Enfoque basado en Distancias de algunos Métodos Estadísticos Multivariantes

Josep Fortiana

October 30, 2001

Contents

Chapter 1

Aplicación de las distancias en Estadística

1.1 Introducción

Desde su principio, la Estadística moderna ha dependido de la Teoría de 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 importante 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 introducida por Mahalanobis [72]

$$(x-y)' \cdot \Sigma^{-1} \cdot (x-y) \tag{1.1}$$

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

La distancia (1.1) 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 1.3.1, interviene en cualquier contraste de hipótesis.

Como esta memoria es una aplicación de ciertas propiedades de las distancias, nos parece oportuno citar los trabajos de Hotelling [59] y Weyl [101], pioneros en la aplicación de la Geometría Diferencial al contraste de hipótesis: Dado el modelo de regresión no lineal

$$y_i = \beta\, f_i(\theta) + e_i \qquad (i = 1, \ldots, n) \tag{1.2}$$

donde las $f_i(\theta)$ son funciones conocidas que dependen de un parámetro θ y los errores $e_1, \ldots, e_n$ son variables aleatorias independientes igualmente distribuidas (iid), con distribución $N(0, \sigma)$, consideremos la hipótesis nula

$$H_0: \quad \beta = 0$$

Es fácil ver que el estadístico Λ de razón de verosimilitud equivale a

$$W = \max_{\theta} \frac{\left(\sum_i f_i(\theta)\, y_i\right)^2}{\sum_i f_i^2(\theta)\ \sum_i y_i^2} \tag{1.3}$$

Sin embargo, puesto que θ no es identificable cuando $\beta = 0$, no es factible aplicar la teoría asintótica sobre la distribución de Λ, ni tampoco los criterios equivalentes de Wald y de Rao, que están asintóticamente distribuidos como Ji–cuadrado. Véase Rao [90, pág. 417] y la sección 1.3.3 de esta memoria.

Empleando las notaciones:

$$\begin{aligned} f(\theta) &= (f_1(\theta), \ldots, f_n(\theta)) \\ y &= (y_i, \ldots, y_n) \\ \gamma(\theta) &= f(\theta)/\|f(\theta)\| \\ U &= y/\|y\| \end{aligned}$$

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 $\mathbf{R}^n$. Véase Knowles and Siegmund [66].

Es éste un ejemplo, nada trivial, de la Estadística y el Análisis de Datos, de entre los innumerables ejemplos en los que se aplica el concepto de distancia. En este capítulo introductorio presentamos una breve exposición de su aplicación en los siguientes campos:

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

1.2 Estimación puntual

1.2.1 En modelos lineales

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

$$y = X \cdot \beta + e \tag{1.4}$$

donde la estimación del vector paramétrico β es aquel $\widehat{\beta}$ tal que $\widehat{y} = X\widehat{\beta}$ verifica

$$R_0{}^2 = \| y - \widehat{y} \|^2 = \text{ mínimo} \tag{1.5}$$

Además, si $e \sim N(0, \sigma I_n)$, entonces se verifica que $R_0{}^2/\sigma^2 \sim \chi^2{}_{n-r}$, siendo $r = \text{rang}\,(X)$, resultado básico del Análisis de la Varianza.

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

$$H_0 : \quad \Psi = \Psi_0 \tag{1.6}$$

se decide mediante el test F

$$F = \frac{(\widehat{\Psi} - \Psi_0)' \cdot (P \cdot (X'X)^- \cdot P')^{-1} \cdot (\widehat{\Psi} - \Psi_0)}{R_0{}^2} \times \frac{n-r}{q} \tag{1.7}$$

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

1.2.2 Divergencia de Kullback–Leibler

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

$$K(p, q) = \int p \, \log(\frac{p}{q}) \, d\mu \tag{1.8}$$

juega un importante papel en el llamado problema de *la especificación* en inferencia estadística.

Supongamos, para concretar, que μ sea la medida de Lebesgue, y sea

$$\Gamma = \{ p\,(x, \theta), \quad \theta \in \Theta \}$$

un modelo estadístico. La verdadera función de densidad es $p\,(x, \theta_0)$, donde θ_0 es el verdadero valor del parámetro. La divergencia entre $p\,(x, \theta)$ y $p\,(x, \theta_0)$

es

$$\begin{aligned} K\left(p\left(x,\theta\right),\, p\left(x,\theta_0\right)\right) &= \int p\left(x,\theta_0\right)\log p\left(x,\theta_0\right)dx \\ &- \int p\left(x,\theta_0\right)\log p\left(x,\theta\right)dx \end{aligned} \tag{1.9}$$

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

$$\int p\left(x,\theta\right)\log p\left(x,\theta_0\right)dx \tag{1.10}$$

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

Dada una muestra $x_1,\ldots,x_n$ de valores *iid* con densidad $p\left(x,\theta_0\right)$, este valor se obtiene mediante el promedio

$$\frac{1}{n}\sum_{i=1}^{n}\log\, p(x_i,\theta) \tag{1.11}$$

El valor $\widehat{\theta}$ que maximiza este promedio nos lleva al estimador máximo verosímil (ML) de θ.

