftarrow \iota \mathrm{inc}_i\}$

b) $D'$ satisface la restricción $W_i$ si existe un árbol-SLDNF fallado finitamente para $A(D) \cup T \cup \{\leftarrow \iota \mathrm{inc}_i\}$.

### 3.4.4. *Método de Das y Williams (Das 1989)*

El método de Das *et al.* utiliza como concepto de satisfacción, el punto de vista de la demostración: $D$ satisface $W$ si y sólo si $\mathrm{comp}(D) \models W$.

En este método, la comprobación de la integridad consiste en buscar un "camino" desde una "actualización" de la transacción $T$ hasta un átomo de inconsistencia.

#### Actualizaciones de $T$

Si $A$ es un átomo base y $H \leftarrow B$ es una regla deductiva el conjunto de actualizaciones de $T$ se define:

$$\{A\text{: insertar\_hecho}(A) \in T\} \cup \{\neg A\text{: borrar\_hecho}(A) \in T\}$$
$$\cup$$
$$\{H\theta\text{: insertar\_regla}(H \leftarrow B) \in T \text{ y } \theta \text{ es una respuesta-SLDNF computada para } D' \cup \{\rightarrow B\}\}$$
$$\cup$$
$$\{\neg H\text{: borrar\_regla}(H \leftarrow B) \in T\}$$

**Camino**

Si $D$ es un estado de la base de datos, un camino se define como:

$$L_0 - R_1 \longrightarrow L_1 - R_2 \longrightarrow \cdots R_n \longrightarrow L_n$$

donde $L_0$ es el origen del camino, $L_n$ es el destino, $n$ es su longitud y $R_1, R_2, \ldots, R_n$ son cláusulas de $D$ utilizadas para construir el camino de $L_0$ a $L_n$. Si $L_0$ es positivo es base y debe existir una refutación-SLDNF para $D \cup \{\leftarrow L_0\}$.

$L_{i+1}$ se define a partir de $L_i$ de la forma:

1. Si:
   - $L_i$ es positivo
   - $L_i$ unifica con un literal positivo del cuerpo de la cláusula $R_i$: $H \leftarrow B$ con mgu $\alpha$
   - $G'$ es el resolvente de $\leftarrow B$ y $L_i$
   - $\theta$ es una respuesta-SLDNF computada para $D \cup \{\leftarrow G'\}$

   entonces $L_{i+1}$ es $H\alpha\theta$
2. Si:
   - $L_i$ es positivo
   - $L_i$ unifica con el complementario de un literal negativo del cuerpo de la cláusula $R_i$: $H \leftarrow B$ con mgu $\alpha$
   - $\neg H\alpha$ no es una instancia de algún $L_j (0 \leq j \leq i)$

   entonces $L_{i+1}$ es $\neg H\alpha$
3. Si:
   - $L_i$ es negativo
   - $L_i$ unifica con un literal negativo del cuerpo de la cláusula $R_i$: $H \leftarrow B$ con mgu $\alpha$

- $\theta$ es una respuesta-SLDNF computada para $D \cup \{G\}$ donde $G = \leftarrow B\alpha$

entonces $L_{i+1}$ es $H\alpha\theta$

4. Si:

   - $L_i$ es negativo
   - $L_i$ unifica con el complementario de un literal $L$ que ocurre en el cuerpo de una cláusula $R_i$: $H \leftarrow B$ con mgu $\alpha$
   - $\neg H\alpha$ no es una instancia de algún $L_j (0 \leq j \leq i)$

   entonces $L_{i+1}$ es $\neg H\alpha$

Un camino que finaliza en algún átomo de inconsistencia se denomina **camino de éxito** en caso contrario se denomina **camino de fallo**.

TEOREMA DE SIMPLIFICACIÓN

Sea:

- ($R$, RI) el esquema de una base de datos deductiva, donde:

  $R$ es un lenguaje relacional y

  RI = $\{\leftarrow \text{inc}_i : \text{inc}_i \neg W_i,\ 1 \leq i \leq n\}$ el conjunto de restricciones de integridad en forma negada (el conjunto de cláusulas resultantes de aplicar el algoritmo de Lloyd a $\{\text{inc}_i \leftarrow \neg W_i : 1 \leq i \leq n\}$ deben ser de rango restringido y forma parte de cualquier estado de la base de datos).

- $D =$ $\{A : A$ es un átomo base $\cup$
  $\{A \leftarrow W$: es de rango restringido y $W$ es una conjunción de literales$\}$.

- $T$ una transacción.

- $D$ y $D'$ son estratificadas.

Se cumple:

a) si existe un camino de éxito en $D'$ con origen alguna de las actualizaciones de $T$ entonces $D'$ viola RI.

b) si no existe un camino de éxito en $D'$ para ninguna de las actualizaciones de $T$ como origen entonces $D'$ satisface RI.

## 4. ANÁLISIS DE LOS MÉTODOS

En este apartado se persiguen dos objetivos:

- resumir los requisitos exigidos por los distintos métodos referentes al tipo y propiedades sintácticas de la base de datos así como a la representación y propiedades sintácticas de las restricciones.
- analizar la estrategia de cada método.

### Tipo de base de datos

El método de Lloyd trabaja con bases de datos "generales" es decir, con cláusulas de la forma: $A \leftarrow W$, donde $A$ es un átomo y $W$ una fórmula bien formada cualquiera.

Los restantes métodos se definen para bases de datos "normales" con reglas de la forma: $A \leftarrow L_1, L_2, \ldots, L_n$, donde $A$ es un átomo y $L_i$ un literal. Este requisito no quita generalidad a un método ya que cualquier base de datos general puede transformarse en una base de datos normal siguiendo el algoritmo de Lloyd sin embargo, en este caso habrá que tener en cuenta que al aplicar el algoritmo se pueden perder ciertas propiedades sintácticas de la base de datos (independencia del dominio).

### Propiedades sintácticas de la base de datos y la restricción

i) Todos los métodos exigen que la base de datos $D$ sea estratificada, asegurando de esta forma que comp($D$) es consistente y que $D$ tiene un único modelo minimal (el modelo estándar).

ii) Los métodos que utilizan el SLDNF como mecanismo procedural para la comprobación de la integridad, exigen a la base de datos y a la restricción propiedades sintácticas que aseguran que en las computaciones SLDNF realizadas no aparece el problema del "tropiezo" (en algún punto de la derivación se obtiene un objetivo que sólo contiene literales negativos no base), Lloyd (1987b).

En el método de Lloyd, el uso de un lenguaje tipado asegura que la forma normal sin tipos de una base de datos y un requerimiento es permitida (Lloyd (1987b)) eliminando por tanto el problema del "tropiezo". En los restantes métodos se exige a cada cláusula normal de la base de datos la propiedad de rango restringido (o alguna propiedad sintáctica que implique ésta). Si las restricciones se representan en forma negada se les exige la

propiedad de rango restringido y si se representa en forma general se les exige alguna propiedad sintáctica (cuantificadores restringidos) que asegure que las cláusulas resultantes al aplicar el algoritmo de Lloyd son de rango restringido.

iii) La simplificación de la comprobación de la integridad basada en consideraciones sintácticas: "instanciación de las restricciones a partir de las actualizaciones inducidas por la transacción a través de las reglas deductivas", sólo es correcta cuando:

- siendo el concepto de satisfacción el de la demostración, la base de datos junto a la restricción cumplen la propiedad de independencia del dominio. Esta propiedad permite asegurar que el conjunto de respuestas correctas para comp($D$)$\cup\{W\}$ es independiente del lenguaje.
- siendo el concepto de satisfacción el canónico, el modelo estándar de comp($D$) es independiente del lenguaje.

**Estrategia de los métodos**

El análisis de las características de los métodos nos permite realizar las siguientes consideraciones:

i) la existencia de una fase de generación potencial es interesante porque permite, sin acceder a los hechos eliminar del proceso de comprobación:

- restricciones no relevantes para la transacción
- actualizaciones no relevantes para la integridad.

ii) la instanciación de las restricciones a partir de actualizaciones potenciales, genera un conjunto de restricciones menos instanciadas que en el caso de trabajar con actualizaciones reales, y por lo tanto en general más costosas de comprobar.

iii) aunque en la fase de evaluación, las restricciones simplificadas obtenidas en la fase de generación potencial se evalúen sólo para aquellas instancias correspondientes a instancias de las actualizaciones potenciales que coinciden con una actualización real (Bry (1988)), el proceso puede ser muy costoso ya que los métodos no hacen uso de la información disponible en la fase generación referente a caminos de derivación para la obtención de estas actualizaciones reales.

iv) los métodos con fase de generación potencial pueden optimizarse intercalando la fase de generación y la fase de evaluación.

v) en los métodos sin fase de generación potencial, la selección como origen de la derivación de actualizaciones que inducen a su vez actualizaciones no relevantes para la integridad puede elevar el coste de la comprobación.

## 5. BIBLIOGRAFÍA


[1] **Apt, K., Blait, H.A.** & **Walker, A.** (1988). "Towards a Theory of Declarative Knowledge." *Foundations of Deductive Databases and Logic Programming.* Morgan Kaufman.

[2] **Asirelli, P., Inverardi, P.** & **Mustaro, A.** (1988). "Improving Integrity Constraint Checking in Deductive Databases". *Proc. 2nd International Conference on Database Theory.*

[3] **Bry, F.** & **Decker, H.** (1987). "Préserver l'Intégrité d'une Base de Données Déductive: une Méthode et son Implémentation". *4emes Journées de Bases de Donées Avancées.* Bénodet (France).

[4] **Bry, F., Decker, H.** & **Manthey, R.** (1988). "A Uniform Approach to Constraint Satisfaction and Constraint Satisfiability in Deductive Databases". *Proc. 1st Conf. Extending Database Technology.*

[5] **Clark. P** (1978). "Negation as Failure". *Logic and Databases.* Plenum.

[6] **Das, S.K.** & **Williams, M.H.** (1989). "A Path Finding Method for Constraint Checking in Deductive Databases". *Data & Knowledge Engineering,* **4**. North Holland.

[7] **Decker, H.** (1986). "Integrity Enforcement in Deductive Databases". *Proc. 1st Int. Conf. on Expert Database Systems.*

[8] **Lloyd, J.W., Soneberg, E.A.** & **Topor, R.W.** (1987a). "Integrity Constraint Checking in Stratified Databases". *Journal of Logic Programming,* 4, 331–343.

[9] **Lloyd, J.W.** (1987b). "Foundations of Logic Programming." Springer-Verlag.

[10] **Nicolas, J.M.** (1982). "Logic for Improving Integrity Checking in Relational Databases". *Acta Informatica,* **18**.

[11] **Olivé, A.** (1991). "Integrity Checking in Deductive Databases". *Proc. del 17th International Conference on Very-Large Databases.*

[12] **Reiter, R.** (1984). "Towards a logical reconstruction of Relational Database Theory." *On Conceptual Modelling.* Springer-Verlag.

[13] **Sadri, F. & Kowalski, R.** (1987). "A Theorem-proving Approach to Database Integrity". *Proc. del Workshop on Foundations of Deductive Databases and Logic Programming.*

[14] **Topor, R.W.** (1987). "Domain Independent Formulas and Databases". *Theoretical Computer Science,* **52**.

# ENGLISH SUMMARY:

## METHODS FOR INTEGRITY CHECKING IN DEDUCTIVE DATABASES

Laura Mota Herranz and Matilde Celma Giménez

Deductive Databases (DDB) are an extension of Relational Databases (RDB) since they include rules that allow us to derive new information from the information explicitly stored.

A *DDB scheme* consists of:

- a set of relation schemes of the form $R(A_1{:}\, D_1, \ldots, A_n{:}\, D_n)$ where $R$ is the name of the relation, $A_1, \ldots, A_n$ are attribute identifiers and $D_1, \ldots, D_n$ are the names of the domains associated with the attribute and
- a set of Integrity Constraints (IC).

A *database state* is a set of facts (tuples of basic relations) plus a set of deductive rules.

An *integrity constraint* is a statement that a database must satisfy at any time in order to faithfully describe the real world represented by the database. The evolution through time of a database can be described by a sequence of states where transitions from one state to the next are accomplished by database transactions (set of insertions and/or deletions of facts and/or rules). According to this evolution scheme, static and dynamic constraints can be distinguished; the former restrict the validity of each state on its own, while the latter relate the validity of a sequence of consecutive states. In the paper, we only deal with static constraints.

In previous database scheme, we can distinguish two types of relations: *basic* relations whose tuples are explicity stored and *derived* ones defined by deductive rules. A RDB is a special case of DDB without derived relations.

From the point of view of logic, a DDB can be formalized as follows (Reiter (1984), Lloyd (1987b)):

- the DDB scheme is represented by a pair ($R$, IC) where:

  $R$ is a relational language defined from the relation schemes

  IC is the set of integrity constraints (closed well formed formulas (wff) of $R$)

- a database state $D$ is represented by a first order theory defined over $R$:

$$D = \{A: A \text{ is a ground atom (fact)}\}$$

$$\cup$$

$$\{A \leftarrow W: A \text{ is an atom and } W \text{ is a wff}\}.$$

To ensure that each integrity constraint is satisfied in the new state, all of them must be checked after every transaction. However, this checking can be very costly if they are evaluated as queries, particularly in large databases. In spite of these difficulties, it is possible to reduce the amount of computation if advantage is taken of the fact that, before the transaction was made, the database was known to satisfy its integrity constraints and hence, any violation of them is due to an update induced by the transaction. Thus, every practical integrity checking method simplifies this process avoiding checking some instances of the constraints that were satisfied before the transaction (in the old state) and have not been affected by this one. This simplification will be only correct if the wff that represents the integrity constraints are domain-independent. Intuitively speaking, a wff is domain-independent if its evaluation only depends on the extensions of the predicated that appear in it (Topor (1987)).

In this paper entitled **Methods for Integrity Checking in Deductive Databases** we present the Nicolas Method for simplifying integrity checking in relational databases (Nicolas (1982)), following this proposal, many methods for simplified integrity checking in deductive databases have been proposed in the last ten years; all of them are based on the idea of evaluating simplified instances of the integrity constraints generated by the updates induced by the transaction. They differ mainly in the way these induced updates are determined and in the strategy used to instantiate and simplify the constraints. From all the methods for simplifying integrity checking in deductive databases we present the methods of Decker (Decker (1986)), Lloyd & Sonnenberg & Topor (Lloyd (1987a)), Sadri

& Kowalski (Sadri (1987)), Bry & Decker & Manthey (Bry (1988)), Das & Williams (Das (1989)) and Olivé (Olivé (1991)).

Independently of the particular strategy followed by the methods, we state that all of them simplify the integrity in two phases: in a first phase, that we will call the *generation phase,* simplified instances of the constraints are obtained using the updates induced by the transaction, in a second phase, the *evaluation phase,* these instances are evaluated in the new state. These two phases can appear separates or interleaves in the different methods.

For each one of all these methods we present:

- point of view of constraints satisfaction assumed by the method
- representation and properties of the constraints
- type and properties of the database
- features of the method

QÜESTIIÓ, Vol 17, 1 pp. 103-134, 1993

# COOPERACIÓN Y DEFENSA*

FRANCESC CARRERAS

Universidad Politécnica de Catalunya

*Se aplican conceptos y técnicas de la teoría de juegos cooperativos a problemas de decisión que afectan a la política de Defensa del país. El análisis permite evaluar las propuestas sobre procedimientos de votación cualificada presentadas al Consejo Europeo en la cumbre de Maastricht de diciembre de 1991. Se ponen así de manifiesto las implicaciones que supondría para la posición estratégica de España la inédita capacidad operativa concedida a la Comunidad por el tratado de unión política.*

**Cooperation and National Defense.**

**Key words:** Cooperative game theory, Shapley value, coalitional value, voting systems.

## 1. INTRODUCCIÓN

¿Hay una forma de valorar con precisión la distribución de poder en un colectivo organizado para tomar decisiones?

---

Francesc Carreras. Departamento de Matemática Aplicada II. ETSEI Terrassa. Universidad Politécnica de Catalunya.

*Este trabajo obtuvo un accésit al III Premio de Investigación Operativa *General Fernández-Chicarro,* concedido por el Ministerio de Defensa en mayo de 1992.

–Article rebut l'octubre de 1992.

–Acceptat el març de 1993.

Esta pregunta adquiere especial relevancia en esta última mitad del siglo XX, cuando no sólo los individuos sino grupos sociales de diverso tipo y tamaño, incluyendo naciones, se ven progresivamente incorporados a organismos de representación regidos por sistemas de votación.

Y en la última década del siglo no solamente desaparece la Unión Soviética, una de las dos partes que han mantenido la llamada guerra fría; se inician conversaciones entre árabes e israelíes y hasta la Comunidad Económica Europea se propone crear un espacio de Defensa común.

Afortunadamente, la paz parece ganar terreno, y probablemente una parte importante de nuestra política de Defensa se va a desarrollar en organismos como Naciones Unidas o la Unión Europea. No parece, pues, mal momento para reflexionar sobre este tipo de centros de decisión y disponer de capacidad de análisis para evaluar nuestra posición exacta dentro de cada uno y nuestras posibilidades estratégicas en este marco.

El argumento es aplicable, mutatis mutandis, a otra parcela fundamental de nuestra política exterior, la económica. El origen de este trabajo se encuentra, sin embargo, en la polémica suscitada por las propuestas formuladas en la reunión de Maastricht de diciembre de 1991 en orden a construir un marco legal de actuación para una política común de Defensa y Relaciones con el exterior de los países integrantes de la Comunidad Económica Europea. Tales propuestas pretendían desarbolar la soberanía de los estados miembros —atrincherada en la regla de unanimidad— introduciendo mecanismos de decisión por mayoría (ponderada y) cualificada, extensibles a corto o medio plazo a temas económicos y sociales.

Dos son los objetivos básicos de este trabajo: (1) presentar dos medidas del poder a nivel de organismos de decisión, una individual y otra coalicional, precisando el tipo de estructuras sobre las que están definidas; (2) aplicarlas al estudio de las propuestas de Maastricht, valorando el alcance y efectos que se derivarían de su aceptación e ilustrando al mismo tiempo el uso de estas técnicas, propias de la rama cooperativa de la Teoría de Juegos.

La Sección 2 servirá al primer objetivo, buscando una exposición de lectura cómoda donde sea primordial el significado y alcance de los conceptos y métodos expuestos a costa, si es necesario, del formalismo matemático subyacente al tema, que puede hallarse en las referencias. La Sección 3 estará dedicada exclusivamente al estudio de las propuestas de Maastricht.

Sin perjuicio de que estos sistemas de análisis sean de utilidad para otros ámbitos (v.g. administrativos), hay una razón muy actual para presentar esta metodología en un campo a priori no cooperativo como es el de la Defensa

nacional. Como se ha sugerido al principio, el mundo parece tender, tras la II Guerra Mundial, a una estabilización pacífica global. El fin de la guerra fría, con la descomposición de la Unión Soviética, parece mucho más trascendente que la existencia de conflictos locales que probablemente perdurarán aún cierto tiempo. Pero los objetivos a corto plazo de nuestra Defensa se centran más bien en asegurarnos una representatividad en los organismos en los que estamos integrados acorde con nuestra posición en el concierto de las naciones. Por ahora, una fracción importante de la actividad de este Departamento ministerial se llevará a cabo dentro de instituciones que tomarán decisiones en un ambiente generalizado de cooperación; parece, pues, importante disponer de una capacidad de análisis de nuestra posición en las instancias políticas supranacionales que colabore a fundamentar nuestra política de Defensa en tiempo de paz.

## 2. ÓRGANOS DE DECISIÓN Y MEDIDAS DE PODER

Se introducen brevemente en esta Sección conceptos y técnicas de la teoría de juegos cooperativos que son de interés para representar y analizar mecanismos de decisión colectiva. Por una parte los juegos de mayoría ponderada y los juegos simples como estructuras básicas; por otra el índice de Shapley-Shubik y el valor coalicional de Owen como medidas canónicas del poder (un desarrollo sistemático puede consultarse en Owen (1982) y Roth (1988)).

### 2.1. Juegos de mayoría ponderada y juegos simples

Un juego de mayoría ponderada consta de un conjunto de agentes o jugadores (accionistas, electores, partidos, naciones, según el contexto) denotado en abstracto por

$$N = \{1, 2, \ldots, n\}$$

a los que se ha asignado una familia de pesos no negativos

$$w_1, w_2, \ldots, w_n \geq 0$$

cuya suma se designará por $T$, y una cuota o condición de mayoría $q > T/2$. La representación abreviada de este tipo de juego es

$$[q;\ w_1, w_2, \ldots, w_n]\,.$$

Un caso teóricamente frecuente es aquél donde todos los pesos son iguales a 1. Por otra parte, también es frecuente que la cuota sea "la mitad más uno", es decir, $q = q_0 = 1 + \mathrm{int}(T/2)$.

Cualquier subconjunto de jugadores se denomina una coalición, actúen o no mancomunadamente. Se dice que $C$ es una coalición vencedora si su peso acumulado es $w(C) \geq q$.