Menos conocida es la siguiente propiedad. Supongamos que la verdadera densidad es q, pero $q \notin \Gamma$. ¿Qué significaría entonces la estimación ML de θ? La divergencia entre q y $p(x,\theta)$ es ahora

$$\int q(x)\log q(x)dx - \int q(x)\log p\left(x,\theta\right)dx \tag{1.12}$$

y el *verdadero valor* θ_0 del parámetro θ se puede definir como aquel θ_0 tal que $p\left(x,\theta_0\right) \in \Gamma$ es la densidad más próxima a q de acuerdo con la divergencia (1.8). θ_0 es entonces solución de

$$E_q\left\{\frac{\partial}{\partial\theta}\log p\left(x,\theta\right)\right\} = 0 \tag{1.13}$$

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

$$\begin{aligned} U(x,\theta) &= \frac{\partial}{\partial\theta}\log p(x,\theta) \\ J(\theta) &= E_q(U\cdot U') \\ H(\theta) &= -E_q\left(\frac{\partial U}{\partial\theta}\right) \end{aligned} \tag{1.14}$$

En un entorno de θ_0 tendremos

$$U(x,\theta) = U(x,\theta_0) + (\theta - \theta_0)\left(\frac{\partial U}{\partial \theta}\right)_{\theta_0} + K$$

y si $x_1, \ldots, x_n$ son iid como q, entonces

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

y haciendo tender $n \longrightarrow \infty$, teniendo en cuenta (1.13) y (1.14), se cumple la identidad asintótica

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

que prueba que $\widehat{\theta}$, el estimador ML que anula $\sum U(x_i,\theta) = 0$, converge a θ_0 en probabilidad.

Además, por el teorema del valor medio podemos escribir

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

donde $\theta^\star$ es un punto entre θ y $\widehat{\theta}$.

Puesto que

$$\begin{aligned}
\sum U(x_i,\widehat{\theta}) &\longrightarrow 0 \\
\widehat{\theta} &\longrightarrow \theta_0 \\
\frac{1}{n}\sum\frac{\partial U(x_i,\theta)}{\partial \theta} &\longrightarrow H(\theta)
\end{aligned}$$

tenemos de nuevo la identidad asintótica

$$\begin{aligned}
\frac{1}{n}\sum U(x_i,\theta) &= n\,(\widehat{\theta} - \theta_0)\, H(\theta_0) \\
&\text{es decir,} \\
\frac{1}{\sqrt{n}}\sum U(x_i,\theta_0) &= \sqrt{n}\,(\widehat{\theta} - \theta_0)\, H(\theta_0)
\end{aligned}$$

Por el teorema central del límite, $\frac{1}{\sqrt{n}}\sum U(x_i,\theta_0)$ es asintóticamente normal de media $E_q\,(U(x,\theta_0) = 0$, que es la condición (1.13), y matriz de covarianzas $J(\theta_0)$. Finalmente tenemos que

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

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

Ventajas y aplicaciones de la estimación ML del *verdadero valor* θ_0 pueden verse en Kent [65] para el estudio de la robustez del test de razón de verosimilitud tomando densidades alternativas, en Royal [94], para la obtención de intervalos de confianza robustos, en Huster [60] para estimar parámetros en modelos bivariantes de supervivencia y en Cuadras [29] para el problema de la estimación de parámetros relativos a densidades multivariantes cuando sólo se conocen las marginales.

1.2.3 El método de la mínima distancia

Es un método de estimación promovido por J. Wolfowitz en una serie de artículos que culminaron en [102]. Supongamos de la forma de la función de distribución de un vector aleatorio es $G \in \Gamma = \{F_\theta,\, \theta \in \Theta\}$. Sean $x_1, \ldots, x_n$ *iid* como G, y sea G_n la función de distribución empírica. Si $\delta\,(G_n, F_\theta)$ es una medida de distancia entre G_n y $G = F_\theta$, el método de la mínima distancia (MD) consiste en tomar como estimación de θ el valor $\widehat{\theta}$ tal que

$$\delta\,(G_n,\, F_{\widehat{\theta}}) = \inf_{\theta \in \Theta} \delta\,(G_n,\, F_\theta)$$

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

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

o la de Cramér–von Mises

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

MD proporciona estimadores que convergen en probabilidad a θ y tienen propiedades de robustez en el caso de desviaciones locales del modelo. Incluso, si $G \notin \Gamma$, tomando δ_C con $w_\theta(x) = 1/f_\theta(x)$, el estimador MD proporciona una estimación $\widehat{\theta}$ tal que $F_{\widehat{\theta}}$ es una proyección L^2 de G_n en Γ. Véase Parr [85].

El método de estimación MD constituye una herramienta 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

$$E\left(\widehat{f}_n(x)\right) = f(x), \qquad \forall x$$

(cfr. Prakasa Rao [86]). Así, la teoría clásica de la estimación no funciona, 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 $h_n \longrightarrow 0$ para $n \longrightarrow \infty$, 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 $f_n(x)$ converge uniformemente a $f(x)$.

Un criterio de proximidad en la estimación de $f(x)$ se basa en la distancia 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. Véase Devroye and Györfi [33].

Finalmente, el método MD es también útil para estimar θ en el modelo de regresión lineal

$$y = A'(x)\,\theta + e$$

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 [47].

1.3 Contraste de hipótesis

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

1.3.1 Distancia de Mahalanobis

La aplicación más clara de esta distancia la encontramos en el test T^2 de Hotelling. Supongamos que $x_1, \ldots, x_n$ son *iid* según $N_p(\mu, \Sigma)$, con Σ desconocido, y estamos interesados en el contraste

$$H_0: \quad \mu = \mu_0 \tag{1.15}$$

Tanto el test de razón de verosimilitud como el principio de unión–intersección (véase Mardia [78]) nos llevan a considerar el estadístico

$$T^2 = n\,(\overline{x} - \mu_0)' \cdot S^{-1} \cdot (\overline{x} - \mu_0) \tag{1.16}$$

donde $\overline{x}$, S son la media y covarianza muestrales. T^2, bajo H_0, es proporcional a una F. Así el test T^2 está basado en la distancia de Mahalanobis entre $\overline{x}$ y μ_0.

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

$$H_0: \quad \mu_1 = \mu_2 \tag{1.17}$$

También los criterios clásicos nos llevan al estadístico

$$T^2 = \frac{nm}{m+n}\,(\overline{x} - \overline{y})' \cdot S^{-1} \cdot (\overline{x} - \overline{y}) \tag{1.18}$$

donde $\overline{x}$, $\overline{y}$, S son los estimadores usuales de μ_1, μ_2, Σ ([80], [78]), es decir, a la T^2 de Hotelling, que es también función de la estimación de la distancia de Mahalanobis entre μ_1 y μ_2.

Más generalmente, consideremos el modelo del Análisis Multivariante de la Varianza (MANOVA), con n observaciones de p variables

$$Y = X \cdot B + E$$

donde Y es $n \times p$, X es $n \times m$ (siendo m el número de parámetros), B es $m \times p$ y E es $n \times p$. Sea

$$\Psi' = (\psi_1, \ldots, \psi_p) = P' \cdot B$$

una función paramétrica estimable multivariante, y consideremos el contraste de hipótesis

$$H_0: \quad \Psi = \Psi_0 \tag{1.19}$$

donde Ψ_0 es conocido. Entonces, si $\widehat{\Psi}$ es el estimador Gauss–Markov

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

y $\widehat{\Sigma}$ es la estimación centrada de Σ (se supone que las filas de E son *iid* $N_p(0, \Sigma)$)

$$\widehat{\Sigma} = \frac{1}{n-r}\,(Y - X\widehat{B})' \cdot (Y - X\widehat{B})$$

donde $r = \operatorname{rang}(\Sigma)$, entonces el test (1.19) se puede decidir mediante el estadístico

$$(\widehat{\Psi} - \Psi_0)' \cdot \widehat{\Sigma}^{-1} \cdot (\widehat{\Psi} - \Psi_0) \tag{1.20}$$

que es una distancia tipo Mahalanobis y cuya distribución bajo H_0 es también proporcional a una F (Cuadras [19] y [20]).

En un contexto parecido, la distancia entre dos modelos MANOVA

$$Y_i = X \cdot B_i + E_i \quad i = 1,\, 2 \tag{1.21}$$

se puede definir como

$$L^2 = \operatorname{tr}\left\{\Sigma^{-1} \cdot (B_1 - B_2)' \cdot X' \cdot X \cdot (B_1 - B_2)\right\} \tag{1.22}$$

que puede justificarse como una distancia de Mahalanobis entre dos distribuciones normales $N_p(I_p \otimes X \cdot B_i, \Sigma \otimes I_n)$, $\quad (i = 1,\, 2)$.