En el ejemplo numérico [5; 4,2,1,1,1] lo son:

12; 13; 14; 15; 123; 124; 125; 134; 135; 145;
y todas las de cuatro o cinco jugadores.

La propiedad básica de esta lista de coaliciones es que si una coalición es vencedora cualquiera que la contenga también lo es. Por tanto basta conservar la colección de las coaliciones vencedoras minimales —por la inclusión— para regenerar el juego. En el ejemplo anterior son:

12; 13; 14; 15; 2345.

El par formado por el conjunto $N$ y la colección $W$ de coaliciones vencedoras (con la propiedad descrita) es un juego simple, que en situaciones como la expuesta se dice asociado al juego de mayoría ponderada original. Se denota por $W^m$ la colección de las coaliciones vencedoras minimales.

Nuestro ejemplo goza de una propiedad especial: cada dos coaliciones vencedoras tienen algún jugador en común, y ello es debido a que $q > T/2$.

Otros pesos y otra mayoría darían lugar al mismo juego simple, por ejemplo [4; 3, 1, 1, 1, 1]. Este segundo juego de mayoría ponderada es más sencillo que el original pero estratégicamente equivalente a él, ya que las coaliciones vencedoras son las mismas.

No todo juego simple puede ser representado mediante un juego de mayoría ponderada. Los problemas de existencia de representaciones y de representaciones minimales son interesantes. Por otra parte, todo juego de mayoría ponderada admite una representación equivalente con pesos y cuota naturales.

### 2.2. El índice de poder de Shapley-Shubik

Así se denomina el valor Shapley, definido para juegos cooperativos en general (cf. Shapley (1953)), cuando se restringe a juegos simples. Según el modelo de regateo propuesto en Shapley (1953) y en Shapley y Shubik (1954), cada jugador recibe el valor esperado de su contribución marginal a la formación de la coalición total $N$ bajo la hipótesis de equiprobabilidad de todos los órdenes de formación posibles. Es conveniente añadir, no obstante, que Shapley planteó el problema de evaluar un juego mediante una distribución de pagos a los jugadores, propuso tres postulados razonables y a la vez simples que debía satisfacer tal

evaluación (eficiencia, simetría y aditividad) y demostró la existencia y unicidad de una regla de asignación, la que hoy conocemos como valor Shapley: éste es un procedimiento extremadamente elegante desde el punto de vista matemático.

En el caso del índice de Shapley-Shubik, el poder de cada jugador es una fracción comprendida entre 0 y 1, que a menudo se expresa como porcentaje ya que por el axioma de eficiencia la suma de los poderes de todos los jugadores es 1 (el 100%).

Si un jugador no interviene en ninguna coalición vencedora minimal —y por tanto es suprimible a efectos estratégicos de cualquier coalición vencedora— se dice que es nulo; equivalentemente, su poder es 0. En el otro extremo, si un jugador $i$ forma por sí solo una coalición vencedora (por ejemplo si $w_i \geq q$ en un juego de mayoría ponderada) se dice que es dictador; equivalentemente, su poder es 1. Por otra parte, el índice muestra una alta sensibilidad que, en su calidad de generalización a situaciones más complejas, hereda el valor coalicional de Owen.

El siguiente ejemplo ilustra aquella afirmación. En él se hace uso del teorema de la representación normalizada, que requiere una breve presentación: en todo juego de mayoría ponderada, a mayor peso corresponde mayor poder, pero es posible (véase el ejemplo propuesto en 2.1) que jugadores de distinto peso tengan igual poder; se dice que un juego de mayoría ponderada es una representación normalizada si igual peso equivale a igual poder. El teorema mencionado (véase Carreras (1989)) demuestra que para todo juego simple representable existen representaciones (naturales y) normalizadas.

### 2.3. El Consejo de Seguridad de las Naciones Unidas

El interés de este ejemplo radica en que por su definición formal se trata de un juego simple, aunque pueda hallarse una representación del mismo como juego de mayoría ponderada. La introducción de pesos pondrá ya de manifiesto de manera cuantitativa el desequilibrio de posiciones observado cualitativamente, pero el índice de Shapley-Shubik determinará con precisión el alcance de este desequilibrio. Asimismo permitirá analizar los efectos de una modificación estructural.

Hasta 1966 formaban el Consejo de Seguridad cinco naciones permanentes y otras seis escogidas por turno rotatorio de la Asamblea General. Cada uno de los once representantes disponía de un voto, pero las cinco naciones permanentes tenían derecho a veto, de forma que los siete votos que requería la aprobación de cualquier resolución debían incluir ineludiblemente los de las cinco naciones permanentes.

El juego simple que describe esta situación consta de un conjunto $N$ de 11 jugadores, cinco del tipo $A$ (permanente) y seis del tipo $B$ (eventual). El único modelo de coalición vencedora minimal es $5A + 2B$, mientras que los modelos de coaliciones perdedoras maximales son $5A + B$ y $4A + 6B$. Para buscar una representación por pesos hay que dar valores a las incógnitas $A$ y $B$ de forma que se verifiquen las desigualdades

$$5A + B < 5A + 2B, \qquad 4A + 6B < 5A + 2B;$$

la primera es obvia y la segunda se reduce a $4B < A$, de modo que la solución más sencilla es el juego

$$[27;\ 5, 5, 5, 5, 5, 1, 1, 1, 1, 1, 1]$$

cuya cuota resulta del modelo de coaliciones vencedoras minimales. Si esto ya muestra la importancia del veto concedido a cinco miembros, el índice de poder marca la diferencia exacta, dando estos valores:

| | |
|---|---|
| Nación permanente | 19.74% |
| Nación rotatoria | 0.22% |

Así, cada nación fija tiene un poder 90 veces superior al de cada rotatoria (éste es el efecto del veto). Juntas, las cinco naciones permanentes acaparan el 98.70% del poder.

La revisión de 1966 substituyó a China Nacionalista (expulsada incluso de la Organización) por China Comunista, aumentó a diez el número de naciones rotatorias y pasó a exigir nueve (incluyendo a los cinco vetos) de los quince votos como condición de aprobación. Un breve estudio similar al anterior daría ahora esta representación por pesos:

$$[39;\ 7, 7, 7, 7, 7, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1].$$

El índice de Shapley-Shubik da para este juego:

| | |
|---|---|
| Nación permanente | 19.63% |
| Nación rotatoria | 0.19% |

Es despreciable la pérdida de poder individual, pero ahora cada nación fija tiene 103 veces el poder de una eventual, y juntas las cinco grandes potencias continúan reuniendo un 98.15% del poder, aunque actualmente sean diez los restantes miembros del Consejo.

## 2.4. Estructuras de coaliciones: el valor coalicional de Owen

La ausencia de un miembro con mayoría absoluta en un organismo de decisión colectiva provoca inmediatamente negociaciones entre jugadores buscando la formación de una coalición. El objetivo se centra en las coaliciones vencedoras, las cuales, pese a soportar cierto lastre de dispersión ideológica, dispondrán de un notable grado de estabilidad por su misma condición. Salvo situaciones excepcionales, las coaliciones vencedoras formadas resultan ser minimales al menos por dos razones: para reducir al máximo las diferencias de criterio y para repartir los beneficios con la mayor eficiencia.

La formación de una coalición, o de varias disjuntas simultáneamente, provoca una alteración del juego. El efecto interno —para los miembros de la coalición— y externo —para los ajenos— requiere una medida numérica con un significado semejante al índice de Shapley-Shubik.

Necesitaríamos conocer el poder de la coalición como superjugador, frente a los restantes jugadores o coaliciones que se hayan formado simultáneamente. También el poder que acaban por tener aquellos jugadores que no se han integrado en ninguna coalición formada. Y, por último, la manera de distribuir entre sus miembros el poder asignado a cada coalición formada. La respuesta a las dos primeras cuestiones se obtiene computando el índice de Shapley-Shubik del llamado juego cociente, cuya construcción no detallaremos porque también resulta del cálculo del valor coalicional.

Pero la tercera cuestión requiere tener en cuenta la negociación intracoalicional condicionada por el exterior, ya que los jugadores argumentan, ante la coalición en la que se han integrado, según las posibilidades estratégicas a las que han renunciado fuera de ella. Owen (1977) proporciona una fundamentación axiomática (eficiencia, simetría, aditividad y simetría en el cociente) del problema entero, siguiendo una vía paralela aunque a un nivel de complejidad superior a la de Shapley (1953), demostrando la existencia y unicidad del valor coalicional asociado a un juego cooperativo bajo una estructura de coaliciones y conservando la elegancia matemática del artículo de Shapley. Aplicado a juegos simples nos da respuesta a las tres preguntas anteriores.

El valor coalicional generaliza el valor Shapley (y por tanto el índice de Shapley-Shubik), ya que coincide con él cuando la estructura de coaliciones es trivial en uno de estos dos sentidos: o no se han formado coaliciones o se ha formado la coalición total. Es aplicable incluso cuando las coaliciones en que se ha estructurado el conjunto de jugadores se subdividen a su vez en familias, y éstas en clanes... Por último, permite una elaboración semejante a la del índice

de poder de Shapley-Shubik, como modelo de regateo para la formación de la coalición total $N$ bajo condiciones restringidas.

El valor coalicional es una herramienta de alta calidad para el estudio de la dinámica de un juego simple, puesto que permite evaluar todas las coaliciones y estructuras de coaliciones posibles, ya desde el punto de vista (digamos egoísta) de cada jugador, que intenta optimizar el beneficio de su participación en uniones, ya desde una posición exterior (arbitral) más interesada tal vez en la estabilidad de las posibles coaliciones.

A modo de ilustración puede consultarse un estudio del Parlamento de Catalunya durante la Legislatura 1980–1984 (Carreras y Owen (1988)) donde se tienen en cuenta ciertas afinidades políticas entre los partidos en liza.

Las afinidades o incompatibilidades ideológicas, que a menudo existen y sin duda condicionan la acción de los jugadores, deben ser consideradas a priori informaciones ajenas al juego: su incorporación modifica la estructura y por tanto el índice de poder asociado. Los problemas teóricos que esto plantea son interesantes (cf. Carreras (1991)), así como su aplicación a problemas reales. Un tratamiento basado en incompatibilidades políticas de otro Parlamento regional español puede consultarse en Carreras y Owen (1992).

Haremos uso del valor coalicional en la siguiente sección cada vez que debamos evaluar alianzas. Al estar restringido el estudio a juegos simples, el axioma de eficiencia permite también expresarlo como porcentaje.

## 3. LAS PROPUESTAS DE MAASTRICHT

La Comunidad Europea (CE) ha venido ocupándose, desde su fundación, de organizar una política económica común para los países que la forman. Sus órganos principales son el Parlamento Europeo, con sede en Estrasburgo, la Comisión Europea, el Consejo de Ministros y el Consejo Europeo. Más adelante se describen sus estructuras.

En la pasada "cumbre" celebrada en Maastricht (Holanda) los días 9 y 10 de diciembre de 1991, el Consejo Europeo, formado por los jefes de Estado o de Gobierno de los doce países miembros y el Presidente de la Comisión Europea Jacques Delors, se enfrentaba a unas propuestas de actuación que suponían un pronunciado salto adelante en cuanto a los objetivos y competencias de la Comunidad como entidad supranacional. El proyecto consistía en dos tratados,

uno sobre unión económica y monetaria y el otro sobre unión política. Entre ambos contenían cinco grandes propuestas: creación de un Banco central y de una moneda única, colegislación por parte del Parlamento Europeo y responsabilidades de la CE en materia de Seguridad y Defensa y Política Exterior. Junto a la cohesión económica y política que esto supondría se reclamaba por parte de algunos países —entre ellos España— la inclusión de un principio de cohesión social.

Las posiciones de los Doce ante esta variedad de temas presentaban serias divergencias. Salvo el Parlamento, los demás organismos de la Comunidad, en particular el Consejo Europeo, toman sus decisiones por unanimidad, de modo que cualquier país puede vetar medidas que considere contrarias a sus intereses. Esta norma, que Buchanan y Tullock llaman "la regla de decisión ideal", es la menos decisiva de todas, aunque tiene la virtud de provocar discusiones, negociaciones y modificaciones transaccionales de los proyectos que, de llevar a un acuerdo, la hacen en efecto la solución perfecta para las elecciones colectivas. De hecho la CE ha alcanzado niveles aceptables de cooperación haciendo uso de la misma.

Sin embargo, el propio Delors considera poco operativa la ley de unanimidad o consenso. Una parte de las propuestas de Maastricht se refería a los procedimientos de decisión que deberían sustituirla. Delors criticaba incluso estos nuevos procedimientos —que después estudiaremos con detalle— por poco ágiles, especialmente en materia de Política Exterior y Seguridad Común (PESC): "el mecanismo de toma de decisiones bloqueará cualquier iniciativa". El comisario español Abel Matutes, cuyas competencias como tal permitirían asimilarle a un "ministro europeo de Relaciones Exteriores", afirmaba días antes que Europa era "un gigante económico y un enano político", subrayando el pobre papel realizado durante la Guerra del Golfo y ante el conflicto interno de Yugoeslavia; señalaba la necesidad de que la CE tuviese capacidad en Política Exterior y por tanto en materia de Seguridad y Defensa, indicando que unión política significa ante todo política exterior común.

Determinados países consideran un atentado a su soberanía nacional cualquier intento por parte de la CE de legislar y actuar en algunas materias, especialmente si ya no se requiere unanimidad para su aprobación. Gran Bretaña, por ejemplo, se opone a perder su dominio absoluto sobre las leyes laborales, mientras España rechaza injerencias en cuestiones de medio ambiente, redes de comunicación o investigación y desarrollo; Alemania se opone a la cohesión social argumentando que el país sufre la incorporación de la antigua República Democrática y no puede contribuir aún más a los fondos comunitarios; por su parte, la típica política exterior presidencialista de Francia difícilmente encaja en las líneas del plan PESC, aunque ya se ha formado la brigada franco-alemana,

germen de un futuro ejército con mando europeo, el punto más ambicioso del proyecto de unión política.

La existencia de la OTAN, con la presencia de los Estados Unidos detrás, y de la hasta ahora poco operativa Unión Europea Occidental (UEO), a las que pertenecen la mayoría de los países de la CE, complica aún más si cabe el entramado de opiniones encontradas en el seno de la Comunidad.

Latente en el fondo de esta compleja situación está la "metadecisión" acerca de los procedimientos de decisión que han de regular la acción de la CE en estas cuestiones, respetando en lo posible la soberanía de cada Estado miembro y asegurando al mismo tiempo la máxima operatividad. En adelante nos centraremos en este delicado problema, analizando las diversas propuestas presentadas para describir, bajo cada hipótesis, la posición estratégica de cada país y, en especial, el papel de España, dando una respuesta cuantificada, y por tanto objetiva, a esta cuestión.

**Unanimidad o consenso:** Las decisiones adaptadas mediante esta regla admiten solamente una coalición vencedora, el conjunto total $N$. Por simetría, el índice de Shapley-Shubik es constante, y asigna a cada uno de los doce jugadores (países) un poder de

$$1/12 = 8.33\%.$$

Adelantamos este resultado en reconocimiento al procedimiento de votación distintivo de la CE —al menos hasta ahora—, señalando que se trata de un valor ceñido exclusivamente a la estructura formal; no contempla por tanto influencias y presiones sin duda existentes de los países principales sobre algunos de los demás. En cualquier caso puede ser utilizado como punto de referencia.

### 3.1. El Parlamento Europeo

El Parlamento Europeo está formado por 518 representantes de los 12 países (se prevé extenderlos a 536 por la reunificación alemana), elegidos internamente en cada uno. Existe una partición transversal en grupos parlamentarios, que se constituyen según afinidades y homologaciones políticas y reflejan con fidelidad las estructuras ideológicas internas de las diferentes naciones. Ambas divisiones —por países y por grupos— influyen poderosamente en las actitudes de los representantes, y en cierta forma esta bidimensionalidad favorece la disensión en lugar del consenso, por lo que no es sorprendente que el Parlamento tome sus decisiones habitualmente por mayoría simple, es decir, mediante la cuota $q = q_0 = 260$.

Sin embargo, aunque la nacionalidad no sea el distintivo excluyente de cada miembro de la Cámara, a efectos de comparar la posición estratégica de cada país en ésta y en las otras instituciones comunitarias debemos computar el poder como si el Parlamento fuese un juego de mayoría ponderada jugado por las 12 naciones. En la Tabla 1 (véase Apéndice) se describe la distribución de escaños en términos absolutos y relativos y el índice de Shapley-Shubik para $q = 260$ (y también para $q = 368$; este segundo valor corresponde a la proporción 54/76 que es la propuesta básica de mayoría cualificada presentada en Maastricht para el Consejo de Ministros y que después valoraremos).

Para $q = 260$, el poder de los cuatros países principales y de España es ligeramente superior a su porcentaje de escaños, mientras que para los países intermedios y menores es ligeramente inferior. No hay jugadores nulos ni, obviamente, jugadores con veto, aunque sí coaliciones de bloqueo más o menos plausibles políticamente, como la formada por tres naciones principales y Dinamarca, o por dos naciones principales, España y los tres países con menor representación; de todas formas, el mínimo margen de votos permitido a tales coaliciones (exactamente 259) y el alto valor del total $T = 518$ las despojan de todo interés práctico. Globalmente, la distribución de poder en el Parlamento Europeo no muestra ninguna desviación notable respecto a la distribución de escaños —que por su parte no corresponde excesivamente a las poblaciones de los países—.

Al elevar la mayoría a 368 (superior incluso a los 2/3, que es una mayoría cualificada habitual en otras Cámaras para cuestiones de gran trascendencia) los cuatros países principales, Holanda y Luxemburgo aumentan su poder, mientras España y los restantes se ven perjudicados, aunque en ningún caso sea destacable la variación.

El último dato que aporta la Tabla 1 es el máximo poder que esperarían alcanzar los cuatro países principales y España. El máximo poder asequible para un jugador $i$ en un juego de mayoría ponderada es su poder de veto, que alcanza en su punto de veto definido por la cuota

$$q_i = T - w_i + 1.$$

En este momento el jugador $i$ comienza a tener veto, lo comparte con los jugadores de peso superior —que lo han alcanzado para un valor anterior de la mayoría— y su poder es máximo. Las dos primeras condiciones se mantienen hasta el final —cuando se exige unanimidad— pero el poder del jugador decrece invariablemente hasta $1/n$.

En nuestro caso se observa una apreciable diferencia entre el poder en mayoría simple y el máximo para los cinco países. Además, para $q = 438$ los cuatro

grandes controlarían el 81.80% del poder, y para $q = 459$ acapararían, con España, el 88.90% (cf. el apartado 2.3).

### 3.2. La Comisión Europea y el Consejo de Ministros

La Comisión Europea es el órgano ejecutivo de la CE, sometido al control del Parlamento. Aunque se pretende que esté constituída por un único representante de cada país, su composición actual se refleja en la Tabla 2 (Apéndice), acompañada de la distribución de poder que daría la mayoría simple, en este caso $q = 9$.

Se aprecia un leve sesgo de la distribución de poder respecto a la de escaños en favor de los cinco países principales. Y sobre todo la imposibilidad de formar coaliciones de bloqueo al ser $T = 17$ (impar) y $q = q_0 = 1 + \text{int}(T/2)$. Sin embargo, la Comisión toma sus decisiones por unanimidad o consenso, de forma que cada país dispone de veto y el poder de cualquiera es igual a 8.33%.

El Consejo de Ministros está compuesto por un representante de cada uno de los Estados que forman la Comunidad. En realidad, la agenda de cada reunión es homogénea, de forma que son convocados los titulares respectivos de una misma (o asimilable) cartera ministerial. Puesto que hasta ahora ha venido rigiéndose por la regla de unanimidad, posturas como la de Gran Bretaña han obstaculizado numerosas iniciativas tendentes a una mayor integración europea.

Por otra parte en el Consejo están representados de hecho los Gobiernos de las naciones. Hay homologaciones políticas entre las ideologías y, aunque en principio son razones de Estado las que diseñan la posición de cada miembro, no puede desdeñarse la influencia del pensamiento político de un partido en la forma de entender esas razones de Estado el gobierno al que sustenta.

Con esto concluye la descripción de las instancias básicas de la CE, que tomaremos como referencia, y pasamos a considerar las alternativas propuestas en Maastricht para agilizar el funcionamiento de este Consejo de Ministros.

### 3.3. El modelo de mayoría cualificada

La propuesta consistía en asignar a cada país un peso en el Consejo de Ministros y exigir una mayoría cualificada, a medio camino entre la mayoría simple —que rige por ejemplo en el Parlamento Europeo— y la unanimidad requerida hasta ahora en el propio Consejo. Una versión más prudente para el plan PESC sugería la aprobación por unanimidad de la decisión de someter un tema

a consideración y, ello supuesto, la aprobación por mayoría cualificada de la necesidad o tipo de intervención. Finalmente, la solución que se esperaba sería de compromiso era decidir por unanimidad qué temas admitirían aprobación de resoluciones votando por mayoría cualificada. Nuestro análisis mostrará a continuación que cualquier mayoría suficientemente alejada de la unanimidad concedería prácticamente idénticas posiciones estratégicas a los distintos países.

En la Tabla 3 (Apéndice) se muestra la distribución de pesos y su porcentaje. Al comparar con la estructura del Parlamento Europeo (Tabla 1) notamos que se han suprimido mínimas diferencias entre Holanda y los siguientes y entre Dinamarca e Irlanda, pero sobre todo que los cinco primeros países salen perjudicados en beneficio de los siete restantes. La tercera columna de la Tabla 3 señala la distribución de poder bajo mayoría simple, y es notable su aproximación a la columna anterior, de forma que también en cuanto a poder la nueva estructura del Consejo mejora las posiciones de los siete países menores y debilita las de los cinco principales con respecto al Parlamento Europeo. Además, la representación resulta ser normalizada, es decir, pesos distintos tienen también distinto poder.

Por su parte, la cuarta columna, que corresponde a la verdadera propuesta del proyecto ($q = 54$) admite los mismos comentarios que la tercera respecto al Parlamento Europeo en las dos versiones de éste presentadas en la Tabla 1, con mayorías $q = 260$ (simple) y $q = 368$ ($368/518 = 54/76$). En cambio, comparada con la columna anterior nos da un ligero retroceso de los cuatro países dominantes y de Luxemburgo compensado con un también pequeño efecto positivo para España y los demás países.

Completando el análisis de este juego,

$$[54;\ 10, 10, 10, 10, 8, 5, 5, 5, 5, 3, 3, 2]$$

y designando por $a, b, c, d, e$ respectivamente los cinco tipos de jugadores según su peso, encontramos 14 modelos de coaliciones vencedoras minimales, que se describen en la Tabla 4. Hay por otra parte gran cantidad de coaliciones de bloqueo, por ejemplo cualquiera formada por seis países, o incluso cualquiera de cinco que contenga al menos un país principal o las de cuatro que incluyan al menos dos países principales.

Una variante del modelo de mayoría cualificada proponía exigir para la aprobación de una medida los 54 votos junto con la concurrencia de al menos 8 países para obtenerlos. Este es un juego simple, cuyos 18 modelos de coaliciones vencedoras minimales se indican en la Tabla 5; entre ellos están los nueve últimos de la tabla anterior y otros nueve que antes eran sólo coaliciones vencedoras y

ahora son minimales. Todo ello resulta de comparar las colecciones respectivas $W$ y $W'$ de coaliciones vencedoras:

$$W' = \{S \in W / s \geq 8\},$$

de donde $(W')^m$ coincide con la unión

$$\{S \in W^m / s \geq 8\} \cup \{S \cup \{i\} / S \in W^m,\ s = 7,\ i \in N - S\},$$

donde $s = |S|$ es el cardinal de la coalición variable $S$. Obviamente hay muchas más coaliciones de bloqueo en este juego, ya que a las del anterior se añaden como mínimo las descritas por los cinco modelos de la Tabla 4 que han dejado de ser coaliciones vencedoras en la variante.

La distribución de poder para esta variante se encuentra en la última columna de la Tabla 3, observándose diferencias muy poco importantes con la columna anterior: ligera disminución de los cinco primeros y ligero incremento para los siete restantes.

Volviendo al modelo original de mayoría cualificada estudiaremos la variación de la distribución de poder al cambiar la mayoría desde su primer valor $q = q_0 = 39$ hasta la unanimidad $q = 76$. Los resultados numéricos se dan en las Tablas 6 y 7, y las gráficas correspondientes aparecen en las Figuras 1 y 2, todo ello en el Apéndice, donde asimismo la Figura 3 muestra la gráfica total aunque con incrementos de dos unidades para la cuota: en los tres casos cada curva describe el poder de un tipo de jugador. El número de coaliciones vencedoras minimales según la mayoría sigue la tendencia que sugiere esta tabla:

| $q$ | 39 | 47 | 48 | 56 | 57 | 65 | 66 | 74 | 76 |
|---|---|---|---|---|---|---|---|---|---|
| cvm | 363 | 252 | 232 | 104 | 93 | 26 | 20 | 1 | 1 |

y debe tenerse en cuenta que la disminución conlleva un aumento casi equivalente de coaliciones de bloqueo, y en particular de jugadores con veto al pasar por los respectivos puntos de veto.

Hay más oscilaciones durante la segunda mitad del recorrido (ver Figura 2). El poder máximo que alcanzarían los cuatro países principales sería del 16.72%, y el de España un 14.70%. Esto confirma, junto con las columnas 3ª y 4ª de la Tabla 3, que no hay variaciones demasiado sorprendentes lejos de la unanimidad. La peor situación de España, en cambio, se halla en el 7.63% justo antes de alcanzar veto y máximo poder, y justo en el momento en que los cuatro países que le preceden alcanzan su propio veto y poder óptimo.

### 3.4. Análisis de alianzas

La regularidad de la distribución de poder en el modelo de mayoría cualificada no parece esconder inesperadas variaciones bajo la formación efectiva de coaliciones. Sin pretender desarrollar un estudio exhaustivo, destacaremos algunas situaciones elementales para poner de manifiesto la tendencia genérica de los resultados.

Las Tablas 8a y 8b recogen la distribución individualizada de poder bajo alianzas que en ningún caso alcanzan la mayoría exigida pero en principio han de mejorar la posición estratégica de sus integrantes. Ello es, en efecto, así. Las columnas (1)–(5) corresponden a la formación de coaliciones que podríamos denominar "motores" de la integración europea: es destacable para España que sería mejor una coalición paritaria con Alemania y Francia —columna (3)— que integrarse a posteriori como tercer socio —columna (4)—.

En general, las naciones no incluídas en la coalición que se forma resultan perjudicadas, excepción hecha de Luxemburgo en (1), (2), (5), (6) y (7) y Holanda, Bélgica y Dinamarca —que las alcanza en poder— en (7).

La columna (6) describe la situación en que los cuatro países dominantes se unen, y el resultado no implica ninguna variación importante respecto a la posición inicial, aunque la coalición controle el 59.48% del poder. La columna (7) corresponde a la alianza de los cuatro países menos fuertes económicamente, y su efecto es igualmente leve.

Finalmente, la Tabla 9 presenta el efecto de la fragmentación que algunas cuestiones originan en el modelo discutido. La cuestión de la cohesión social agrupa solamente a los cuatro países menos ricos en su favor y al contribuyente principal, Alemania, y el país más crítico ante los gastos comunitarios, Gran Bretaña, en contra. El juego cociente presenta cierta complejidad; los países coaligados en uno u otro sentido aumentan su poder —comparando con el poder desligado, suma de los poderes de los miembros de la coalición en el juego de partida—; los dos países principales no comprometidos pierden calidad estratégica y los restantes experimentan variaciones insignificantes.

El debate entre los partidarios de convertir la UEO en el órgano adecuado para poner en marcha el plan PESC y los que prefieren seguir otorgando a la OTAN el papel principal en la Defensa de la Europa comunitaria polariza radicalmente el modelo, como pone de manifiesto la distribución de poder. Es difícil llegar a un acuerdo en estas condiciones, que son consecuencia en parte de las reglas de votación y en parte del decidido enfrentamiento entre las dos posturas. Al declararse neutral, Irlanda pierde todo su poder, y el beneficio principal se lo llevan los atlantistas.

Por último, una perspectiva de las ideologías actualmente presentes en el Consejo nos muestra una situación de claro predominio de los democristianos y afines: sin llegar a disponer de la mayoría requerida forman una coalición de bloqueo; el juego entre las tres líneas de pensamiento es de veto radical, ya que las coaliciones vencedoras minimales son

Democristianos + Conservadores
Democristianos + Socialistas.

### 3.5. Conclusión

Los resultados de la cumbre de Maastricht son suficientemente conocidos. Se ha llegado a ciertos acuerdos después de duras negociaciones que han diluído el contenido de los proyectos. Los europeístas consideran que se ha avanzado, aunque sea con la lentitud característica de la Comunidad. Gran Bretaña aparece como el vencedor, al haber conseguido mantener sus condiciones sin cesiones no deseadas. España, por su parte, ha logrado incluir una idea de cohesión social en la redacción final.

En lo político, el adelanto en el plan de Seguridad y Defensa común ha sido mínimo. La sombra de la OTAN planea todavía con fuerza sobre la Europa comunitaria. Las decisiones por mayoría cualificada han salido vencidas por ahora y deberán esperar tiempos mejores para ser incorporadas al procedimiento. El estudio que aquí se ha presentado conservará su validez, pero hoy por hoy la CE sigue rigiéndose por unanimidad o, al menos, consenso. El juego, de momento, ha terminado.

## APÉNDICE: TABLAS Y GRÁFICAS

**Tabla 1**

Estructura por naciones del Parlamento Europeo

| País | escaños | % esc. | poder $q = 260$ | poder $q = 368$ |
|---|---|---|---|---|
| Alemania | 81 | 15.64 | 16.59 | 16.98 |
| Francia | 81 | 15.64 | 16.59 | 16.98 |
| Italia | 81 | 15.64 | 16.59 | 16.98 |
| Gran Bretaña | 81 | 15.64 | 16.59 | 16.98 |
| España | 60 | 11.58 | 12.62 | 11.96 |
| Holanda | 25 | 4.83 | 3.83 | 3.96 |
| Bélgica | 24 | 4.63 | 3.83 | 3.61 |
| Portugal | 24 | 4.63 | 3.83 | 3.61 |
| Grecia | 24 | 4.63 | 3.83 | 3.61 |
| Dinamarca | 16 | 3.09 | 2.79 | 2.38 |
| Irlanda | 15 | 2.90 | 2.49 | 1.86 |
| Luxemburgo | 6 | 1.16 | 0.41 | 1.05 |
| TOTAL | 518 | | | |
| coaliciones vencedoras minimales | | | 167 | 99 |

Alemania, Francia, Italia y Gran Bretaña:
máximo poder 20.45% en el punto de veto $q = 438$

España:
máximo poder 17.78% en el punto de veto $q = 459$.

**Tabla 2**

Comisión Europea

| País | repres. | poder si $q = 9$ |
|---|---|---|
| Alemania | 2 | 12.07 |
| Francia | 2 | 12.07 |
| Italia | 2 | 12.07 |
| Gran Bretaña | 2 | 12.07 |
| España | 2 | 12.07 |
| Holanda | 1 | 5.66 |
| Bélgica | 1 | 5.66 |
| Portugal | 1 | 5.66 |
| Grecia | 1 | 5.66 |
| Dinamarca | 1 | 5.66 |
| Irlanda | 1 | 5.66 |
| Luxemburgo | 1 | 5.66 |
| TOTAL | 17 | |
| coaliciones vencedoras minimales | | 601 |

**Tabla 3**

La modificación del Consejo de Ministros

| País | peso | % peso | poder $q = 39$ | poder $q = 54$ | poder[(*)] $q = 54, c \geq 8$ |
|---|---|---|---|---|---|
| Alemania | 10 | 13.16 | 13.55 | 13.42 | 13.01 |
| Francia | 10 | 13.16 | 13.55 | 13.42 | 13.01 |
| Italia | 10 | 13.16 | 13.55 | 13.42 | 13.01 |
| Gran Bretaña | 10 | 13.16 | 13.55 | 13.42 | 13.01 |
| España | 8 | 10.53 | 10.68 | 11.13 | 10.88 |
| Holanda | 5 | 6.58 | 6.31 | 6.37 | 6.57 |
| Bélgica | 5 | 6.58 | 6.31 | 6.37 | 6.57 |
| Portugal | 5 | 6.58 | 6.31 | 6.37 | 6.57 |
| Grecia | 5 | 6.58 | 6.31 | 6.37 | 6.57 |
| Dinamarca | 3 | 3.95 | 3.82 | 4.26 | 4.62 |
| Irlanda | 3 | 3.95 | 3.82 | 4.26 | 4.62 |
| Luxemburgo | 2 | 2.63 | 2.27 | 1.18 | 1.59 |
| TOTAL | 76 | | | | |
| coaliciones vencedoras minimales | | | 363 | 135 | 160 |

(*) mayoría cualificada $q = 54$ y coaliciones vencedoras con $c \geq 8$ jugadores.

**Tabla 4**

Coaliciones vencedoras minimales en el juego

[54; 10, 10, 10, 10, 8, 5, 5, 5, 5, 3, 3, 2]

| | |
|---|---|
| $4a + b + 2c$ | $3a + b + 4c$ |
| $4a + b + c + d$ | $3a + b + 3c + d$ |
| $4a + b + c + e$ | $3a + b + 3c + e$ |
| $4a + b + 2d$ | $3a + b + 2c + 2d$ |
| $4a + 3c$ | $3a + 4c + 2d$ |
| $4a + 2c + 2d$ | $3a + 4c + d + e$ |
| $4a + 2c + d + e$ | $2a + b + 4c + 2d$ |

**Tabla 5**

Coaliciones vencedoras minimales en la variante

| | |
|---|---|
| $4a + b + 3c$ | $4a + 2c + 2d$ |
| $4a + b + 2c + d$ | $4a + 2c + d + e$ |
| $4a + b + 2c + e$ | $3a + b + 4c$ |
| $4a + b + c + 2d$ | $3a + b + 3c + d$ |
| $4a + b + c + d + e$ | $3a + b + 3c + e$ |
| $4a + b + 2d + e$ | $3a + b + 2c + 2d$ |
| $4a + 4c$ | $3a + 4c + 2d$ |
| $4a + 3c + d$ | $3a + 4c + d + e$ |
| $4a + 3c + e$ | $2a + b + 4c + 2d$ |

**Tabla 6**

Variación del poder según la mayoría (1ª parte)

| | Mayorías | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 76 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
| 10 | 13.55 | 13.78 | 13.26 | 13.80 | 13.58 | 13.54 | 13.76 | 13.22 | 13.82 |
| 10 | 13.55 | 13.78 | 13.26 | 13.80 | 13.58 | 13.54 | 13.76 | 13.22 | 13.82 |
| 10 | 13.55 | 13.78 | 13.26 | 13.80 | 13.58 | 13.54 | 13.76 | 13.22 | 13.82 |
| 10 | 13.55 | 13.78 | 13.26 | 13.80 | 13.58 | 13.54 | 13.76 | 13.22 | 13.82 |
| 8 | 10.68 | 9.96 | 11.49 | 9.90 | 10.51 | 10.84 | 10.02 | 11.52 | 9.82 |
| 5 | 6.31 | 6.41 | 6.18 | 6.40 | 6.31 | 6.31 | 6.41 | 6.17 | 6.42 |
| 5 | 6.31 | 6.41 | 6.18 | 6.40 | 6.31 | 6.31 | 6.41 | 6.17 | 6.42 |
| 5 | 6.31 | 6.41 | 6.18 | 6.40 | 6.31 | 6.31 | 6.41 | 6.17 | 6.42 |
| 5 | 6.31 | 6.41 | 6.18 | 6.40 | 6.31 | 6.31 | 6.41 | 6.17 | 6.42 |
| 3 | 3.82 | 3.10 | 4.66 | 3.11 | 3.71 | 3.93 | 3.11 | 4.69 | 3.08 |
| 3 | 3.82 | 3.10 | 4.66 | 3.11 | 3.71 | 3.93 | 3.11 | 4.69 | 3.08 |
| 2 | 2.27 | 3.10 | 1.40 | 3.11 | 2.54 | 1.93 | 3.11 | 1.54 | 3.08 |

| | Mayorías | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 76 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 |
| 10 | 13.63 | 13.55 | 13.73 | 13.13 | 13.84 | 13.71 | 13.42 | 13.55 | 13.01 |
| 10 | 13.63 | 13.55 | 13.73 | 13.13 | 13.84 | 13.71 | 13.42 | 13.55 | 13.01 |
| 10 | 13.63 | 13.55 | 13.73 | 13.13 | 13.84 | 13.71 | 13.42 | 13.55 | 13.01 |
| 10 | 13.63 | 13.55 | 13.73 | 13.13 | 13.84 | 13.71 | 13.42 | 13.55 | 13.01 |
| 8 | 10.29 | 11.04 | 10.14 | 11.57 | 9.65 | 10.01 | 11.13 | 10.20 | 11.64 |
| 5 | 6.34 | 6.27 | 6.35 | 6.15 | 6.52 | 6.44 | 6.37 | 6.40 | 6.08 |
| 5 | 6.34 | 6.27 | 6.35 | 6.15 | 6.52 | 6.44 | 6.37 | 6.40 | 6.08 |
| 5 | 6.34 | 6.27 | 6.35 | 6.15 | 6.52 | 6.44 | 6.37 | 6.40 | 6.08 |
| 5 | 6.34 | 6.27 | 6.35 | 6.15 | 6.52 | 6.44 | 6.37 | 6.40 | 6.08 |
| 3 | 3.55 | 4.07 | 3.17 | 4.74 | 2.97 | 3.33 | 4.26 | 3.33 | 4.90 |
| 3 | 3.55 | 4.07 | 3.17 | 4.74 | 2.97 | 3.33 | 4.26 | 3.33 | 4.90 |
| 2 | 2.72 | 1.55 | 3.17 | 1.83 | 2.97 | 2.76 | 1.18 | 3.33 | 2.25 |

**Tabla 7**

Variación del poder según la mayoría (2ª parte)

| | Mayorías | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 76 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 |
| 10 | 14.30 | 14.20 | 13.66 | 13.56 | 12.98 | 13.48 | 13.43 | 12.63 | 12.63 |
| 10 | 14.30 | 14.20 | 13.66 | 13.56 | 12.98 | 13.48 | 13.43 | 12.63 | 12.63 |
| 10 | 14.30 | 14.20 | 13.66 | 13.56 | 12.98 | 13.48 | 13.43 | 12.63 | 12.63 |
| 10 | 14.30 | 14.20 | 13.66 | 13.56 | 12.98 | 13.48 | 13.43 | 12.63 | 12.63 |
| 8 | 9.45 | 9.65 | 11.97 | 11.06 | 11.97 | 9.24 | 9.39 | 10.81 | 9.90 |
| 5 | 6.22 | 6.19 | 6.01 | 6.14 | 5.98 | 7.35 | 7.30 | 6.87 | 6.57 |
| 5 | 6.22 | 6.19 | 6.01 | 6.14 | 5.98 | 7.35 | 7.30 | 6.87 | 6.57 |
| 5 | 6.22 | 6.19 | 6.01 | 6.14 | 5.98 | 7.35 | 7.30 | 6.87 | 6.57 |
| 5 | 6.22 | 6.19 | 6.01 | 6.14 | 5.98 | 7.35 | 7.30 | 6.87 | 6.57 |
| 3 | 2.83 | 3.03 | 4.29 | 3.38 | 4.75 | 2.47 | 2.63 | 5.35 | 4.44 |
| 3 | 2.83 | 3.03 | 4.29 | 3.38 | 4.75 | 2.47 | 2.63 | 5.35 | 4.44 |
| 2 | 2.83 | 2.73 | 0.76 | 3.38 | 2.68 | 2.47 | 2.42 | 0.51 | 4.44 |

| | Mayorías | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 76 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 |
| 10 | 12.12 | 16.72 | 16.72 | 14.70 | 14.09 | 13.94 | 10.61 | 10.61 | 9.09 |
| 10 | 12.12 | 16.72 | 16.72 | 14.70 | 14.09 | 13.94 | 10.61 | 10.61 | 9.09 |
| 10 | 12.12 | 16.72 | 16.72 | 14.70 | 14.09 | 13.94 | 10.61 | 10.61 | 9.09 |
| 10 | 12.12 | 16.72 | 16.72 | 14.70 | 14.09 | 13.94 | 10.61 | 10.61 | 9.09 |
| 8 | 11.21 | 7.63 | 7.63 | 14.70 | 14.09 | 13.94 | 10.61 | 10.61 | 9.09 |
| 5 | 6.26 | 4.90 | 4.90 | 4.70 | 5.00 | 4.85 | 10.61 | 10.61 | 9.09 |
| 5 | 6.26 | 4.90 | 4.90 | 4.70 | 5.00 | 4.85 | 10.61 | 10.61 | 9.09 |
| 5 | 6.26 | 4.90 | 4.90 | 4.70 | 5.00 | 4.85 | 10.61 | 10.61 | 9.09 |
| 5 | 6.26 | 4.90 | 4.90 | 4.70 | 5.00 | 4.85 | 10.61 | 10.61 | 9.09 |
| 3 | 5.66 | 1.97 | 1.97 | 3.79 | 3.18 | 3.94 | 1.52 | 1.52 | 9.09 |
| 3 | 5.66 | 1.97 | 1.97 | 3.79 | 3.18 | 3.94 | 1.52 | 1.52 | 9.09 |
| 2 | 3.94 | 1.97 | 1.97 | 0.15 | 3.18 | 3.03 | 1.52 | 1.52 | 0.00 |