Como $L^2 = 0$ si y sólo si $X \cdot B_1 = X \cdot B_2$, la distancia (1.22) puede servirnos para contrastar la hipótesis

$$H_0: \quad X \cdot B_1 = X \cdot B_2$$

de que los dos modelos de regresión (1.21) son iguales. Para más detalles y generalizaciones, véase Ríos y Cuadras [92].

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

$$H_0: \quad \theta \in \Theta_0 \subset \Theta \tag{1.23}$$

Dada una muestra $x_1, \ldots, x_n$, el procedimiento clásico iniciado por Neyman y Pearson [81] para decidir acerca de H_0 utiliza el test de razón de verosimilitud

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}}{\sup_{\theta \in \Theta} \mathcal{L}} \tag{1.24}$$

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

$$U = -2 \log \Lambda \tag{1.25}$$

el cual, 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 [89]. (Véase, por ejemplo, Rao [90]). Se basa en los llamados *efficient scores*

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

y en el comportamiento de

$$V_\theta = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(\theta)$$

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 ${\chi^2}_q$.

El estadístico que propone Rao es

$$S = {V_{\theta^\star}}' \cdot {\mathcal{F}_{\theta^\star}}^{-1} \cdot V_{\theta^\star} \tag{1.26}$$

donde $\theta^\star$ representa la estimación máximo verosímil de θ dentro de Θ_0.

Podemos poner

$$V_{\theta^\star} = \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} Z_i(\theta^\star) = \sqrt{n} \cdot \overline{Z}_{\theta^\star}$$

y como bajo H_0, $\mathcal{F}_{\theta^\star}$ puede considerarse una estimación de $\mathcal{F}_{\theta_0}$, donde θ_0 representa el verdadero valor del parámetro, tenemos que la proximidad de $V_{\theta^\star}$ a $V_{\theta_0} = 0$ favorece la hipótesis nula. Pero tal proximidad la podemos medir mediante la distancia de Mahalanobis entre $\overline{Z}_{\theta^\star}$ y la media esperada 0, es decir, mediante

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

Ahora bien, según se muestra en Rao [90], se cumple la siguiente igualdad asintótica:

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

que viene a probarnos que la razón de verosimilitud, el estadístico más utilizado en contrastes de hipótesis, es asintóticamente equivalente a una distancia de Mahalanobis, pues en definitiva, $\mathcal{F}_{\theta^\star}$ es la estimación de una matriz de covarianzas.

Como ilustración, sea

$$p\,(x,\theta) = \theta^{-1}\exp(-\theta^{-1}\,x), \quad \theta > 0$$

y consideramos la hipótesis nula

$$H_0: \quad \theta = \theta_0,$$

donde $\theta_0 \in \Theta = \mathbf{R}_+$. La razón de verosimilitud es

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

Mientras que el estadístico de Rao es

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

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

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

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

1.3.2 Distancia de Matusita

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

$$\delta^2(F_1, F_2) = \int \left\{ \sqrt{f_1(x)} - \sqrt{f_2(x)} \right\}^2 dx = 2\,(1-\rho), \tag{1.27}$$

donde

$$\rho = \int \sqrt{f_1(x)\, f_2(x)}\, dx$$

es la llamada *afinidad* entre F_1 y F_2.

La distancia (1.27), introducida por Matusita [74], aunque es también conocida como distancia de Hellinger, ha sido aplicada en problemas de estimación, decisión y análisis discriminante. Por ejemplo, teniendo en cuenta que

$$F_1 = F_2 \iff \delta^2(F_1, F_2) = 0$$

el contraste de hipótesis, en el caso univariante,

$$H_0 : F_1 = F_2$$

es equivalente a

$$H_0 : \delta^2(F_1, F_2) = 0$$

Se acepta H_0 si $\delta^2(F_1, F_2) \leq \delta_\epsilon$, donde δ_ϵ es una cantidad positiva que dependerá del nivel de significación ϵ y de los tamaños muestrales m, n. La decisión se toma, sin embargo, trabajando con la distancia $\delta^2(S_1, S_2)$ entre las funciones de distribución empíricas.

Matusita [76] discute extensamente la utilización de la distancia (1.27) en el caso normal univariante $N(\mu, \Sigma)$. Consideremos algunos ejemplos:

1) La hipótesis $H_0 : \mu = \mu_0$ se decide a través de $\delta^2(F, S_n)$, donde F en $N(\mu_0, \Sigma)$, y S_n es $N(\widehat{x}, S)$, siendo $\widehat{x}$ 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 $N(\mu, \Sigma_0)$

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

3) La hipótesis de que $X = (x_1, \ldots, x_p)$ es $N(\mu, \Sigma)$, donde

$$\Sigma = \mathrm{diag}\left(\Sigma_{1\,1}, \ldots, \Sigma_{p\,p}\right),$$

es decir, que los vectores aleatorios $x_1, \ldots, x_p$ son estocásticamente independientes, se decide calculando el supremo

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

siendo M_0 la clase de matrices con cajas en la diagonal y cero en el resto, es decir, tal que $\Sigma \in M_0$.