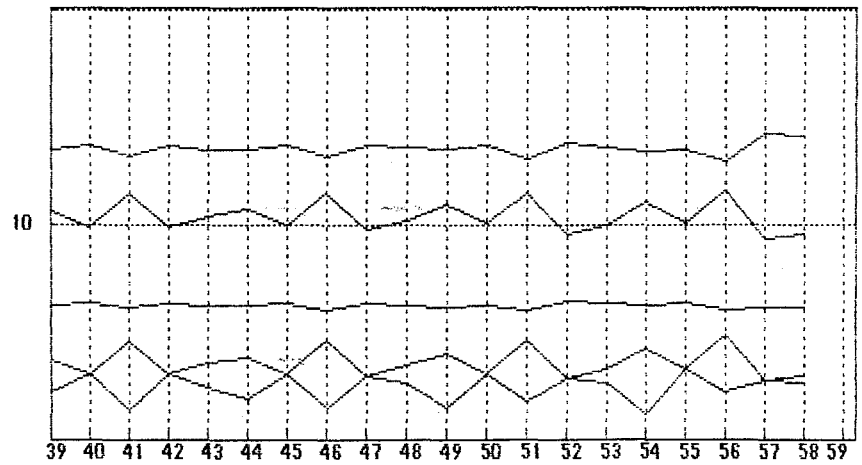



**Figura 1**

Variación del poder según la mayoría (1ª parte)

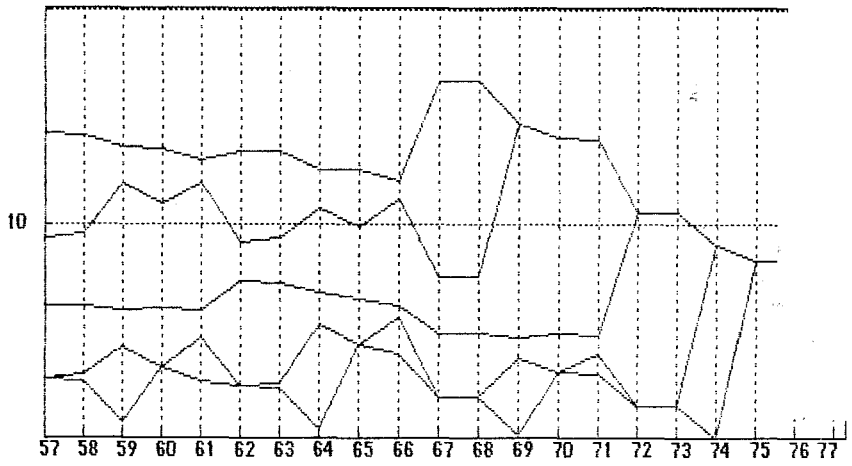



**Figura 2**

Variación del poder según la mayoría (2ª parte)

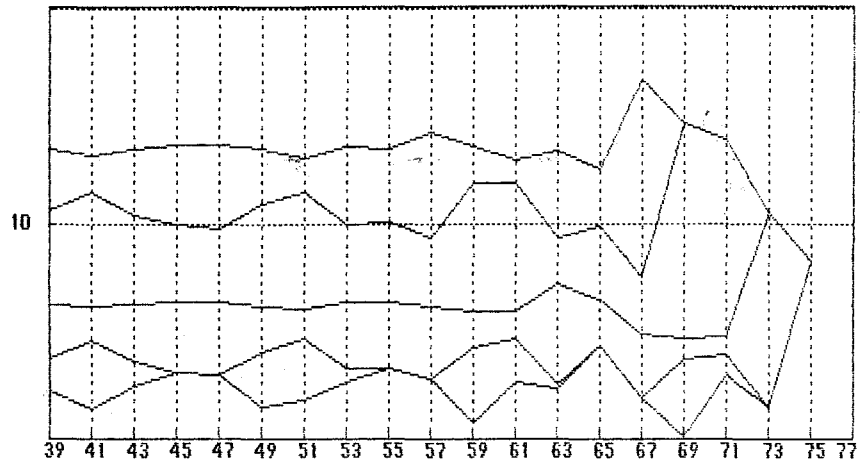



**Figura 3**

Variación del poder según la mayoría (global)

**Tabla 8a**

Efectos individuales de algunas alianzas

| País | (0) | (1) | (2) | (3) |
|---|---|---|---|---|
| Alemania | 13.42 | (15.28) | (16.03) | (16.38) |
| Francia | 13.42 | (15.28) | (16.03) | (16.38) |
| Italia | 13.42 | 12.46 | (16.03) | 11.87 |
| Gran Bretaña | 13.42 | 12.46 | 12.58 | 11.87 |
| España | 11.13 | 10.16 | 9.56 | (13.84) |
| Holanda | 6.37 | 6.15 | 5.40 | 5.24 |
| Bélgica | 6.37 | 6.15 | 5.40 | 5.24 |
| Portugal | 6.37 | 6.15 | 5.40 | 5.24 |
| Grecia | 6.37 | 6.15 | 5.40 | 5.24 |
| Dinamarca | 4.26 | 4.21 | 3.17 | 3.89 |
| Irlanda | 4.26 | 4.21 | 3.17 | 3.89 |
| Luxemburgo | 1.18 | 1.31 | 1.83 | 0.95 |

(0) estructura original sin formación de alianzas, $q = 54$.

(1) Alemania + Francia.

(2) Alemania + Francia + Italia.

(3) Alemania + Francia + España.

**Tabla 8b**

Efectos individuales de algunas alianzas

| País | (4) | (5) | (6) | (7) |
|---|---|---|---|---|
| Alemania | (16.75) | (14.75) | (14.87) | 12.10 |
| Francia | (16.75) | (14.75) | (14.87) | 12.10 |
| Italia | 11.87 | (14.75) | (14.87) | 12.10 |
| Gran Bretaña | 11.87 | 12.50 | (14.87) | 12.10 |
| España | (13.10) | (12.55) | 9.48 | (11.78) |
| Holanda | 5.24 | 5.36 | 5.56 | 6.39 |
| Bélgica | 5.24 | 5.36 | 5.56 | 6.39 |
| Portugal | 5.24 | 5.36 | 5.56 | (7.40) |
| Grecia | 5.24 | 5.36 | 5.56 | (7.40) |
| Dinamarca | 3.89 | 3.93 | 3.06 | 6.39 |
| Irlanda | 3.89 | 3.93 | 3.06 | (4.45) |
| Luxemburgo | 0.95 | 1.43 | 2.70 | 1.39 |

(4) (Alemania + Francia) + España.

(5) Alemania + Francia + Italia + España.

(6) Alemania + Francia + Italia + Gran Bretaña.

(7) España + Portugal + Grecia + Irlanda.

**Tabla 9**

Efectos globales de algunas alianzas

| **Cuestión : cohesión social** | | peso | poder | poder desligado |
|---|---|---|---|---|
| A favor: | España, Portugal, | | | |
| | Grecia, Irlanda: | 21 | 30.60 | 28.13 |
| Predispuestos: | Francia | 10 | 9.88 | 13.42 |
| | Italia | 10 | 9.88 | 13.42 |
| | Bélgica | 5 | 6.55 | 6.37 |
| Poco dispuestos: | Holanda | 5 | 6.55 | 6.37 |
| | Dinamarca | 3 | 6.55 | 4.26 |
| | Luxemburgo | 2 | 1.79 | 1.18 |
| En contra: | Alemania, Gran Bretaña: | 20 | 28.21 | 26.84 |

| **Cuestión: UEO/OTAN** | | peso | poder | poder desligado |
|---|---|---|---|---|
| UEO: | Alemania, Francia, España, | | | |
| | Bélgica, Grecia, Luxemburgo: | 40 | 50.00 | 51.89 |
| neutral: | Irlanda | 3 | 0 | 4.26 |
| OTAN: | Italia, Gran Bretaña, Holanda, | | | |
| | Portugal, Dinamarca: | 33 | 50.00 | 43.84 |

| **Cuestión: afinidad ideológica** | | peso | poder | poder desligado |
|---|---|---|---|---|
| Democristianos: | Alemania, Italia | | | |
| | Holanda, Bélgica, Portugal, | | | |
| | Grecia, Irlanda, Luxemburgo: | 45 | 66.67 | 57.76 |
| Socialistas: | Francia, España: | 18 | 16.67 | 24.55 |
| Conservadores: | Gran Bretaña, Dinamarca: | 13 | 16.67 | 17.68 |

## REFERENCIAS


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

[2] **Carreras, F.** (1989). "Normalización de juegos de mayoría ponderada". *Actas de las XIV Jornadas Hispano-Lusas de Matemáticas,* 783–787.

[3] **Carreras, F.** (1991). "Restriction of Simple Games". *Math. Soc. Sci.,* **21**, 245–260.

[4] **Carreras, F. y Owen G.** (1988). "Evaluation of the Catalonian Parliament, 1980–1984". *Math. Soc. Sci.,* **15**, 87–92.

[5] **Carreras, F. y Owen, G.** (1992). "An Analysis of the Euskarian Parliament", en *Coalition Theory and Coalition Governments.* Kluwer.

[6] **Colomer, J.M.** (1990). *El arte de la manipulación política.* Anagrama.

[7] **Neumann, J. Von y Morgenstern, O.** (1944). *Theory of Games and Economic Behavior.* Princeton U. Press.

[8] **Owen, G.** (1977). "Values of Games with a Priori Unions", en *Mathematical Economics and Game Theory.* Springer-Verlag, 76–88.

[9] **Owen, G.** (1982). *Game Theory.* Academic Press, 2ª ed.

[10] **Roth, A.E.** (1988). *The Shapley value: Essays in honor of Lloyd S. Shapley.* Cambridge U. Press.

[11] **Shapley, L.S.** (1953). "A Value for $n-$Person Games". Contributions to the Theory of Games II. *Annals of Math. Studies,* **28**, 307–317.

[12] **Shapley, L.S.** (1962). "Simple Games: An outline of the descriptive theory". *Behavioral Science,* **7**, 59–66.

[13] **Shapley, L.S. y Shubik, M.** (1954). "A method for evaluating the distribution of power in a committee system". *Am. Pol. Sci. Rev.* XLVIII, 787–792.

# ENGLISH SUMMARY:

## COOPERATION AND NATIONAL DEFENSE

Francesc Carreras

## 1. INTRODUCTION

As a remarkable fact of the period between the end of the II World War and this final decade of the 20th century, there is a growing tendency for individuals and social groups of diverse sizes and entities, including nations, to find themselves incorporated in representative organisms, ruled by voting systems.

This increases our interest in descriptive and analytical techniques to study such systems and, in particular, in methods to measure the distribution of power among the members of these decision-making structures. Section 2 is devoted to describing two measures of power, one of them at individual level and the other in terms of coalition formation, specifying the kind of structures in which they are defined. We have tried to make the reading easy, focussing on the significance and scope of the exposed concepts rather than on their mathematical formulation, which can be found in the references.

The good relationships existing between Spain and all other countries in the world suggest that an important fraction of our defense policy will basically develop, at least for some time —which we hope will be for as long as possible—, within institutions that will take their decisions in a generalized cooperative ambient as, e.g., United Nations Organization of the Western European Union.

Among the proposal formulated to the European Council during the meeting held at Maastricht (the Netherlands) in December 1991, there was a project to build a legal framework of actuation for a defense and external relations common policy concerning all the members of the European Economic Community. In particular, the project intended to strip countries of their national sovereignty —entrenched beyond the unanimity rule— by introducing (weighted and) qualified majority decision mechanisms. The ideas discussed in Section 2 are applied to studying these proposals and their effects in Section 3. The corresponding tables and figures have been collected in the Appendix.

## 2. DECISION-MAKING BODIES AND MEASURES OF POWER

Concepts and techniques of cooperative game theory are introduced, which are of interest to represent and analyze mechanisms of collective choice.

### 2.1. Weighted majority games and simple games

Weighted majority games are the structure most frequently encountered in the real world. By using the notion of winning coalition, a more flexible structure, that of simple game, is associated to every weighted majority game which exactly reflects the player's strategic positions. Not all simple games are representable in this way (see, e.g., the U.S. Congress structure).

### 2.2. The Shapley-Shubik index of power

This index is, simply, the restriction to simple games of the Shapley value, which is defined for any cooperative game, it has been axiomatically characterized and is commonly accepted as an evaluation of the game which describes numerically the players' strategic strength. When applied to weighted majority games, a useful theorem on normalized representation is interesting.

### 2.3. The Security Council of United Nations Organization

This example illustrates both the diaphaneity and the sensibility of the Shapley-Shubik index. The original structure of the Council, as well as the modification made in 1966, are studied. The crucial role of the right to veto awarded to the five permanent members is clearly shown.

### 2.4. Coalition structures: Owen's coalitional value

Coalition formation, and hence a valuation of its effects, is the fundamental problem to be studied in cooperative games. Essentially, we would like to determine the power of each actually formed coalition, say, as a player in the new situation, the sharing of this power among the members of such a coalition and, finally, the new power of the remaining players —i.e. those who have not entered any coalition. Owen's coalitional value is an elegant generalization of the Shapley value which provides an answer to these questions. It is an excellent tool that has been successfully applied to the study of several real-world situations.

## 3. THE MAASTRICHT PROPOSALS

We sketch the outlines of the ambitious project that Maastricht's meeting was expected to turn into a great treatise of political and economic union. We emphasize, however, on the disagreements that were known to exist between the twelve about a lot of questions, and also on the highly important *metadecision* that was to be taken about the decision procedures that were suggested to replace the unanimity rule.

### 3.1. The European Parliament

This house is the only communitary organism that does not use unanimity; instead, it follows the straight majority rule. To compare its structure with the projected one for the Council of Ministers we take it as a game among the nations (as players), disregarding the existing transversal partition into parliamentary groups formed by ideological affinities or homologations.

### 3.2. European Comission and the Council of Ministers

These two committees are ruled by unanimity law, exactly as the European Council is.

### 3.3. The qualified majority model

After discussing some aspects of the project, two main models are studied, comparing them with the situation found in the European Parliament and with each other. There is a large family of blocking coalitions, while, on the other hand, an analysis of the range of the quota shows that any majority far enough from unanimity would have given a very stable distribution of power. Hence, once representativity (absence of null players) and fidelity (adequacy of weights to power) are guaranteed, it seems that the main objective in designing the game (using a so very high quota) has been to keep a wide spectrum of blocking possibilities.

### 3.4. Analysis of alliances

When illustrating the use of the coalitional value, several feasible alliances are studied which do not affect the power distribution too much because they

are not formed by winning coalitions. We consider, moreover, polarizations due to some basic questions.

### 3.5. Conclusion

Maastritcht's results are well known. Once again, unanimity rule wins and, at the moment, its opponents, the weighted and qualified majority mechanisms, have been left aside by the European Economic Community.

# SECCIÓ DOCENT I PROBLEMES

La introducció de la nova "SECCIÓ DOCENT I PROBLEMES" a la revista QÜESTIIÓ es fa amb l'objectiu d'incloure una secció on es publiquen articles de caire docent, difícilment publicables en revistes de recerca. Alhora es continua amb l'antiga secció de problemes. A cada número de QÜESTIIÓ s'inclourà d'un a tres problemes i les solucions es donaran en el número següent.

Els lectors poden, si ho volen, proposar problemes amb les solucions pertinents i enviar-los a QÜESTIIÓ, que farà una selecció i en publicarà els més adequats, fent la corresponent referència a l'autor.

També seran ben rebudes solucions alternatives a les propostes fetes per l'autor dels problemes; l'editorial es reservarà, però, el dret a publicar-les.

# PROBLEMES PROPOSATS

## PROBLEMA N° 47

Sea $H(x,y)$ una función de distribución de probabilidad bivariante con marginales univariantes $F(x) = H(x,\infty)$, $G(y) = H(\infty,y)$. Consideremos la curva de regresión de la media de $Y$ sobre $X$:

$$y = m(x) = \mathrm{E}(Y/X = x)$$

que suponemos existe. Se trata de probar las siguientes propiedades:

1) Si $\overline{X} = aX + b$, $\overline{Y} = cY + d$, con $a \neq 0$, entonces la curva

$$\overline{m}(x) = \mathrm{E}(\overline{Y}/\overline{X} = x)$$

es

$$\overline{m}(x) = \mathrm{c.m}\left(\frac{(x-b)}{a}\right) + d.$$

2) Si $m_0(x)$ y $m_1(x)$ son las curvas de regresión de la media para dos distribuciones bivariantes $H_0, H_1$ con las mismas marginales $F, G$, entonces la combinación lineal convexa

$$m_\lambda(x) = \lambda m_1(x) + (1-\lambda)m_0(x), \qquad 0 \leq \lambda \leq 1,$$

es también curva de regresión de la media para una cierta distribución bivariante $H_\lambda$ con las mismas marginales $F, G$.

3) Supongamos ahora que las distribuciones marginales son exponenciales negativas

$$\begin{aligned} F(x) = G(x) &= 1 - e^{-x} \qquad x > 0, \\ &= 0 \qquad\qquad x \leq 0. \end{aligned}$$

Entonces se cumple la siguiente desigualdad

$$\int_0^x m\left[-\log(1-t)\right] dt \geq (1-x)\log(1-x) + x, \qquad 0 < x < 1,$$

cualquiera que sea la curva de regresión de la media $m(x)$ de una distribución bivariante con las mismas marginales $F, G$.

4) Interpretar geométricamente la desigualdad anterior.

C.M. Cuadras

Universitat de Barcelona

**PROBLEMA N° 48**

Sean $X_1, X_2, X_3$ variables aleatorias independientes $\mathrm{N}(0,1)$, y sea $Z$ una variable aleatoria arbitraria. Probar que las variables aleatorias:

$$U_1 = \frac{X_1 + ZX_2}{\sqrt{1+Z^2}}$$

$$U_2 = \frac{X_1 + ZX_2 + Z^2X_3}{\sqrt{1+Z^2+Z^4}}$$

son de nuevo $\mathrm{N}(0,1)$.

José Mª Sarabia

Universidad de Cantabria

# APLICACIÓ DE L'ANÀLISI MULTIVARIANT A UN ESTUDI SOBRE LES LLENGÜES EUROPEES

F. OLIVA, C. BOLANCE i L. DIAZ*

Universitat de Barcelona

*Utilitzant una informació complexa i qualitativa (l'escriptura dels deu primers nombres) es presenta un mètode que permet quantificar adequadament les diferències entre catorze llengües europees construint una matriu de distàncies. Per realitzar l'estudi comparatiu s'empren dues tècniques d'anàlisi multivariant: l'anàlisi de proximitats ("multidimensional scaling") i l'anàlisi de conglomerats jeràrquica ("hierarchical cluster analysis").*

## 1. INTRODUCCIÓ

L'estudi que es presenta a continuació és una comparació de catorze llengües europees a partir d'una informació fàcilment accessible però complexa pel que fa a la seva quantificació: l'escriptura dels deu primers nombres naturals (vegeu la taula 1). Cal dir ja des d'ara mateix que els autors no pretenen extreure conclusions rigoroses des d'un punt de vista lingüístic, sinó mostrar com l'anàlisi multivariant pot proporcionar eines interessants si prèviament s'ha realitzat un esforç per quantificar la informació. La utilització dels deu primers nombres es justifica perquè són paraules força representatives i que no han experimentat canvis de significat en el curs de la història (recordem que solen formar part indefectiblement de la primera lliçó dels llibres de text per aprendre una llengua).

Un antecedent d'aquest estudi pot trobar-se a Johnson i Wichern (1988), on a partir de la informació esmentada anteriorment construeixen una matriu de distàncies entre onze llengües i realitzen posteriorment una anàlisi de conglo-

*Dept. d'Estadística. Facultat de Biologia. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona.

**Taula 1.**
Escriptura dels deu primers nombres de les catorze llengües considerades en l'estudi. S'ha prescindit dels accents i d'altres signes i s'han conservat sols els caràcters alfabètics. Entre parèntesi figuren les abreviatures que seran utilitzades en les representacions gràfiques.

| Alemany | Anglès | Basc | Castellà | Català | Danès | Finlandès | Francès | Gallec | Holandès | Hongarès | Italià | Noruec | Polonès |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (Al) | (An) | (Ba) | (Cs) | (Ca) | (Da) | (Fi) | (Fa) | (Ga) | (Ho) | (Hg) | (It) | (No) | (Po) |
| ein | one | bat | uno | u | en | yksi | un | unho | een | egy | uno | en | jeden |
| zwei | two | bi | dos | dos | to | kaksi | deux | dous | twee | ketto | due | to | dwa |
| drei | three | iru | tres | tres | tre | kolme | trois | tres | drie | harom | tre | tre | trzy |
| vier | four | lau | cuatro | quatre | fire | neua | quatre | catro | vier | negy | quattro | fire | cztery |
| funf | five | bost | cinco | cinc | fem | viisi | cinq | cinco | vijf | ot | cinque | fem | piec |
| sechs | six | tzei | seis | sis | seks | kuusi | six | seis | zes | hat | sei | seks | szesc |
| sieben | seven | tzaspi | siete | set | syv | seitseman | sept | sete | zeven | het | sette | sju | siedem |
| acht | eight | txortzi | ocho | vuit | otte | kahdeksan | huit | oito | acht | nyolc | otto | atte | osiem |
| neun | nine | beratzi | nueve | nou | ni | yhdeksan | neuf | nove | negen | kylenc | nove | ni | dziewiec |
| zehn | ten | amar | diez | deu | ti | kymmenen | dix | dez | tien | tiz | diez | ti | dziesiec |

merats jeràrquica que permet observar algunes agrupacions. El mètode emprat per quantificar les diferències és ben senzill: consideren com a distància entre dues llengües el nombre de paraules que no comencen per la mateixa lletra. Per exemple, la distància entre el polonès i el castellà seria de tres, ja que en tres nombres ("1", "5" i "9" ) la lletra inicial és diferent; en canvi, la distància entre el castellà i l'italià és d'1, ja que sols el "4" no té la primera lletra igual. Aquest algorisme binari per comparar dues paraules desaprofita bona part de la informació disponible i comporta avaluacions poc afortunades: paraules força diferents aporten distància zero, mentre que paraules gairebé idèntiques són valorades com a diferència u. Un exemple ben esclaridor del que pot passar és el nombre "4" en català, castellà i polonès (*quatre, cuatro i cztery* respectivament). Segons la regla anteriorment descrita, per a aquesta paraula hem de considerar el català igual de diferent del castellà i del polonès (!?), mentre que el castellà i el polonès "coincideixen" (per un cop no valdrà l'acudit que els catalans sembla que parlem polonès!). És evident que el mètode resulta poc adequat perquè valora com a iguals paraules que només coincideixen en la primera lletra per casualitat i com a diferents paraules similars quant a l'escriptura i la pronunciació.