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. Como esta última está particularmente recomendada para variables contínuas, una ventaja de la distancia de Matusita es que puede ser aplicada a variables discretas (Dillon y Goldstein [36]), y a variables mixtas (Krzanowski [68]). De todos modos, sus aplicaciones se centran más bien en el área del Análisis Discriminante (véase Krzanowski [69]), como se va a comentar en la sección 1.6.3.

1.3.3 Distancia de Rao

Aunque introducida por Rao [88] hace bastante tiempo, ha sido estudiada más recientemente por Atkinson y Mitchell [6], Burbea y Rao [10], Oller y Cuadras [82],[83], Burbea y Oller [11], y otros.

Sea $S = \{p\,(x, \theta)\}$ un modelo estadístico, donde $p(x, \theta)$ está parametrizado por θ que pertenece a una variedad diferenciable Θ, dotada de una métrica en la que la matriz de información de Fisher

$$\mathcal{F}_\theta = E\left[\frac{\partial}{\partial \theta} \log p\,(x, \theta) \cdot \frac{\partial}{\partial \theta'} \log p\,(x, \theta)\right] \tag{1.28}$$

juega el papel de tensor métrico fundamental sobre Θ. La distancia de Rao es la distancia entre dos parámetros θ_A, θ_B de Θ. Se conoce la distancia de Rao para bastantes distribuciones (Cuadras [23]), aunque el caso normal multivariante ha sido sólo en parte resuelto (Calvo y Oller [13]).

Dadas dos distribuciones, F y G, pertenecientes a una misma familia paramétrica, la distancia de Rao puede ser utilizada para contrastar la hipótesis nula $H_0 : F = G$, transformándola en $H_0 : \delta(F, G) = 0$, de modo parecido al descrito en la sección anterior con la distancia de Matusita. Un resultado general dice que

$$V = \frac{n_1\, n_2}{n_1 + n_2}\, \widehat{\delta}^2(F, G)$$

sigue asintóticamente una ${\chi^2}_p$, siendo p el número de variables y $\widehat{\delta}^2$ una estimación de δ^2, en el caso $F = G$.

Como ejemplo interesante de aplicación, consideremos el modelo lineal normal $Y \sim N(X \cdot \beta, \sigma^2 I_n)$. Entonces $\theta = (\beta, \sigma) \in \mathbf{R}^m \times \mathbf{R}_+$. Consideremos la hipótesis nula

$$H_0 : \quad H \cdot \beta = 0$$

donde $\mathcal{F}(H) \subset \mathcal{F}(X)$. Sea $\widehat{\gamma} = (\widehat{\beta}, \widehat{\sigma})$ la estimación ML de (β, σ), y consideremos la subvariedad

$$\Theta_H = \{\gamma = (\beta, \sigma) : H\,\beta = 0\}$$

La distancia de Rao entre $\widehat{\gamma}$ y la subvariedad Θ_H es

$$R_H(\widehat{\gamma}) = \inf\{R(\widehat{\gamma}, \gamma) : \gamma \in \Theta_H\}$$

Una región crítica para decidir sobre $H_0 : H\,\beta = 0$ es de la forma

$$W = \{x \in \mathbf{R}^n : R_H(\widehat{\gamma}) > \delta_\epsilon\}$$

y puede probarse que es equivalente al clásico test F.

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

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

donde F es una función no negativa sobre $\mathbf{R}_+$ satisfaciendo la condición de normalización, Σ_0 y X son matrices fijas, se debe a Burbea y Oller [11]. Véase también Oller [84].

Aunque este planteamiento y el de Matusita son muy parecidos, conviene observar que para dos funciones de distribución F, G se cumple que

$$M(F, G) \leq R(F, G)$$

pues la distancia de Rao está definida en una subvariedad Θ de la variedad (diferenciable de dimensión infinita)

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

es decir, la esfera unidad (o el espacio proyectivo) del espacio L^2, con la estructura diferenciable inducida por la estructura natural de espacio de Hilbert con producto

$$\langle f, g\rangle = \int f\,g\,dx$$

Así, la distancia de Rao, que aprovecha el conocimiento de una parametrización y el cambio de información al variar los parámetros, tiene mayor poder de separación que la distancia de Matusita, que puede interpretarse como la distancia en línea recta entre las dos densidades en L^2, y es, por tanto, más apropiada en un contexto no paramétrico. No obstante, justo es añadir que las distancias de Matusita, Rao, y otras medidas de divergencia, coinciden localmente (Burbea y Rao [10]).

1.4 Representación de conjuntos

La representación de un conjunto finito 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.

Sea I un conjunto finito con n elementos que, por economía de notación, indicaremos

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

Presentamos en esta sección las formas de representación más usuales del conjunto I, 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 enfasis en el punto (1), puesto que proporciona una forma general de predicción, como estudiaremos a lo largo de esta memoria.

Definición 1.4.1 *Una* matriz de disimilaridades $\boldsymbol{\Delta} = (\boldsymbol{\delta}_{ij})$ *es una matriz real simétrica* $n \times n$ *cuyos elementos* $\boldsymbol{\delta}_{ij}$ *satisfacen*

$$\boldsymbol{\delta}_{ij} = \boldsymbol{\delta}_{ji} \geq \boldsymbol{\delta}_{ii} = 0 \qquad \forall\, i,\, j \in I$$

Se conocen muchos métodos para construir disimilaridades $\boldsymbol{\delta}_{ij}$ sobre I. Sin embargo, nos centraremos más en las propiedades de $\boldsymbol{\delta}_{ij}$ y en el tipo de representación de I que permiten.

Definición 1.4.2 *La matriz* $\boldsymbol{\Delta}$ *es llamada* Euclídea *si existe una configuración de puntos en un espacio euclídeo* $\mathbf{R}^p$ *cuyas interdistancias coincidan con las contenidas en* $\boldsymbol{\Delta}$*, es decir, si existen* $x_1, \ldots, x_n \in \mathbf{R}^p$ *tales que*

$$\boldsymbol{\delta}_{ij}{}^2 = (x_i - x_j)' \cdot (x_i - x_j) \qquad \forall\, i,\, j \in I$$

Definición 1.4.3 *La matriz de disimilaridades* $\boldsymbol{\Delta}$ *es llamada* ultramétrica *si, para todas las ternas* i, j, $k \in I$ *se verifica que*

$$\boldsymbol{\delta}_{ij} \leq \max\{\boldsymbol{\delta}_{ik}, \boldsymbol{\delta}_{jk}\}$$

Definición 1.4.4 *La matriz de disimilaridades* $\boldsymbol{\Delta}$ *es llamada* cuadripolar *si para todas las cuaternas* $i, j, k, l \in I$ *se verifica que*

$$\boldsymbol{\delta}^{+}{}_{ij} \leq \max\{\boldsymbol{\delta}^{+}{}_{ik}, \boldsymbol{\delta}^{+}{}_{jk}\}$$

siendo

$$\begin{aligned} \boldsymbol{\delta}^{+}{}_{ij} &= \boldsymbol{\delta}_{ij} + \boldsymbol{\delta}_{kl} \\ \boldsymbol{\delta}^{+}{}_{ik} &= \boldsymbol{\delta}_{ik} + \boldsymbol{\delta}_{jl} \\ \boldsymbol{\delta}^{+}{}_{jk} &= \boldsymbol{\delta}_{jk} + \boldsymbol{\delta}_{il} \end{aligned}$$

Definición 1.4.5 *La matriz de disimilaridades* $\boldsymbol{\Delta}$ *es llamada* de Robinson *si, para todas las ternas* $i, j, k \in I$ *con* $i \leq j \leq k$ *se verifica que*

$$\max\{\boldsymbol{\delta}_{ij}, \boldsymbol{\delta}_{jk}\} \leq \boldsymbol{\delta}_{ik}$$

1.4.1 Representación Euclídea

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

La matriz $\boldsymbol{\Delta}$ euclídea permite una proyección de I en $\mathbf{R}^p$

$$\begin{aligned} I &\longrightarrow \mathbf{R}^p \\ i &\longrightarrow x_i \end{aligned}$$

de modo que cada elemento i está representado por un punto x_i.

Cuando tal representación se hace mediante la técnica del análisis de Coordenadas Principales, la proyección es óptima en dimensión reducida. Las aplicaciones de esta técnica son numerosísimas, y ya son consideradas clásicas en el contexto del Análisis Multivariante. Véase por ejemplo Mardia et al. [78], Seber [96].

1.4.2 Representación Ultramétrica

El concepto de matriz $\boldsymbol{\Delta}$ ultramétrica está ligado a la necesidad de obtener clasificaciones jerárquicas objetivas, especialmente a partir de los trabajos de Benzécri [8], Hartigan [55], Jardine et al. [61] [62], y Johnson [63].