Per tal d'evitar aquest problema es proposa un mètode més laboriós però que permeterà considerar tota la informació per avaluar la diferència entre dues llengües. S'ha incorporat a l'estudi per raons d'interès evident el català, el basc i el gallec. Un cop obtinguda la matriu d'interdistàncies, l'anàlisi de proximitats o MDS (*multidimensional scaling*) serà utilitzada per aconseguir una representació adequada en dimensió reduïda que permeti relacionar les diferents llengües. Així mateix, s'empraran tres algorismes d'anàlisi de conglomerats jeràrquica i es construiran els dendrogrames corresponents. Es presentaran i es discutiran diverses mesures de l'ajust en ambdós casos per valorar la fiabilitat de les representacions obtingudes.

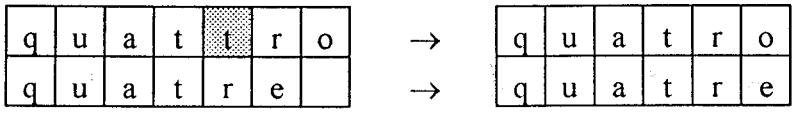
## 2. CONSTRUCCIÓ DE LA MATRIU D'INTERDISTÀNCIES

La construcció d'una dissimilitud entre dues llengües aprofitant al màxim possible la informació proporcionada per la taula 1 no és una tasca fàcil. Pot observar-se que s'ha realitzat una primera simplificació conservant sols les lletres i obviant en l'escriptura de les paraules els accents i la resta de signes. El criteri escollit per calcular la distància està basat en el nombre de no coincidències quan es consideren les paraules senceres, però amb unes normes addicionals que tot seguit exposem.

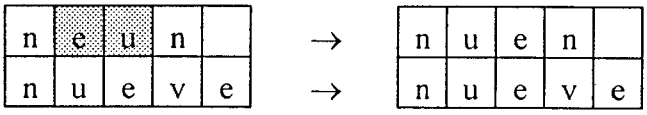
Suposem que volem calcular la dissimilitud entre dues llengües. Per a cada un dels deu nombres realitzarem el següent procés:

1) Utilitzarem si és convenient les regles:

   a) L'addició, deleció o duplicació d'una lletra és considerada com una diferència. Per exemple, considerem el "4" en italià i català. Suprimint una "t" de la paraula italiana (o bé addicionant una "t" al català) aconseguim amb una diferència




   b) La transposició entre dues lletres consecutives és considerada com una diferència. Per exemple, el nombre "9" en alemany i castellà. Transposant les lletres "e" i "u" en qualsevol de les dues paraules obtenim

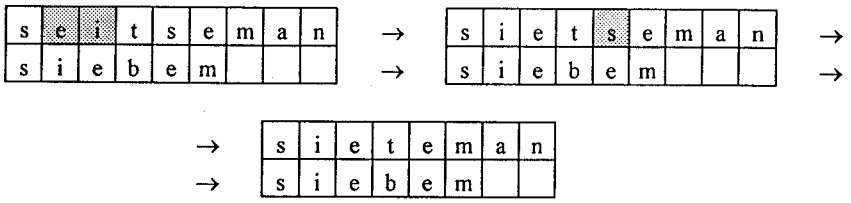



2) Superposarem ambdues paraules i cada lletra no coincident serà considerada una diferència (per exemple, una per al "4" en italià-català i dues per al "9" en alemany-castellà).

   Sumarem les diferències trobades en 1) i 2) (dues i quatre respectivament en els exemples anteriors). En alguns casos, segons l'ordre i el nombre de vegades que són utilitzades les regles, és possible obtenir diferents resultats; aleshores triarem el camí òptim, és a dir, el nombre mínim de diferències.

3) Repetirem el procés per a cada un dels nombres i considerarem com a mesura de dissimilitud entre dues llengües el nombre mitjà de diferències per paraula.

Pot comprovar-se com l'aplicació de les regles a) i b) permeten reduir el nombre de diferències d'una manera lògica. En efecte, sense la seva utilització els dos exemples abans presentats tindrien una diferència més en cada cas. Un bon exemple de l'aplicació completa del procés per a una paraula és el nombre "7" en finlandès (*seitseman*) i polonès (*siebem*):




El nombre de diferències és cinc, mentre que hauria estat vuit si la comparació fos considerant sols les lletres no coincidents.

Utilitzant aquest mètode s'ha construït la matriu de distàncies entre les catorze llengües. El resultat ha estat el següent (vegeu la taula 1 per al significat de les abreviatures)

| | Al | An | Ba | Ca | Cs | Da | Fi | Fr | Ga | Ho | Hg | It | No | Po |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Al | 0.0 | | | | | | | | | | | | | |
| An | 2.9 | 0.0 | | | | | | | | | | | | |
| Ba | 4.5 | 4.4 | 0.0 | | | | | | | | | | | |
| Ca | 3.4 | 2.8 | 4.5 | 0.0 | | | | | | | | | | |
| Cs | 3.2 | 2.9 | 4.6 | 1.7 | 0.0 | | | | | | | | | |
| Da | 3.0 | 2.6 | 4.3 | 2.7 | 3.1 | 0.0 | | | | | | | | |
| Fi | 5.8 | 5.5 | 5.9 | 5.7 | 5.5 | 5.9 | 0.0 | | | | | | | |
| Fr | 3.3 | 3.2 | 4.6 | 1.3 | 2.4 | 3.3 | 5.9 | 0.0 | | | | | | |
| Ga | 3.2 | 2.7 | 4.4 | 1.3 | 0.7 | 2.6 | 5.5 | 2.3 | 0.0 | | | | | |
| Ho | 1.9 | 2.5 | 4.3 | 4.3 | 3.2 | 2.9 | 5.6 | 3.3 | 3.3 | 0.0 | | | | |
| Hg | 4.2 | 3.8 | 4.5 | 4.0 | 4.2 | 3.6 | 5.6 | 3.8 | 4.0 | 3.7 | 0.0 | | | |
| It | 3.7 | 3.5 | 4.6 | 2.2 | 1.7 | 3.2 | 6.0 | 2.4 | 1.5 | 3.6 | 4.5 | 0.0 | | |
| No | 2.9 | 2.7 | 4.3 | 2.9 | 3.2 | 0.3 | 5.8 | 3.3 | 2.7 | 2.8 | 3.6 | 3.3 | 0.0 | |
| Po | 4.5 | 4.4 | 5.3 | 4.4 | 3.6 | 4.4 | 5.6 | 4.5 | 3.8 | 4.2 | 5.2 | 4.2 | 4.4 | 0.0 |

## 3. L'ANÀLISI DE PROXIMITATS O MDS

Considerem un conjunt finit de $n$ objectes (individus, poblacions, ...) $O_1, \ldots, O_n$.. La MDS (*multidimensional scaling*) és una tècnica multivariant d'anàlisi de dades que permet trobar una configuració de $n$ punts en un espai euclidià utilitzant com a informació les proximitats (similituds o dissimilituds) entre els $n$ objectes.

Direm que $\boldsymbol{D}_{n\times n} = (d_{ij})$ on $d_{ij} = d(O_i, O_j)$ és una matriu de distàncies si compleix les següents propietats

(a) simetria: $d_{ij} = d_{ji}$ (b) no negativitat: $d_{ij} \geq 0$, $i \neq j$ (c) $d_{ii} = 0$

Si a més compleix les propietats

(e) $d_{ij} = 0 \Leftrightarrow O_i \equiv O_j$ (f) desigualtat triangular: $d_{ij} \leq d_{ik} + d_{jk}$

aleshores $\boldsymbol{D}$ és una matriu de distàncies mètrica. Si una distància sols presenta les tres primeres propietats molts autors l'anomenen dissimilitud. Finalment una matriu de distàncies $\boldsymbol{D}$ és euclidiana si compleix les cinc propietats i a més és possible trobar una configuració de punts $P_1, \ldots, P_n$ en un espai euclidià amb coordenades $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ tals que

$$d_{ij}^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)'(\boldsymbol{x}_i - \boldsymbol{x}_j)$$

En algunes situacions, es disposa de les similituds entre els objectes i no pas de distàncies. Diem que $\boldsymbol{S}_{n\times n} = (s_{ij})$ on $s_{ij} = s(O_i, O_j)$ és una matriu de similituds si

(a) $s_{ij} = s_{ji}$ (b) $s_{ij} \leq s_{ii}$

És fàcil però comprovar que podem obtenir una matriu de dissimilituds a partir de les similituds mitjançant la transformació

$$d_{ij} = (s_{ii} + s_{jj} - 2s_{ij})^{1/2}$$

A partir de les dissimilituds o distàncies $d_{ij}$, l'objectiu de la MDS és trobar una configuració de $n$ punts $P_1, \ldots, P_n$ en dimensió $k$ tal que si denominem $d_{ij(k)}$ la distància euclidiana entre $P_i$ i $P_j$, aleshores $\boldsymbol{D}_{(k)}$ sigui "semblant" a $\boldsymbol{D}$. La dimensió $k$ és desconeguda, però a la pràctica es limita generalment a $k \leq 3$ per possibilitar la interpretació. És important assenyalar que la solució obtinguda és invariant quant a translacions, rotacions i reflexions. Hi ha nombrosos mètodes de MDS, però poden distingir-se dos grans grups:

- *mètodes mètrics:* intenten aconseguir la configuració de punts $P_i$ utilitzant directament les distàncies entre els objectes.

- *mètodes no mètrics:* la informació utilitzada és sols el rang de les $n(n-1)/2$ distàncies entre tots els parells d'objectes

$$d_{i_1j_1} < d_{i_2j_2} < \cdots < d_{i_mj_m} \qquad m = n(n-1)/2$$

Atès que els rangs no varien per transformacions monòtones de les distàncies, la solució obtinguda serà també invariant respecte l'expansió o contracció uniforme.

En aquest estudi s'ha utilitzat el mètode mètric clàssic que serà breument descrit a continuació.

### Solució mètrica clàssica

Sigui $\boldsymbol{D}$ la matriu d'interdistàncies entre els $n$ objectes. Considerem les matrius $\boldsymbol{A}$ i $\boldsymbol{B}$ d'ordre $n$

$$\boldsymbol{A} = (a_{ij}), \quad a_{ij} = -\frac{1}{2}d_{ij}^2 \qquad\qquad \boldsymbol{B} = (b_{ij}), \quad b_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..}$$

on

$$a_{i.} = \frac{1}{n}\sum_{j=1}^{n} a_{ij} \qquad a_{.j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij} \qquad a_{..} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}$$

O expressat en forma matricial $\boldsymbol{B} = \boldsymbol{HAH}$ on $\boldsymbol{H} = \boldsymbol{I} - n^{-1}\mathbf{1}\mathbf{1}'$ és l'anomenada matriu centradora de dades d'ordre $n$. Es compleixen aleshores els següents resultats:

a) Si $\boldsymbol{D}$ és la matriu de distàncies euclidianas per a una configuració de punts $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ llavors $b_{ij} = (\boldsymbol{z}_i - \overline{\boldsymbol{z}})(\boldsymbol{z}_j - \overline{\boldsymbol{z}})'$, $i, j = 1, \ldots, n$. En forma matricial $\boldsymbol{B} = (\boldsymbol{HZ})(\boldsymbol{HZ})'$ i per tant $\boldsymbol{B} \geq 0$, és a dir, és semidefinida positiva (s.d.p.).

b) I a l'inrevés, si $\boldsymbol{B}$ és s.d.p. de rang $r \leq n-1$ pot aleshores construir-se una configuració de $n$ punts $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, on $\boldsymbol{x}_i \in \mathbb{R}^r$ és la fila $i$–èsima de $\boldsymbol{X}$, tals que $d_{ij}^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)'(\boldsymbol{x}_i - \boldsymbol{x}_j)$. L'obtenció de $\boldsymbol{X}$ es immediata diagonalitzant $\boldsymbol{B}$. En efecte,

$$\boldsymbol{B} = \boldsymbol{T\Lambda T}' = (\boldsymbol{T\Lambda}^{1/2})(\boldsymbol{T\Lambda}^{1/2})' = \boldsymbol{XX}'$$

on $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_r)$ és la matriu diagonal de valors propis diferents de zero $\lambda_1 \geq \cdots \geq \lambda_r > 0$ i $\boldsymbol{T}$ és la matriu de vectors propis associats.

Algunes propietats importants són:

a) Les columnes de $\boldsymbol{X}$ són els vectors propis $\lambda$–normalitzats, és a dir, $\boldsymbol{x}'_{(i)}\boldsymbol{x}'_{(i)} = \lambda_i$, $\boldsymbol{x}'_{(i)}\boldsymbol{x}_{(j)} = 0$, $i \neq j$ $(i, j = 1, \ldots, r)$.

b) El centre de gravetat és l'origen de coordenades. Efectivament, $\boldsymbol{B}\mathbf{1} = \boldsymbol{HAH}\mathbf{1} = \mathbf{0}$ i per tant $\mathbf{1}$ és vector propi de valor propi 0. Llavors

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = \boldsymbol{X}'\mathbf{1} = \mathbf{0}$$

c) Es compleix la relació $\sum_{i,j}^{n} d_{ij}^2 = \text{tr } \boldsymbol{D}^2 = 2n \text{ tr } \boldsymbol{B} = 2n\sum_{i=1}^{r}\lambda_i$ on $\boldsymbol{D}^2 = \boldsymbol{DD}$.

d) Si volem representar els objectes en un espai de dimensió reduïda $k$, la màxima resolució (separació entre els objectes) serà aconseguida utilitzant les $k$ primeres coordenades (columnes de $\boldsymbol{X}$): $\boldsymbol{x}_{i(k)} = (x_{i1} \cdots \boldsymbol{x}_{ik})'$, $i = 1, \ldots, n$. La dispersió dels objectes en aquest espai euclidià serà

$$\sum_{i,j}^{n} d_{ij(k)}^2 = 2n \text{ tr } \boldsymbol{B}_{(k)} = 2n\sum_{i=1}^{k}\lambda_i \quad \text{on } \boldsymbol{B}_{(k)} = \boldsymbol{X}_{(k)}\boldsymbol{X}'_{(k)}$$

És a dir, de totes les possibles configuracions $\overline{\boldsymbol{x}}_{i(k)}$ dels $n$ objectes en un espai euclidià de dimensió $k$, pot demostrar-se que la mesura de discrepància $\phi = \sum_{i,j}(d_{ij}^2 - \tilde{d}_{ij(k)}^2)$ és mínima si $\tilde{\boldsymbol{x}}_{i(k)} = \boldsymbol{x}_{i(k)}$. Aleshores $\min\phi = \text{ tr } (\boldsymbol{B} - \boldsymbol{B}_{(k)})$ i la mesura

$$c = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{r}\lambda_i} \times 100 = \frac{\text{tr } \boldsymbol{B}_{(k)}}{\text{tr } \boldsymbol{B}} \times 100 = 1 - \left[\frac{\text{tr } (\boldsymbol{B} - \boldsymbol{B}_{(k)})}{\text{tr } \boldsymbol{B}}\right] \times 100.$$

que ens indica el percentatge de dispersió (variabilitat), explicada per les $k$ primeres coordenades, és màxima.

e) Si la informació inicial és una matriu de similituds $\boldsymbol{S}$ entre els $n$ objectes, no és necessari transformar-les a dissimilituds, atès que és fàcil comprovar que $\boldsymbol{B} = \boldsymbol{HSH}$.

Aquesta solució, generalment coneguda com a mètode clàssic de la MDS, fou demostrada per Schoenberg (1935) i Richardson (1938) però ha estat popularitzada per Torgerson (1952, 1958), que introduí el terme *multidimensional scaling*. Més tard, Gower (1966) l'anomenà anàlisi de coordenades principals i va mostrar l'estreta connexió amb l'anàlisi de components principals.

Què passa però si la distància no és euclidiana? En aquest cas $\boldsymbol{B}$ tindrà valors propis negatius i no podem definir en els reals la potència $\boldsymbol{\Lambda}^{1/2}$, per tant no és possible $\lambda-$normalitzar els vectors propis. Suposem que $r(\boldsymbol{B}) = r$, $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_r)$ amb $p$ valors propis positius i $q = r - p$ valors propis negatius

$$\lambda_1 \geq \cdots \geq \lambda_p > 0 > \lambda_{p+1} \geq \cdots \geq \lambda_r$$

Per aconseguir $\boldsymbol{x}'_{(i)}\boldsymbol{x}_{(i)} = \lambda_i\,(i = 1, \ldots, r)$ les últimes $q$ coordenades han de ser imaginàries. En efecte,

$$\boldsymbol{x}_{(h)} = i\,|\lambda_h|^{1/2}\,\boldsymbol{t}_{(h)}, \quad h = p+1, \ldots, r \quad \text{on} \quad i = \sqrt{-1}$$

i les coordenades dels objectes serien $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jp}, ix_{jp+1}, \ldots, ix_{jr})'$, $j = 1, \ldots, n$. Les interdistàncies entre els $n$ objectes poden aleshores ser expressades com

$$d_{ij}^2 = \sum_{h=1}^{p}(x_{ih} - x_{jh})^2 - \sum_{h=p+1}^{r}(x_{ih} - x_{jh})^2$$

distàncies corresponents a una geometria ortogonal no representable en l'espai euclidià real (es pot observar que les dispersions dels objectes per a les $q$ últimes coordenades representen de fet anti-distàncies).

La solució més senzilla al problema plantejat és considerar sols les $p$ coordenades reals i obviar les $q$ imaginàries, la qual cosa implica aproximar la matriu $\boldsymbol{B}$ per una altra semidefinida positiva de rang inferior $\boldsymbol{B}^* = \boldsymbol{B}_{(p)} = \boldsymbol{X}_{(p)}\boldsymbol{X}'_{(p)}$. Aquest problema fou estudiat per Eckart i Young (1936) en un context general i per Mardia (1978) en el context de la MDS, obtenint el següent resultat:

Si considerem

$$\psi = \sum_{i,j=1}^{n}(b_{ij} - \tilde{b}_{ij})^2 = \text{tr}\,(\boldsymbol{B} - \tilde{\boldsymbol{B}})^2$$

la mesura de la discrepància entre $\boldsymbol{B}$ i una altra matriu simètrica s.d.p. de rang inferior $\tilde{\boldsymbol{B}}$ es demostra aleshores que $\psi$ es minimitza si $\tilde{\boldsymbol{B}} = \boldsymbol{B}^*$ i, per tant,

$$\min\psi = \text{tr}\,(\boldsymbol{B} - \boldsymbol{B}^*)^2 = \sum_{i=p+1}^{r}\lambda_i^2$$

Aquest resultat comporta la següent generalització: si volem aconseguir una configuració dels objectes en un espai euclidià de dimensió $k \leq p$ l'expressió $\psi$ serà minimitzada quan $\tilde{\boldsymbol{B}} = \boldsymbol{B}_{(k)}$. Mardia (1978) proposa dues mesures de la dispersió explicada per la representació dels objectes en $\mathbb{R}^k$

$$c_1 = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{r}|\lambda_i|}\times 100 \qquad c_2 = \frac{\sum_{i=1}^{k}\lambda_i^2}{\sum_{i=1}^{r}|\lambda_i^2|}\times 100 = \left[1 - \frac{\text{tr}\,(\boldsymbol{B} - \boldsymbol{B}_{(k)})^2}{\text{tr}\,\boldsymbol{B}^2}\right]\times 100$$

La justificació de $c_2$ (mesura proposada també per Saito (1978)) és evident donada la seva relació amb $\psi$, mentre que $c_1$ és una generalització de la mesura $c$ presentada anteriorment en el cas que $\boldsymbol{D}$ sigui euclidiana. No obstant això, en l'opinió dels autors d'aquest estudi, $c_1$ és menys "natural" atès que el denominador no correspon a cap dispersió global. En efecte

$$c_1 = \frac{\text{tr } \boldsymbol{B}_{(k)}}{\text{tr } \boldsymbol{B} + 2 \text{ tr } (\boldsymbol{B}^* - \boldsymbol{B})} \times 100 = \left[1 - \frac{\text{tr } (\boldsymbol{B} - \boldsymbol{B}_{(k)}) + 2 \text{ tr } (\boldsymbol{B}^* - \boldsymbol{B})}{\text{tr } \boldsymbol{B} + 2 \text{ tr } (\boldsymbol{B}^* - \boldsymbol{B})}\right] \times 100$$

Cal finalment assenyalar que el procediment exposat serà poc aconsellable si el "pes" dels valors propis negatius és alt o fins i tot impossible d'aplicar si $k > p$. En aquest cas poden emprar-se altres mètodes més adequats, com per exemple la solució de Mardia (1978), el mètode iteratiu lineal i els mètodes no mètrics (el lector interessat pot consultar els capítols sobre el tema de Cuadras (1991), Dillon i Goldstein (1984), Mardia *et al* (1979), Seber (1984) o l'obra específica sobre el tema de Davison (1983)).