La importancia de $\boldsymbol{\Delta}$ ultramétrica reside en:

a) $\boldsymbol{\Delta}$ define sobre I una jerarquía indexada (C, α), es decir, $C \subset \mathcal{P}(I)$, de modo que

1. $I \in C$
2. $\{i\} \in C \quad \forall i \in I$

3. $\forall\, c_1, c_2 \in C$, la intersección $c_1 \cap c_2$ es $\emptyset$, o bien uno de los dos conjuntos c_1, c_2 está contenido en el otro.

4. Todo $c \in C$ es igual a la reunión de los elementos de C que contiene, o bien no contiene ningún otro elemento de C.

5. Existe una aplicación no negativa $\alpha : C \longrightarrow \mathbf{R}$ tal que $\alpha(\{i\}) = 0$, y $\alpha(c) < \alpha(c')$ si $c \subset c'$. Esta aplicación α recibe el nombre de *índice de la jerarquía*.

Así, C es una clase de conjuntos no solapantes, con un orden compatible con el índice α.

b) Para todo $r \in \mathbf{R}_+$, la relación binaria

$$i \sim_r j \iff \boldsymbol{\delta}_{i\,j} \leq r$$

es de equivalencia si, y sólo si $\boldsymbol{\delta}_{i\,j}$ es ultramétrica.

Recíprocamente, una jerarquía indexada (C, α) sobre I define una matriz ultramétrica $\boldsymbol{\Delta}$, pues basta definir

$$\boldsymbol{\delta}_{i\,j} = \alpha(c_{i\,j})$$

donde $c_{i\,j}$ es la mínima clase de C que contiene $\{i\}$ y $\{j\}$.

Tales propiedades confieren a las ultramétricas un papel fundamental en el estudio teórico de las clasificaciones, tal como fuera iniciado por C. Linneo en su famoso *Sistema Natural* y continuado, bajo la perspectiva matemática, por Jardine, Sibson, Sokal, Rohlf, Sneath y otros, creadores de la llamada Taxonomía Numérica de las especies vegetales y animales.

La representación geométrica de I se realiza mediante un grafo llamado dendrograma, y es a través del dendrograma que el taxonomista construye la jerarquía indexada.

Por ejemplo, la matriz

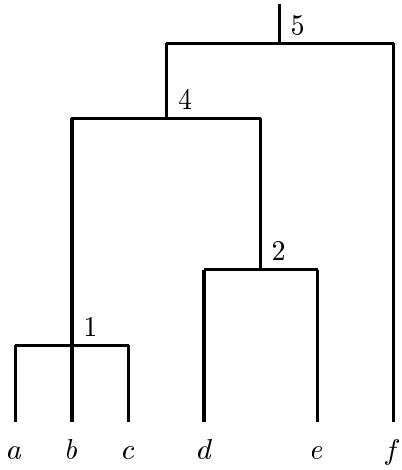
$$\Delta_u = \begin{pmatrix} 0 & 1 & 1 & 4 & 4 & 5 \\ & 0 & 1 & 4 & 4 & 5 \\ & & 0 & 4 & 4 & 6 \\ & & & 0 & 2 & 5 \\ & & & & 0 & 5 \\ & & & & & 0 \end{pmatrix}$$

sobre el conjunto $I = \{\, a,\, b,\, c,\, d,\, e,\, f \,\}$ es ultramétrica. I puede representarse mediante el dendrograma de la figura 1.1, que visualiza fácilmente la jerarquía de conjuntos

$$C = \{\{a\}_0,\, \ldots,\, \{f\}_0,\, \{a,\, b,\, c\,\}_1,\, \{d,\, e\,\}_2,\, \{a,\, b,\, c,\, d,\, e\,\}_4,\, I_5\}$$

donde se ha indicado el índice de la jerarquía como subíndice.

Figure 1.1: Dendrograma representando la matriz ultramétrica Δ_u



1.4.3 Representación Cuadripolar

Si la motivación de las matrices 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).

Consideremos un grafo conexo sin ciclos cuyos ejes tienen longitudes no negativas y cuyos extremos son los elementos de I. Las longitudes de los caminos que unen los extremos generan una matriz de distancias de tipo cuadripolar. Este tipo de grafo, junto con la métrica considerada, recibe el nombre de *árbol aditivo*. También se dice que $\boldsymbol{\delta}_{i\,j}$ es una distancia aditiva o que cumple el axioma de los cuatro puntos (Definición 1.4.4).

Son propiedades fundamentales de los árboles aditivos y las matrices cuadripolares:

a) Si $\boldsymbol{\Delta}$ es cuadripolar, entonces I se puede representar mediante un único árbol aditivo y reciprocamente (Buneman, [9]).