## 4. L'ANÀLISI DE CONGLOMERATS JERÀRQUICA

Donat un conjunt de $n$ objectes $=_1, \ldots, O_n$ dels quals es disposa una informació quantificable, l'anàlisi de conglomerats (*cluster analysis*) engloba una sèrie de tècniques que tenen com a finalitat l'agrupació dels objectes més semblants (mètodes aglomeratius) o bé la formació de subconjunts a partir del conjunt inicial (mètodes divisius). L'objectiu és classificar els $n$ objectes de manera que, respecte a la informació coneguda, siguin el més homogenis possible dins dels grups i heterogenis entre els grups. Els mètodes d'anàlisi de conglomerats poden classificar-se també d'acord amb un altre criteri:

- *mètodes jeràrquics*: estableixen successives fusions o divisions dels objectes depenent de la seva homogeneïtat, formant una estructura de grups jerarquitzada. Presenten la particularitat que quan un objecte ha estat assignat a un grup no és ja possible reconsiderar la seva classificació. El resultat obtingut pot ser representat mitjançant un diagrama de dues dimensions en forma d'arbre anomenat dendrograma.

- *mètodes no jeràrquics*: pretenen aconseguir grups homogenis sense establir entre ells relacions de jerarquia. Són tècniques que realitzen una partició del conjunt dels n objectes optimitzant un determinat criteri formal predeterminat. En oposició als mètodes jeràrquics, un individu inicialment assignat a un grup pot ser posteriorment reclassificat.

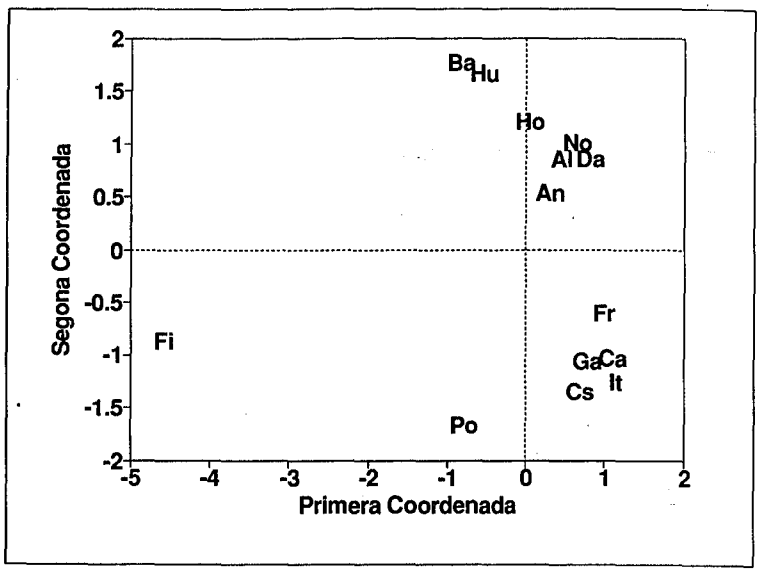



**Figura 1.** Representació de les llengües emprant les dues primeres coordenades obtingudes amb la MDS.

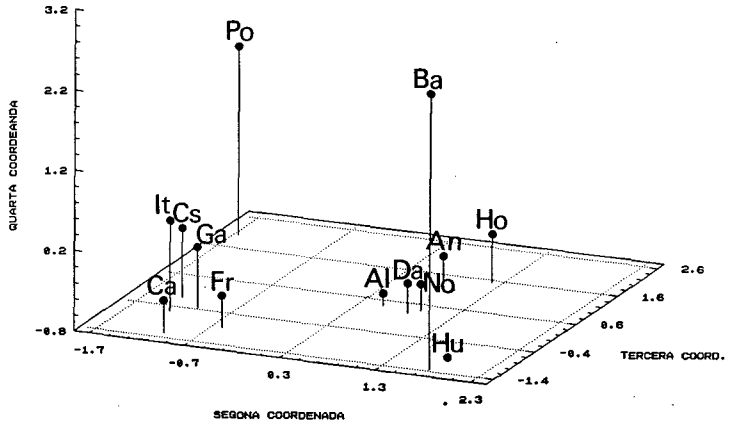



**Figura 2.** Representació de les llengües, exceptuant el finlandès, utilitzant la segona, tercera i quarta coordenades de la MDS.

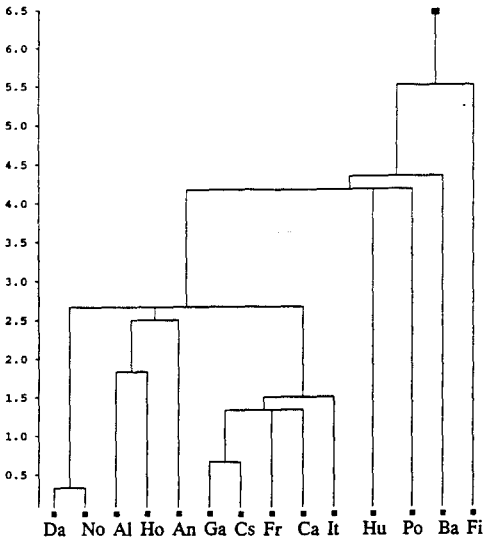



**Figura 3a)**

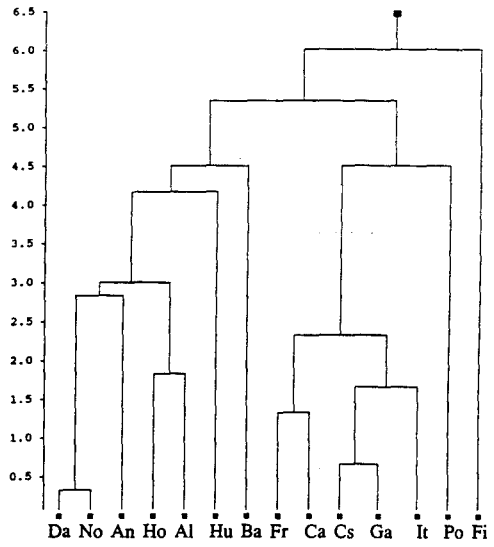



**Figura 3b)**

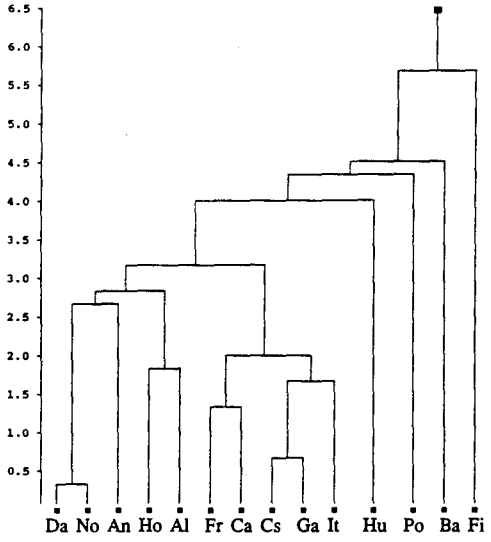



**Figura 3c)**

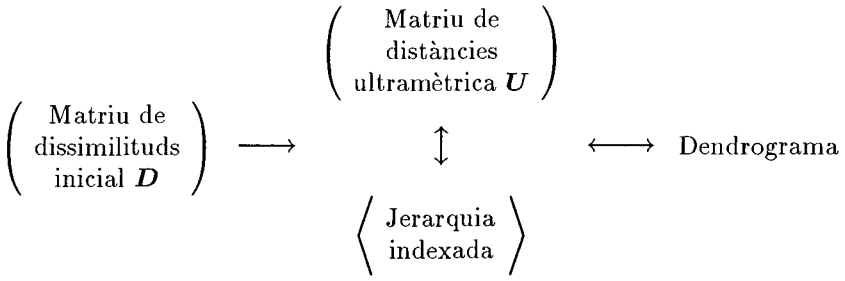
**Figura 3.** Dendrogrames obtinguts amb l'anàlisi de conglomerats jeràrquica: a) mètode del mínim, b) mètode del màxim, i c) mètode UPGMA.

En aquest estudi lingüístic s'han utilitzat mètodes jeràrquics aglomeratius, atès que la finalitat és aconseguir una agrupació successiva de les llengües europees per observar les "relacions de parentiu". Presentem a continuació una breu exposició dels mètodes utilitzats.

Sigui $\boldsymbol{D}$ una matriu de dissimilituds entre $n$ objectes. Els mètodes d'anàlisi de conglomerats jeràrquica es basen en algorismes de construcció d'una matriu de distàncies ultramètrica $\boldsymbol{U} = (u_{ij})$, $u_{ij} = u(O_i, O_j)$, que sigui el més semblant possible a $\boldsymbol{D}$. Perquè una distància sigui ultramètrica ha de verificar la següent propietat:

$$u_{ij} \leq \max\{u_{ik}, u_{jk}\}$$

coneguda com a axioma ultramètric i que és més restrictiva que la desigualtat triangular. Com a conseqüència geomètrica d'aquesta propietat, tot triangle definit per les distàncies ultramètriques entre tres objectes és isòsceles, la qual cosa pot ser comprovada amb els dendrogrames representats en la figura 3. No és l'objectiu d'aquesta exposició aprofundir en aspectes formals i per tant només citarem que tota distància ultramètrica $u$ definida sobre un conjunt finit de $n$ objectes defineix una jerarquia indexada en $\{O_1, \ldots, O_n\}$; així mateix, tota jerarquia indexada implica una distància ultramètrica definida entre els $n$ objectes. Finalment, un dendrograma no és més que la representació geomètrica d'una jerarquia indexada i, en conseqüència, d'una distància ultramètrica. Aquest procés pot ser esquematitzat de la següent manera




Hi ha nombrosos algorismes per construir una distància ultramètrica adequada (i, per tant, una jerarquia indexada) a partir d'una matriu de dissimilituds $\boldsymbol{D}$. S'han emprat en aquest estudi tres dels més freqüentment utilitzats:

a) **mètode del mínim**, *single linkage* o *nearest-neighbour* (Sneath (1957), Sokal i Sneath (1963), Johnson (1967)): si $C_1$ i $C_2$ són dos conglomerats o grups, la distància entre ells es defineix com la mínima distància observada entre un membre de $C_1$ i un altre de $C_2$, és a dir,

$$d_{(C_1)(C_2)} = \min\{d_{ij} : i \in C_1,\ j \in C_2\}$$

Mostrarem el procés de fusió amb un exemple senzill. Sigui

$$\boldsymbol{D} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{pmatrix} 0 & 7 & 1 & 9 \\ 7 & 0 & 6 & 3 \\ 1 & 6 & 0 & 8 \\ 9 & 3 & 8 & 0 \end{pmatrix} \end{array}$$

la matriu d'interdistàncies entre quatre objectes $O_1, O_2, O_3$ i $O_4$. La fusió ha de començar entre $O_1$ i $O_3$, ja que són els objectes més propers (homogenis). Els grups són ara $(O_1, O_3), O_2$ i $O_4$. Aleshores

$$d_{(2)(1,3)} = \min\{d_{21}, d_{23}\} = d_{23} = 6, \quad d_{(4)(1,3)} = \min\{d_{41}, d_{43}\} = d_{43} = 8$$

i la matriu resultant és

$$\boldsymbol{D}_1 = \begin{array}{c} \begin{matrix} (1,3) & 2 & 4 \end{matrix} \\ \begin{pmatrix} 0 & 6 & 8 \\ 6 & 0 & 3 \\ 8 & 3 & 0 \end{pmatrix} \end{array}$$

La unió serà ara entre $O_2$ i $O_4$ i es formaran els grups $(O_1, O_3)$ i $(O_2, O_4)$, amb distància $d_{(1,3)(2,4)} = \min\{d_{2(1,3)}, d_{4(1,3)}\} = d_{2(1,3)} = 6$. Finalment unirem els dos conglomerats en un sol conjunt global $(O_1, O_2, O_3, O_4)$. El resultat ha estat una jerarquia i una distància ultramètrica definida entre el objectes

$$\boldsymbol{D} = \begin{pmatrix} 0 & 7 & 1 & 9 \\ 7 & 0 & 6 & 3 \\ 1 & 6 & 0 & 8 \\ 9 & 3 & 8 & 0 \end{pmatrix} \longrightarrow \boldsymbol{U} = \begin{pmatrix} 0 & 6 & 1 & 6 \\ 6 & 0 & 6 & 3 \\ 1 & 6 & 0 & 6 \\ 6 & 3 & 6 & 0 \end{pmatrix}$$

que permeterà la representació en un dendrograma com el de la figura 2. El mètode del mínim és un algorisme espai contractiu: la ultramètrica associada a la classificació jeràrquica tendeix a aproximar els objectes respecte les dissimilituds inicials.

b) **mètode del màxim**, *complete linkage* o *farthest-neighbour* (Sokal i Sneath (1963), McQuitty (1964)): oposat al mètode anterior, un cop units els dos grups més propers $C_1$ i $C_2$, la distància es defineix per

$$d_{(C_1)(C_2)} = \max\{d_{ij} : i \in C_1,\ j \in C_2\}$$

És un algorisme espai dilatant: tendeix a allunyar els objectes respecte la dissimilitud inicial.

c) **mètode UPGMA** (*Unweighted Pair Group Method Using Arithmetic Averages*) o *group average* (Sokal i Michener (1958), McQuitty (1964), Lance i Williams (1966)): la distància entre $C_1$ i $C_2$ és definida en aquest cas com la mitjana aritmètica de les $n_1 n_2$ distàncies entre totes les parelles d'objectes formades per un element de $C_1$ i un altre de $C_2$

$$d_{(C_1)(C_2)} = \frac{1}{n_1 n_2} \sum_{i \in C_1} \sum_{j \in C_2} d_{ij}$$

És un algorisme espai conservador i no modifica substancialment les dissimilituds inicials.

Algunes altres propietats importants d'aquests mètodes són:

- invariància monòtona: un algorisme verifica aquesta propietat si no varia l'estructura jeràrquica quan $\boldsymbol{D}$ és substituïda per una nova matriu de dissimilituds $\hat{\boldsymbol{D}}$ obtinguda mitjançant una transformació monòtona de $\boldsymbol{D}$. Els tres mètodes esmentats presenten invariància monòtona.

- no arbitrarietat: si dues o més dissimilituds són iguals, la classificació jeràrquica no depèn de l'ordre en què han estat agrupats els objectes. Dels tres mètodes, sols el del mínim garanteix aquesta propietat.

- continuïtat: petites pertorbacions en les dissimilituds inicials haurien de provocar sols petites modificacions en el dendrograma resultant. Novament sols el mètode del mínim assegura aquesta propietat.

- no encadenament: l'encadenament es produeix quan diversos conglomerats s'ajunten ràpidament a causa de l'existència de pocs individus intermedis. Aquest és un problema que presenta amb freqüència el mètode del mínim, perquè és espai contractiu.

Finalment, cal considerar alguna mesura de la qualitat de la classificació, ja que el procés $\boldsymbol{D} \longrightarrow \boldsymbol{U}$ provoca una distorsió (excepte si $\boldsymbol{D}$ ja és ultramètrica). Un procediment força utilitzat consisteix a calcular la correlació entre les $n(n-1)/2$ parelles $d_{ij}, u_{ij}$) i rep el nom de "correlació cofenètica" $0 \leq r_c \leq 1$). Introduïda per Sokal i Rohlf (1962) i analitzades amb profunditat les seves propietats per Farris (1969 ), proporciona una mesura de la distorsió: valors baixos adverteixen d'una forta discrepància entre les dissimilituds inicials i les distàncies ultramètriques, mentre que valors propers a 1 indiquen una evident estructura jeràrquica entre els $n$ objectes. Una altra mesura, proposada per Jardine i Sibson (1968), és el coeficient

$$\lambda_\alpha = \frac{\left[\sum_{i,j} |d_{ij} - u_{ij}|^{1/\alpha}\right]^\alpha}{\left[\sum_{i,j} d_{ij}^{1/\alpha}\right]^\alpha} \quad 0 < \alpha < 1 \quad 0 \leq \lambda_\alpha \leq 1$$

que depèn del paràmetre $\alpha$. Una bona classificació quedarà reflectida per un valor de $\lambda_\alpha$ proper a zero. Si escollim $\alpha = 1/2$ coincideix amb la popular mesura de *STRESS* utilitzada en diversos mètodes no mètrics de MDS.

## 5. MESURES DE L'AJUST

En l'exposició de la tècnica clàssica de la MDS s'han comentat dues mesures de la dispersió, $c_1$ i $c_2$, explicada en representar els $n$ objectes en $\mathbb{R}^k$; en l'apartat anterior s'ha parlat de la correlació cofenètica $r_c$ i del coeficient $\lambda_\alpha$ (infinites mesures depenent del valor de $\alpha$) com a mesura de la distorsió d'un dendrograma. Es plantegen aleshores diverses qüestions: quan aquestes mesures indiquen un bon ajust? Hi ha altres mesures de l'ajust adequades o que s'hagin de considerar? Per què no utilitzar també altres mesures de la distorsió emprades en altres mètodes? Preguntes en efecte interessants, però que no tenen una fàcil resposta.

No ha estat pretensió dels autors solucionar aquestes qüestions, sinó evidenciar la prudència amb la qual han de ser considerades aquestes mesures. Amb aquesta finalitat s'han emprat, a part de les mesures ja esmentades, els següents coeficients:

- El *STRESS*, proposat per Kruskal (1964) i utilitzat com a coeficient a minimitzar per diversos programes de MDS no mètrica, com per exemple el KYST (Kruskal *et al.*, 1973)

$$S = \left[ \frac{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2} \right]^{1/2} = \left[ \frac{\sum_{i,j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2} \right]^{1/2}$$

  on $\hat{d}_{ij}$ seran en el nostre cas les interdistàncies en dimensió $k$ per a la MDS i les ultramètriques en les anàlisis de conglomerats. La utilització com a mesura d'ajust està ampliament justificada per la següent raó: si definim l'error quadràtic de la representació per una parella d'objectes $(O_i, O_j)$ com $e_{ij}^2 = (d_{ij} - \hat{d}_{ij})^2$, el numerador de $S^2$ és aleshores l'error quadràtic total. El terme del denominador "normalitza" $S$ i permet que sigui una mesura estandarditzada. Poden establir-se dues altres relacions interessants:

  a) $S$ pot ser expressat com el quocient de les normes de dos vectors

$$S = \frac{\|\boldsymbol{d} - \hat{\boldsymbol{d}}\|}{\|\boldsymbol{d}\|}$$

on $\boldsymbol{d} = (d_{12}, \ldots, d_{1n}, d_{23}, \ldots, d_{2n}, d_{34}, \ldots, d_{nn})'$ i $\hat{\boldsymbol{d}}$ es defineix anàlogament.

b) $$S^2 = \frac{\text{tr } (\boldsymbol{D} - \hat{\boldsymbol{D}})^2}{\text{tr } \boldsymbol{D}^2} = \frac{\text{tr } (\boldsymbol{B} - \hat{\boldsymbol{B}})}{\text{tr } \boldsymbol{B}} + \frac{2 \text{ tr } (\hat{\boldsymbol{D}}^2 - \boldsymbol{D}\hat{\boldsymbol{D}})}{\text{tr } \boldsymbol{D}^2}$$

- El *S−STRESS*, plantejat per Takane *et al.* (1977) i Young i Lewyckyj (1979), utilitzat en l'algorisme ALSCAL de MDS no mètric

$$SS = \left[ \frac{\sum_{i<j} (d_{ij}^2 - \hat{d}_{ij}^2)^2}{\sum_{i<j} (d_{ij}^2)^2} \right]^{1/2}$$

S'observa que la diferència consisteix en el fet que les distàncies són elevades al quadrat. Poden realitzar-se consideracions similars a la mesura anterior

a) $$SS = \frac{\|\boldsymbol{d}^2 - \hat{\boldsymbol{d}}^2\|}{\|\boldsymbol{d}^2\|} \qquad \text{b)} \quad SS^2 = \frac{\text{tr } (\boldsymbol{A} - \hat{\boldsymbol{A}})^2}{\text{tr } \boldsymbol{A}^2}$$

on $\boldsymbol{d}^2 = (d_{i2}^2, \ldots, d_{1n}^2, d_{23}^2, \ldots, d_{2n}^2, d_{34}^2, \ldots, d_{nn}^2)'$ i $\hat{\boldsymbol{d}}^2$ es defineix anàlogament.

- coeficient d'alienació $K$, proposat per Guttman (1968)

$$K = \sqrt{1 - \mu^2}$$

on $\mu$ és l'anomenat coeficient de monotonicitat

$$\mu = \frac{\sum_{i<j} d_{ij} \hat{d}_{ij}}{\left[ \sum_{i<j} d_{ij}^2 \sum_{i<j} \hat{d}_{ij}^2 \right]^{1/2}}$$

Pot observar-se que $\mu = \cos \phi$ on $\phi$ és l'angle format pels vectors $\boldsymbol{d}$ i $\hat{\boldsymbol{d}}$ i per tant $K = \sin \phi$. L'única diferència entre $\mu$ i $r_c$ és que en aquest darrer coeficient les distàncies són centrades. S'acompleix també la relació:

$$K^2 = \frac{\text{tr } (\boldsymbol{B} - \hat{\boldsymbol{B}})}{\text{tr } \boldsymbol{B}} - \frac{\text{tr } (\boldsymbol{D}\hat{\boldsymbol{D}} + \hat{\boldsymbol{D}}^2) \text{ tr } (\boldsymbol{D}\hat{\boldsymbol{D}} - \hat{\boldsymbol{D}}^2)}{\text{tr } \boldsymbol{D}^2 \text{ tr } \hat{\boldsymbol{D}}^2}$$

Hem trobat per tant les relacions entre tres mesures considerades característiques de mètodes no mètrics de MDS i que poden ser també adequades per a mètodes mètrics i per a l'anàlisi de conglomerats. Així mateix, la correlació cofenètica pot ser una mesura més a considerar en el cas de la MDS.

## 6. PRESENTACIÓ I ANÀLISI DELS RESULTATS

En la taula 2 poden trobar-se els valors propis obtinguts en diagonalitzar $\boldsymbol{B} = \boldsymbol{HAH}$ i els coeficients $c_1$ i $c_2$. Pot observar-se que el rang de $\boldsymbol{B}$ és 13 i que $\boldsymbol{D}$ no és euclidiana, ja que han aparegut valors propis negatius. No obstant això, si considerem sols les 11 coordenades reals, la distorsió provocada és pràcticament negligible. Efectivament, $c_1 = 98,1\%$ i $c_2 = 99.8\%$, per tant podem obviar els valors propis negatius i aproximar $\boldsymbol{B}$ per $\boldsymbol{B}^*$.

**Taula 2.**
Valors propis obtinguts en diagonalitzar $\boldsymbol{B} = \boldsymbol{HAH}$ i coeficients $c_1$ i $c_2$ indicadors de la dispersió explicada (vegeu-ne la definició en el text).

| $\lambda$ | $c_1$ | $c_2$ | $\lambda$ | $c_1$ | $c_2$ |
|---|---|---|---|---|---|
| 28,24 | 27,3 | 47,0 | 3,83 | 95,0 | 99,4 |
| 19,44 | 46,0 | 69,3 | 2,38 | 97,3 | 99,7 |
| 13,80 | 59,4 | 80,5 | 0,76 | 98,0 | 99,8 |
| 12,35 | 71,3 | 89,5 | 0,06 | 98,1 | 99,8 |
| 8,80 | 79,8 | 94,0 | 0 | 98,1 | 99,8 |
| 7,64 | 87,2 | 97,5 | −0,06 | 98,1 | 99,8 |
| 4,24 | 91,3 | 98,6 | −1,94 | 100,0 | 100,0 |

La figura 1 mostra la representació MDS en $\mathbb{R}^2$ obtinguda al considerar les dues primeres coordenades. La primera coordenada presenta un "pes" considerable, però és fonamentalment degut a la clara separació entre el finlandès i la resta de llengües. Examinant les mesures d'ajust (vegeu les taules 2, 3 i 4), podem veure que considerar sols les dues primeres coordenades comporta una elevada pèrdua d'informació. Atès que la primera coordenada correspon pràcticament a la patent separació del finlandès, hem representat la resta de llengües en un diagrama tridimensional on els eixos són la segona, tercera i quarta coordenades. El significat i aportació de cada una de les coordenades pot sintetitzar-se de la manera següent:

- Primera coordenada: separació indubtable del finlandès com a llengua no comparable a cap de les altres.
- Segona coordenada: poden observar-se dos grups força evidents. Un grup està constituït pel català, castellà, gallec, francès, italià i polonès; l'altre és format per la resta de les llengües exceptuant el finlandès.

- Tercera coordenada: mostra una clara separació del polonès enfront del català, castellà, francès, gallec i italià. Per l'altre costat, basc i hongarès es mantenen properes, però ja es diferencien clarament de l'alemany, anglès, danès holandès i noruec.
- Quarta coordenada: basc i hongarès són una a cada extrem i per tant es fa evident que són dues llengües ben diferents.

**Taula 3.**
Diferents mesures d'ajust calculades per la representació MDS ($k = 2$ i $k = 4$) i els dendrogrames (vegeu la definició i el significat de cada un dels coeficients en el text).

| | MDS $k$=2 | MDS $k$=4 | MÍNIM | MÀXIM | UPGMA |
|---|---|---|---|---|---|
| $c_1$ | 46,0 | 71,3 | | | |
| $c_2$ | 69,3 | 89,5 | | | |
| $S$ | 0,426 | 0,232 | 0,142 | 0,311 | 0,077 |
| $SS$ | 0,516 | 0,266 | 0,221 | 0,579 | 0,128 |
| $K$ | 0,348 | 0,200 | 0,108 | 0,198 | 0,076 |
| $r_c$ | 0,851 | 0,943 | 0,956 | 0,788 | 0,971 |

**Taula 4.**
Mesures d'ajust per a la representació MDS i els dendrogrames adequadament transformades per permetre la seva comparació.

| | MDS $k$=2 | MDS $k$=4 | MÍNIM | MÀXIM | UPGMA |
|---|---|---|---|---|---|
| $1-c_1/100$ | 0,540 | 0,287 | | | |
| $1-c_2/100$ | 0,307 | 0,105 | | | |
| $S^2$ | 0,182 | 0,054 | 0,020 | 0,097 | 0,006 |
| $SS^2$ | 0,266 | 0,070 | 0,049 | 0,335 | 0,016 |
| $K^2$ | 0,121 | 0,040 | 0,012 | 0,039 | 0,006 |
| $1-r_c^2$ | 0,275 | 0,111 | 0,086 | 0,379 | 0,057 |

L'aplicació dels tres algorismes d'anàlisi de conglomerats jeràrquica (mètode del mínim, del màxim i UPGMA) ha permès construir els dendrogrames representats en la figura 3. La correlació cofenètica i les altres mesures discutides en l'apartat anterior es mostren en la taula 3 i 4. Els resultats obtinguts són semblants qualitativament i no difereixen substancialment de les conclusions que s'han obtingut amb la MDS: un grup format per les llengües romàniques (català, castellà, francès, gallec i italià), el grup de llengües germàniques (alemany, anglès, danès, holandès i noruec) i quatre llengües que no poden ser agrupades (basc, finlandès, hongarès i polonès). Pot observar-se clarament en els dendrogrames fins i tot la divisió entre llengües germàniques del grup septentrional (noruec i danès) i del grup occidental (alemany, anglès i holandès). Novament s'ha fet palès l'enigma basc, testimoni viu del passat lingüístic d'occident i que tradueix la tenaç adhesió d'una petita comunitat a la seva llengua contra les fortes pressions exteriors suportades durant més de dos mil·lenis.

Finalment caldria algun comentari respecte les mesures d'ajust. Les principals conclusions que es poden extreure són les següents:

- La primera impressió que produeixen les taules 3 i 4 és una certa sorpresa: les diferències entre les distintes mesures són molt notables! Pot observar-se la variació entre $c_1$ i $c_2$, $K$ i $r_c$, $S$ i $SS$. La comparació alhora de totes les mesures provoca una certa incomoditat. Quina mesura és més adequada? És l'ajust bo o dolent? Com hem mencionat anteriorment no pretenem respondre aquestes preguntes, sinó fer palès el problema amb unes dades que són prou evidents. En tot cas, es desprèn d'aquests resultats que és convenient tenir en compte diverses mesures i no pas una de sola.
- De les representacions obtingudes, el dendrograma aconseguit mitjançant el mètode UPGMA és el que més s'ajusta a les dissimilituds inicials. El mètode del màxim i la representació MDS per $k = 2$ presenten unes discrepàncies considerables (es poden observar però les diferències entre els valors de $K$ i $r_c$). L'ajust de S millora substancialment augmentant la dimensió i considerant les quatre primeres coordenades.

## 7. CONCLUSIONS

Hem pogut comprovar en aquest estudi com l'anàlisi multivariant pot esdevenir una eina extraordinàriament eficaç en una àrea de recerca aparentment llunyana dels seus objectius com és el cas d'un estudi lingüístic. La quantificació acurada d'una informació qualitativa bastant reduïda (l'escriptura dels deu

primers nombres) ha permès agrupar adequadament catorze llengües europees, separant clarament les romàniques de les germàniques septentrionals i orientals. No ha estat l'objectiu obtenir conclusions rigoroses des d'un punt de vista lingüístic, sinó mostrar una metodologia que s'ha revelat apropiada. Un estudi més ampli i acurat, incorporant més llengües i un ventall més ampli de paraules, podria probablement aportar resultats encara més interessants. Tècniques semblants poden estendre's a d'altres estudis com ara l'evolució històrica d'una llengua o les variacions dialectals. La possibiliat de digitalitzar el so obre un altre possible camp d'aplicacions més sofisticades però essencialment fonamentades en mètodes similars. No obstant això, és el col·lectiu de lingüistes qui de ben segur sabrà trobar noves i més profitoses aplicacions.

## 8. AGRAÏMENTS

Els autors volen agrair molt especialment al Dr. C. M. Cuadras els suggeriments efectuats i el seu suport en el decurs de l'elaboració d'aquest treball. També agraïm la col·laboració de R. Serrano.

## 9. REFERÈNCIES BIBLIOGRÀFIQUES

[1] **Borg, I.**, i **Lingoes, J.** (1987). *Multidimensional Similarity Structure Analysis.* Springer-Verlag: New York.

[2] **Cuadras, C.M.** (1991). *Métodos de Análisis Multivariante.* PPU: Barcelona.

[3] **Davison, M.L.** (1983). *Multidimensional Scaling.* Wiley: New York.

[4] **Dillon, W.R.**, i **Goldstein, M.** (1984). *Multivariate Analysis. Methods and Applications.* Wiley: New York.

[5] **Eckart, C.**, i **Young, G.** (1936). "Approximation of one matrix by another of lower rank". *Psychometrika,* **1**, 211–218.

[6] **Farris, J.S.** (1969). "On the cophenetic correlation coefficient". *Syst. Zool.,* **18(3)**, 279–285.

[7] **Gower, J.C.** (1966). "Some distances properties of latent roots and vector methods used in multivariate analysis". *Biometrika,* **53**, 325–338.

[8] **Guttman, L.** (1968). "A general nonmetric technique for finding the smallest coordinate space for a configuration of points". *Psychometrika,* **33**, 469–504.

[9] **Jardine, N.**, i **Sibson, R.** (1968). "The construction of hierarchic and non-hierarchic classifications". *Comput. J.*, **11**, 177–184.
[10] **Johnson, S.C.** (1967). "Hierarchical clustering schemes". *Psychometrika*, **32**, 241–254.
[11] **Johnson, R.A.**, i **Wichern, D.W.** (1988). *Applied Multivariate Statistical Analysis.* Prentice-Hall International: New Jersey.
[12] **Kruskal, J.B.** (1964). "Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis". *Psychometrika*, **29**, 1–28, 115–129.
[13] **Kruskal, J.B., Young, F.W.**, i **Seery, S.B.** (1973). *How to use KYST, a very flexible program to do multidimensional scaling and unfolding.* Murray Hill: Unpublished manuscript, Bell Laboratories.
[14] **Lance, G.N.**, i **Williams, W.T.** (1966). "Computer programs for hierarchical polythetic classification ('similarity analysis')". *Comput. J.*, **9**, 60–64.
[15] **McQuitty, L.L.** (1964). "Capabilities and improvements of linkage analysis as a clustering method". *Educ. Psychol. Meas.*, **24**, 441–456.
[16] **Mardia, K.V.** (1978). "Some properties of classical multidimensional scaling". *Comm. Statist.-Theor. Meth.*, **A 7**, 1233–1241.
[17] **Mardia, K.V., Kent, J.T.**, i **Bibby, J.M.** (1979). *Multivariate Analysis.* Academic Press: London.
[18] **Richardson, M.W.** (1938). "Multidimensional psychophysics". *Psychol. Bull.*, **35**, 659–660.
[19] **Saito, T.** (1978). "The problem of the additive constant and eigenvalues in metric multidimensional scaling". *Psychometrika*, **43**, 193–201.
[20] **Schoenberg, I.J.** (1935). "Remarks to Maurice Fréchet's article 'Sur la definition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert'". *Ann. Math.*, **36**, 724–732.
[21] **Seber, G.A.F.** (1984). *Multivariate Observations.* Wiley: New York.
[22] **Sneath, P.H.A.** (1957). "The application of computers to taxonomy". *J. Gen. Microbiol.*, **17**, 201–226.
[23] **Sokal, R.R.**, i **Michener, C.D.** (1958). "A statistical method for evaluating systematic relationships". *Univ. Kansas Sci. Bull.*, **38**, 1409–1438.
[24] **Sokal, R.R.**, i **Sneath, P.H.A.** (1963). *Principles of Numerical Taxonomy.* Freeman: San Francisco.
[25] **Sokal, R.R.**, i **Rohlf, F.J.** (1962). "The comparison of dendrograms by objective methods". *Taxon*, **11**, 33–40.
[26] **Takane, Y., Young, F.W.**, i **De Leeuw, J.** (1977). "Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features". *Psychometrika*, **42**, 7–67.

[27] **Torgerson, W.S.** (1952). "Multidimensional scaling: I-theory and method". *Psychometrika,* **17**, 401–419.

[28] **Torgerson, W.S.** (1958). *Theory and Methods of Scaling.* Wiley: New York.

[29] **Young, F.W.**, i **Lewyckyj, R.** (1979). *ALSCAL 4 User's Guide.* Data Analysis and Theory Associates: Chapel Hill, NC.

# SOLUCIONS ALS PROBLEMES PROPOSATS AL VOLUM 16. N° 1, 2, 3

## PROBLEMA N° 44

*a)* El modelo lineal correspondiente a este problema es:

$$\begin{pmatrix} 10 \\ 35 \\ 45 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} \alpha \\ \\ \beta \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

o equivalentemente en forma esquemática:

$$y = X\boldsymbol{\beta} + e \tag{1}$$

donde $y$ es el vector de observaciones, $X$ la matriz de diseño, $\boldsymbol{\beta} = (\alpha, \beta)'$ el vector de parámetros y $e = (e_1, e_2, e_3)'$ el vector de errores.

Se supone que el vector $e$ sigue una distribución normal multivariante de vector de medias $\mathrm{E}(e) = 0$ y matriz de varianzas-covarianzas $\Sigma_e = \sigma^2\Omega$; es decir, $e \sim \mathrm{N}_3\left(0, \Sigma_e = \sigma^2\Omega\right)$.

Puesto que las covarianzas son

$$\begin{aligned} \operatorname{cov}(e_1, e_2) &= 0.6\sigma^2 \\ \operatorname{cov}(e_1, e_3) &= 0.2\sigma^2 \\ \operatorname{cov}(e_2, e_3) &= 0.2\sigma^2 \end{aligned}$$

concluimos que $e_1, e_2$ y $e_3$ no son estocásticamente independientes.

*b)* Estimación de los parámetros $\alpha$ y $\beta$ por MCO (mínimos cuadrados ordinarios).

La solución mínimo-cuadrática del vector de parámetros $\boldsymbol{\beta}$ viene dada por las ecuaciones normales:

$$X'X\hat{\boldsymbol{\beta}} = X'y$$

cuya solución es

$$\hat{\boldsymbol{\beta}}_{\mathrm{MCO}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X'X)^{-1}X'y$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{MCO}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \left[\begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}\right]^{-1} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 10 \\ 35 \\ 45 \end{pmatrix} = \begin{pmatrix} 3.75 \\ 8.75 \end{pmatrix}$$

*c)* Estimación de $\sigma^2$ y de la matriz de varianzas-covarianzas de $\hat{\beta}_{\text{MCO}}$ bajo las hipótesis del modelo lineal básico.

En el modelo lineal básico se supone que

$$e \sim \text{N}\left(0, \Sigma_e = \sigma^2 I\right)$$

es decir, la matriz de varianzas-covarianzas es escalar.

Un estimador insesgado de la varianza $\sigma^2$ del modelo lineal básico es:

$$\hat{\sigma}^2 = \frac{\mathbb{R}_0^2}{n-r} \tag{2}$$

siendo $\mathbb{R}_0^2$ la norma al cuadrado del vector de errores mínimo-cuadrático, $n$ es el número de observaciones y $r$ es el rango de la matriz de diseño.

Teniendo en cuenta (1)

$$y = X\hat{\beta} + \hat{e} \quad \text{y entonces}$$

$$\begin{aligned}
\mathbb{R}_0^2 &= \|\hat{e}\|^2 = \hat{e}'\hat{e} = \left(y - X\hat{\beta}\right)' \left(y - X\hat{\beta}\right) = \\
&= y'y - \hat{\beta}' X'y \\
\hat{\sigma}^2 &= \frac{\mathbb{R}_0^2}{n-r} = \frac{y'y - \hat{\beta}' X'y}{3-2} = \\
&= (10 \quad 35 \quad 45) \begin{pmatrix} 10 \\ 35 \\ 45 \end{pmatrix} - (3.75 \quad 8.75) \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 10 \\ 35 \\ 45 \end{pmatrix} = 37.5
\end{aligned}$$

La fórmula (2) no nos proporciona una estimación insesgada de la varianza $\sigma^2$ del modelo propuesto en que la matriz de varianzas-covarianzas de los errores no es escalar.

La estimación de la matriz de varianzas-covarianzas de $\hat{\beta}_{\text{MCO}}$ bajo las hipótesis del modelo lineal básico viene dada por

$$\widehat{\text{var}}\left(\hat{\beta}_{\text{MCO}}\right) = \hat{\sigma}^2 (X'X)^{-1} =$$

$$= 37.5 \left[ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix} \right]^{-1} =$$

$$= \begin{pmatrix} 54.69 & -14.06 \\ -14.06 & 4.69 \end{pmatrix}$$

*d)* Estimación de $\sigma^2$ utilizando los errores MCO.

Puesto que

$$\hat{e} = \mathcal{P}e, \quad \text{siendo} \quad \mathcal{P} = I - X(X'X)^{-1}X'$$

un proyector (matriz simétrica idempotente), se tiene

$$\begin{aligned} \text{E}(\hat{e}'\hat{e}) &= \text{E}(e'\mathcal{P}'\mathcal{P}e) = \\ &= \text{E}\,(e'\mathcal{P}e) = \\ &= \text{tr}\ \text{E}\,(e'\mathcal{P}e) = \\ &= \text{E}\,[\,\text{tr}\,(e'\mathcal{P}e)] = \\ &= \text{E}\,[\,\text{tr}\,\mathcal{P}ee'] = \\ &= \text{tr}\,\mathcal{P}\ \text{E}\,(ee') = \\ &= \text{tr}\,\mathcal{P}\sigma^2\Omega = \\ &= \sigma^2\ \text{tr}\,\mathcal{P}\Omega \end{aligned}$$

de forma que un estimador insesgado de $\sigma^2$ será:

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{\text{tr}\mathcal{P}\Omega} \tag{3}$$

La fórmula (3) nos proporciona un estimador insesgado de la varianza del modelo propuesto con matriz de varianzas-covarianzas de los errores no escalar, pero no es un estimador insesgado de la varianza del modelo lineal básico con matriz de varianzas-covarianzas de los errores escalar.

Como

$$\mathcal{P}\Omega = \left[I - X(X'X)^{-1}X'\right]\Omega =$$

$$= \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix} \left[ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \right\}.$$

$$\cdot \begin{pmatrix} 1 & 0.6 & 0.2 \\ 0.6 & 0.8 & 0.6 \\ 0.2 & 0.6 & 0.9 \end{pmatrix} = \begin{pmatrix} 0 & -0.067 & -0.017 \\ 0 & 0.133 & 0.033 \\ 0 & -0.067 & -0.017 \end{pmatrix}$$

la traza de $\mathcal{P}\Omega$ es:

$$\text{tr } \mathcal{P}\Omega = 0.133 - 0.017 = 0.116$$

y el estimador pedido viene dado por

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{\text{tr } \mathcal{P}\Omega} = \frac{37.5}{0.116} = 323.28$$

*e)* Estimación insesgada de la matriz de varianzas-covarianzas de los estimadores obtenidos en el apartado b).

La solución mínimo-cuadrática (MCO) del vector de parámetros $\boldsymbol{\beta}$ es

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (X'X)^{-1}X'y = \\ &= (X'X)^{-1}X'(X\boldsymbol{\beta} + e) = \\ &= \boldsymbol{\beta} + (X'X)^{-1}X'e \end{aligned}$$

de donde se obtiene

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (X'X)^{-1}X'e$$

La matriz de varianzas-covarianzas de $\hat{\boldsymbol{\beta}}$ viene dada por

$$\begin{aligned} \text{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right] &= \text{E}\left[(X'X)^{-1}X'ee'X(X'X)^{-1}\right] = \\ &= (X'X)^{-1}X' \text{ E}\,(ee')X(X'X)^{-1} = \\ &= (X'X)^{-1}X'\sigma^2\Omega X(X'X)^{-1} = \\ &= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} \end{aligned}$$

y utilizando (3) la estimación insesgada de esta matriz será

$$\begin{aligned} \widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_{\text{MCO}}\right) &= \hat{\sigma}^2(X'X)^{-1}X'\Omega X(X'X)^{-1} = \\ &= 323.28\left[\begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}\right]^{-1} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \cdot \\ &\quad \cdot \begin{pmatrix} 1 & 0.6 & 0.2 \\ 0.6 & 0.8 & 0.6 \\ 0.2 & 0.6 & 0.9 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}\left[\begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}\right]^{-1} = \\ &= \begin{pmatrix} 486.492 & -93.617 \\ -93.617 & 30.308 \end{pmatrix} \end{aligned}$$

*f)* Estimación de $\alpha$ y $\beta$ por MCG.

El modelo propuesto es

$$y = X\beta + e$$

con matriz de varianzas-covarianzas de los errores no escalar, $\Sigma_e = \sigma^2\omega$.

Como $\Omega$ es simétrica y definida positiva, también lo es $\Omega^{-1}$, y por tanto, existe una matriz cuadrada de orden $n$, no singular, $T$ tal que

$$T'T = \Omega^{-1}$$

La afirmación anterior es evidente, pues por la diagonalización de la matriz $\Omega^{-1}$ se tiene:

$$\Omega^{-1} = \mathcal{V}\mathcal{D}\mathcal{V}'$$

con $\mathcal{V}$ matriz ortogonal ($\mathcal{V}^{-1} = \mathcal{V}'$)

$$\Omega^{-1} = \mathcal{V}\mathcal{D}^{\frac{1}{2}}\mathcal{D}^{\frac{1}{2}}\mathcal{V}'$$

y haciendo $\mathcal{D}^{\frac{1}{2}}\mathcal{V}' = T$, resulta:

$$\Omega^{-1} = T'T$$

Si premultiplicamos el modelo propuesto

(4) $$y = X\beta + e$$

por $T$ tenemos que:

(5) $$Ty = TX\beta + Te$$

En (5) se cumple

$$\mathrm{E}\,(Te) = 0$$

y

$$\begin{aligned}
\Sigma_{Te} &= \mathrm{E}\,[Te(Te)'] = \\
&= \mathrm{E}\,(Tee'T') = \\
&= T\,\mathrm{E}\,(ee')T' = \\
&= \sigma^2 T\Omega T' = \\
&= \sigma^2 TT^{-1}(T')^{-1}T' = \\
&= \sigma^2 I \qquad (6)
\end{aligned}$$

Se ha comprobado en (6) que el término de error del modelo transformado tiene una matriz de varianzas-covarianzas escalar. La aplicación de MCO a (5) da lugar a las siguientes ecuciones normales:

$$X'T'TX\hat{\beta} = X'T'Ty \tag{7}$$

Teniendo en cuenta que $T'T = \Omega^{-1}$, la solución a este sistema de ecuaciones es:

$$\hat{\beta}_{MCG} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}y \tag{8}$$

Esta última expresión es el estimador por "mínimos cuadrados generalizados" (MCG), o estimador de Aitken.

Aplicando (8):

$$\hat{\beta}_{MCG} = \begin{pmatrix} \hat{\alpha}_{MCG} \\ \hat{\beta}_{MCG} \end{pmatrix} = \left[ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & 0.6 & 0.2 \\ 0.6 & 0.8 & 0.6 \\ 0.2 & 0.6 & 0.9 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix} \right]^{-1} \cdot$$

$$\cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & 0.6 & 0.2 \\ 0.6 & 0.8 & 0.6 \\ 0.2 & 0.6 & 0.9 \end{pmatrix}^{-1} \begin{pmatrix} 10 \\ 35 \\ 45 \end{pmatrix} =$$

$$= \begin{pmatrix} 1.786 \\ 8.214 \end{pmatrix}$$

*g)* Estimación de $\sigma^2$ y de la matriz de varianzas-covarianzas de los estimadores $\beta$ por MCG.

$$\hat{\sigma}^2_{MCG} = \frac{\hat{e}'_{MCG}\hat{e}_{MCG}}{n-r} =$$

$$= \frac{\left(y - X\hat{\beta}_{MCG}\right)'\left(y - X\hat{\beta}_{MCG}\right)}{n-r} =$$

$$= \frac{\left(y' - \hat{\beta}'_{MCG}X'\right)\left(y - X\hat{\beta}_{MCG}\right)}{n-r} =$$

$$= \frac{1}{3 \cdot 2}\left[(10 \quad 35 \quad 45) - (1.786 \quad 8.214)\begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix}\right] \cdot$$

$$\cdot \left[\begin{pmatrix} 10 \\ 35 \\ 45 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}\begin{pmatrix} 1.786 \\ 8.214 \end{pmatrix}\right] =$$

$$= 78.061$$

Y finalmente, la estimación de la matriz de varianzas-covarianzas de $\hat{\boldsymbol{\beta}}_{\text{MCG}}$ es

$$\begin{aligned}
\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_{\text{MCG}}\right) &= \hat{\sigma}^2_{\text{MCG}}\left(X'\Omega^{-1}X\right)^{-1} = \\
&= 78.061\left[\begin{pmatrix}1 & 1 & 1\\ 1 & 3 & 5\end{pmatrix}\begin{pmatrix}1 & 0.6 & 0.2\\ 0.6 & 0.8 & 0.6\\ 0.2 & 0.6 & 0.9\end{pmatrix}^{-1}\begin{pmatrix}1 & 1\\ 1 & 3\\ 1 & 5\end{pmatrix}\right]^{-1} = \\
&= \begin{pmatrix}116.534 & -22.861\\ -22.861 & 7.249\end{pmatrix}.
\end{aligned}$$

Pedro Sánchez
Universitat de Barcelona

## PROBLEMA N° 45

Llamemos "msc" y "mas" a los diseños comparados. Si la población $U$ es de tamaño $N = 5$ de manera que la variable de interés $y_p = p$, $p = 1, 2, \ldots .5$. Sea la muestra de tamaño efectivo fijo $n = 2$:

$$V(\text{mas}, \overline{y}) = 3/4 \text{ para } 10 \text{ muestras posibles.}$$

Si $U_1 = U = \{1, 2, 3, 4, 5\}$ o cualquier permutación circular de ella, $V(\text{msc}, \overline{y}) = 1$ para 5 muestras posibles, pero si el orden en $U$ fuera $U_2 = \{1, 4, 2, 5, 3\}$,

$$V(\text{msc}, \overline{y}) = 3/10 \text{ para } 5 \text{ muestras posibles,}$$

lo cual muestra que la estrategia $(\text{mas}, \overline{y})$ puede ser menos eficiente que $(\text{msc}, \overline{y})$.

En la práctica, la reordenación de las unidades en $U$ se puede basar en el conocimiento de una variable auxiliar "$x$" fuertemente correlacionada con "$y$".

M. Ruiz Espejo
Universidad Complutense de Madrid

**PROBLEMA N° 46**

Sea $S = \{s_{ij} \subset U: 1 \leq i < j \leq 2k,\ s_{ij}$ muestra sistemática de doble arranque$\}$,

$p(s_{ij}) = \dfrac{1}{k(2k-1)}$ siendo $i$ y $j$ números naturales para toda $s_{ij} \in S$.

La media muestral será

$$\overline{y}_{ij} = \frac{1}{n} \sum_{p \in s_{ij}} y_p .$$

Ahora,

$$(1) \qquad V(\overline{y}_{ij}) = \mathrm{E}(\overline{y}_{ij}^2 - \overline{y}^2 .$$

$$\mathrm{E}(\overline{y}_{ij}) = \mathrm{E}\left[\left(\frac{1}{n}\sum_{p \in s_{ij}} y_p\right)^2\right] = \frac{1}{n^2} \sum_{s_{ij} \in S} \left(\sum_{p \in s_{ij}} y_p\right)^2 p(s_{ij}) =$$

$$= \frac{1}{n^2} \cdot \frac{1}{k(2k-1)} \sum_{s_{ij} \in S} \left(\sum_{p \in s_{ij}} \sum_{m \in s_{ij}} y_p y_m\right) =$$

$$(2) \qquad = \frac{1}{n^2 k(2k-1)} \sum_{p \in U} \sum_{m \in U} y_p y_m \,\mathrm{card}\,\{s_{ij} \in S: p, m \in s_{ij}\}$$

donde

$$\mathrm{card}\,\{s_{ij} \in S: p, m \in s_{ij}\} = \begin{cases} 2k-1 & \text{si } p = m \text{ mód } (2k) \\ 1 & \text{si } p \neq m \text{ mód } (2k). \end{cases}$$

Por tanto, de (1) y (2),

$$V(\overline{y}_{ij}) = \sum_{p \in U} \sum_{m \in U} f_{pm} y_p y_m$$

donde

$$f_{pm} = \frac{1}{n^2 k(2k-1)} \,\mathrm{card}\,\{s_{ij} \in S: p, m \in s_{ij}\} - \frac{1}{N^2}.$$

Un estimador insesgado de $V(\overline{y}_{ij})$ será (Ruiz, 1986)

$$\hat{V}(\overline{y}_{ij}) = \sum_{p \in U} \sum_{m \in U} f_{pm} y_p y_m e_{pm} / \pi_{pm}$$

donde $e_{pm}$ es una variable aleatoria definida

$$e_{pm} = \begin{cases} 1 & \text{si } p, m \in s_{ij} \\ 0 & \text{si } p \text{ ó } m \notin s_{ij}, \end{cases}$$

que verifica $\mathrm{E}(e_{pm}) = \pi_{pm}$.

Es inmediato comprobar que $\pi_{pm} > 0$ para todo $p, m \in U$ pues

$$\pi_{pm} = \begin{cases} 1/k = n/N & \text{si } p = m \text{ mód } (2k) \\ 1/[k(2k-1)] & \text{si } p \neq m \text{ mód } (2k). \end{cases}$$

## Referencias


**Gautschi, W.** (1957). "Some remarks on systematic sampling". *Ann. Math. Statist.*, **28**, 385–394.

**Ruiz, M.** (1986). "Funciones paramétricas estimables en teoría de muestras". *Estadíst. Española,* **28**, nº 112–113, 69–73.



M. Ruiz Espejo
Universidad Complutense de Madrid

## COMENTARI DE LLIBRE

**T.P. Hutchinson and C.D. Lai.**

**THE ENGINEERING STATISTICIAN'S GUIDE TO CONTINUOUS BIVARIATE DISTRIBUTIONS**

Rumsby Scientific Publishing, Adelaide, XXII + 346 pp. + 19 tablas.

Estando próxima la celebración de la conferencia sobre Distribuciones con Marginales fijas, Medidas doblemente Estocásticas y Operadores de Markov (Seattle, USA, agosto 1993) , parece oportuno un comentario sobre el libro de Hutchinson y Lay. Como su título indica, lo dice casi todo sobre las distribuciones de probabilidad bivariantes y sus múltiples aplicaciones.

En el capítulo 1 se resalta la cantidad de modelos probabilísticos sobre distribuciones univariantes que han sido bien estudiados, pero que el caso de dos dimensiones casi se reduce a la distribución normal bivariante. Seguidamente se citan algunos ejemplos tempranos con datos que no siguen tal omnipresente distribución.

El capítulo 2 introduce las notaciones, las principales funciones (distribución $H$, densidad, características, jacobianos, etc.). El capítulo 3 explica los métodos de construir distribuciones (mixturas, composición, sumas aleatorias, expansiones, etc.). Desde luego se introduce el caso de independencia y las cotas de Fréchet, haciendo énfasis que éstas proporcionan máxima correlación, que equivalen a una relación funcional entre las variables, y se advierte sobre los métodos "poco elegantes" de construir distribuciones bivariantes.

El capítulo 4 se destina a los modelos de fiabilidad, colas de espera, etc., donde las marginales son de tipo exponencial, Weibull o valores extremos. La distribución central es la de Marshall-Olkin, que se relaciona con muchas otras (Cuadras-Augé, Block-Basu, Sarkar, Friday-Patil, etc.), las cuales se pueden obtener de ésta por traslación, mixtura, etc. Otras distribuciones con marginales exponenciales también comentadas son las de Raftery y Downton. Se mencionan además numerosas aplicaciones. El capítulo 5 es una continuación del anterior, pero haciendo mayor énfasis en el concepto de dependencia estocástica, sus diferentes versiones y sus interrelaciones e interpretación geométrica.

El otro principal método de construir distribuciones bivariantes (modelando la densidad de $Y$ condicionada a $X = x$) es el tema del capítulo 6, en el que se destaca la distribución de McKay (marginales gamma y condicional beta) y se citan numerosas aplicaciones (datos sobre la velocidad del viento, lluvias, inundaciones). Se hace especial mención a las contribuciones de Arnold, Castillo y Galambos. El capítulo 7 está destinado a la distribución Pareto bivariante, que se introduce como una exponencial compuesta con una gamma, algunas generalizaciones y aplicaciones. Las distribuciones de valores extremos aparecen en el capítulo 8, empezando con el caso univariante (que se justifica por las áreas de interés: metereología, ingeniería). El caso bivariante se estudia centrándose en la propiedad de que la función de distribución es de valores extremos si las potencias $H^n$ son también distribuciones (obsérvese que esto siempre es cierto en el caso univariante). El análisis de datos composicionales o estudio de proporciones bivariantes con marginales esencialmente tipo beta, viene en el capítulo 9, que contiene numerosos ejemplos.

De interés fundamental es el capítulo 10, dedicado a las "cópulas", distribuciones bivariantes con marginales uniformes (también se habla de "representación uniforme", "función de dependencia") y a las "cópulas arquimedianas". Las cópulas más estudiadas son la Farlie-Gumbel-Morgenstern, Ali-Mikhail-Haq, Frank, Plackett y Pareto. Sorprende que esta última cópula, que creíamos atribuida a Clayton y Oakes (de reconocida importancia en el estudio de datos de supervivencia) se conecta con la Pareto vía representación uniforme. Las cópulas que además son arquimedianas (es decir, $g(H) = g(F) + g(G)$, siendo $H$ la conjunta, $F, G$ las marginales y $g$ una función) son interesantes porque ciertas propiedades pueden estudiarse fácilmente a partir de la función $g$. El estudio más interesante se debe a Genet-McKay, posteriormente extendido por Marshall-Olkin. Las cópula de Pareto-Clayton-Oakes, Frank y Ali-Mikhail-Haq son arquimedianas.

El estudio de la distribución normal bivariante, la más importante, nos llega en el capítulo 11, del que se hace un estudio bastante completo (regresión, simulación, propiedades estadísticas, truncación, aplicaciones) no sin antes resaltar que a veces es cuestionable el ajuste de datos empíricos a esta distribución. El siguiente capítulo es una miscelánea de otras distribuciones algo difícil de resumir.

La medida de la correlación (Pearson, Kendall, Spearman), en relación con las diferentes distribuciones, es objeto del capítulo 13. Quizás faltaría en este capítulo una referencia al resultado de Kimeldorf-Sampson: la cópula independencia $H(x, y) = xy$ se puede aproximar arbitrariamente por otra expresando perfecta dependencia entre ambas variables. No se puede distinguir estadísticamente entre ambas cópulas. Este sorprendente resultado fué generalizado por Vitale (cualquier cópula admite esta aproximación).

Finalmente, los capítulos 14 y 15 están destinados a la simulación de distribuciones bivariantes y a cómo entender e interpretar los datos bivariantes.

Se trata de una obra recomendable, más orientada a la consulta que a la didáctica, relacionando muchas distribuciones entre sí (la principal baza del libro) y que es una versión más manejable del libro CONTINUOUS BIVARIATE DISTRIBUTIONS, EMPHASISING APPLICATIONS (1990) de los mismos autores y misma editorial, con contenido similar pero orientación más académica, y que también contiene muchos ejemplos.

C.M. Cuadras

# NOVETATS DE SOFTWARE

La introducció de la nova secció de "NOVETATS DE SOFTWARE" a la revista QÜESTIIÓ es fa amb la finalitat de promoure l'intercanvi d'informació relacionada amb programes d'ordinador disponibles, destinats a l'implementació de metodologia estadística, d'informàtica o d'investigació operativa.

A causa de l'important creixement que ha experimentat darrerament la utilització dels ordinadors a totes les àrees científiques i técniques i, a les esmentades més amunt, en particular, hi ha un bon nombre d'investigadors que han desenvolupat un software propi, l'existència del qual és desconeguda, de vegades, per a molts lectors que el podríen aprofitar. Per això, creiem que és convenient i útil fer-lo conèixer mitjançant aquesta revista, amb el benentés que només actuaria com a mitjà de difusió.

Per tal d'uniformitzar la descripció del software, adjuntem una butlleta que ha de ser omplenada i tramesa a l'editorial de QÜESTIIÓ.

Amb tota certesa, la vostra col.laboració serà d'utilitat per a molts lectors als qui facilitarà el treball i que, alhora, podràn ajudar els autors del programes suggerint-los possibles millores.

**Nom del programa:**

**Area/àrees d'aplicació (Estadística, Sistemes, etc.):**

**Descripció del software:**

- Llenguatge:
- Ordinador/s:
- Sistema operatiu:

**Està disponible en els suports següents:**

Floppy disk/diskette. Assenyaleu:

Mida: Densitat: una dues cares

Cinta magnètica. Assenyaleu:

Mida Densitat Codi

**Distribuït per:**

**Configuració mínima de hardware requerida:**

**Requereix l'ensinistrament de l'usuari:**

**Documentació**

**Llistat, font disponible:**

**Grau de desenvolupament:**

**Es fa servir aquest software normalment?**

**En cas afirmatiu**

des de quan?

a quants llocs?

**L'autor d'aquest software està disponible per atendre les preguntes dels usuaris?**

**Descripció del que fa l'esmentat software:** (200 paraules aproximadament).

**Posibles usuaris:**

**Camps d'interès:**

**Nom de l'autor/s:**

**Institució:**

**Adreça:**

**Número de telèfon:**

# RESUMS EN ANGLÈS

JOSÉ A. VILAR and JUAN M. VILAR

**Estimation of the Probability density from random sampling.**

Let $X(t)$ be a stationary continuous-time processes, from discrete-time samples $X(\tau_1), X(\tau_2), \ldots, X(\tau_n)$, with sampling instants, $\tau_i$, irregularly spaced or random, recursive estimation of the univariate probability density function, $f(x)$, for process $X(t)$ is studied when the observations satisfy a strong mixing condition. Estimators of the kernel type are considered.

Some asymptotic expressions are obtained for the bias and variance-covariance and asymptotic normality is proven.

SADAO TOMIZAWA

**A simple statistic to test generalized palindromic symmetry model in a 4 × 4 contingency table.**

For a 4 × 4 contingency table, this note gives a simple statistic to test the goodness-of-fit of the generalized palindromic symetry (GPS) model considered by McCullagh (1978). Also an asymptotic confidence interval for a parameter of interest in the GPS model is given. Two sets of unaided vision data are used as example.

C.M. CUADRAS and J. FORTIANA

**Applying distances in statistics.**

This paper is a review of the relevance in statistics of some geometrical ideas based on the concept of distance. Its application to four statistical topics is presented and discussed: point estimation, testing hypotheses, geometric representation of sets and prediction methods.

LAURA MOTA HERRANZ and MATILDE CELMA GIMÉNEZ

**Methods for Integrity Checking in Deductive Databases.**

Integrity checking is a classical problem in field of databases: The first methods were proposed to simplify static constraint checking in relational databases being extended to deductive databases later. These method are based on the idea of evaluating simplified instances of the integrity constraints generated by the updates induced by the transaction and differ mainly in the strategy used to instance and simplify the constraints.

In this paper, we classify and analyze the most important methods found in the literature.

FRANCESC CARRERAS

**Cooperation and National Defense.**

Concepts and techniques of cooperative game theory are applied to decision problems concerning national defense policy. Our analysis includes an evaluation of qualified voting procedures that were submitted to the European Council at Maastricht (the Netherlands) in December 1991. Implications for Spain's strategic position, which would have derived from the inedited operative capability appointed to the European Community by the political union project, are also described.