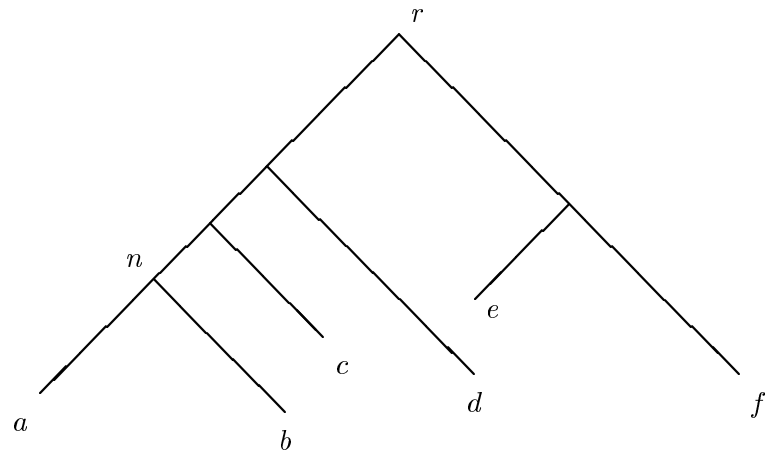
b) Si $\boldsymbol{\Delta} = (\boldsymbol{\delta}_{i\,j})$ es cuadripolar, existe entonces una matriz ultramétrica $U = (u_{i\,j})$, y una aplicación $\psi : I \longrightarrow \mathbf{R}$ tal que

$$\boldsymbol{\delta}_{i\,j} = u_{i\,j} + \psi(i) + \psi(j)$$

(Sattah y Tversky [95]).

Un árbol aditivo permite una fácil visualización de I, que mantiene diferencias esenciales con un dendrograma, pues un árbol aditivo no define una

Figure 1.2: Arbol aditivo representando la matriz cuadripolar Δ_c



jerarquía indexada ni sus extremos equidistan de un punto raiz (figura 1.2). Cuando un árbol aditivo representa un árbol evolutivo, los vértices o nodos representan elementos no pertenecientes a I, de modo que pueden ser interpretados como predecesores. La siguiente matriz sobre $I = \{a, b, c, d, e, f\}$

$$\Delta_c = \begin{pmatrix} 0 & 3 & 4 & 6 & 7 & 8 \\ & 0 & 3 & 7 & 8 & 9 \\ & & 0 & 6 & 7 & 8 \\ & & & 0 & 7 & 8 \\ & & & & 0 & 3 \\ & & & & & 0 \end{pmatrix}$$

es cuadripolar, y el árbol aditivo que representa I viene en la figura 1.2. Si se tratara de una árbol evolutivo, entonces las especies a y b tendrían un ancestro común representado por el nodo n.

1.4.4 Representación de Robinson

La motivación de las matrices de Robinson proviene de la necesidad de ordenar cronológicamente los elementos de un conjunto I. Los términos de las filas o columnas de $\boldsymbol{\Delta}$ no decrecen cuando nos apartamos de la diagonal principal a lo largo de cualquier fila o columna. Tales matrices surgen tras una adecuada ordenación de $\boldsymbol{\Delta}$, y reflejan esencialmente una estructura unidimensional de los datos.

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, la estructura unidimensional está dominada por el tiempo. Este problema equivale a ordenar I para obtener una matriz Δ de tipo cuadripolar. Éste fue el planteamiento original de Robinson [93].

Pero las matrices de Robinson juegan también un papel fundamental en el estudio de las *pirámides* introducidas por Diday [34], y Fichet [41], que son una generalización de las jerarquías indexadas,

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

1. $\{i\} \in P, \qquad \forall\, i \in I$
2. $I \in P$
3. La intersección de cualquier par $p,\, p' \in P$ puede ser $\emptyset$, o bien $p \cap p' \in P$
4. Existe un orden que es compatible con P

La última propiedad significa que si, por ejemplo, adoptamos el orden natural $I = \{1,\, 2,\, \ldots,\, n\}$ en I, y $p = \{i_1,\, \ldots,\, i_k\} \in P$, entonces $i_1 < \cdots < i_k$.

Una *pirámide indexada* (P, α) es una pirámide con un índice α tal que $\alpha(\{i\}) = 0$, para todos los $i \in P$, i $\alpha(p) \leq \alpha(p')$ si $p \subset p'$.

Se dice que 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 $p = p_1 \cap p_1$.

Una matriz de Robinson $\boldsymbol{\Delta} = (\boldsymbol{\delta}_{i\,j})$ recibe el nombre de *Robinson fuerte* si para todas las cuaternas ordenadas $i \leq j \leq k \leq l \in I$ se verifica que

$$\begin{array}{rcrcrcll} \boldsymbol{\delta}_{i\,j} & = & \boldsymbol{\delta}_{i\,k} & \Longrightarrow & \boldsymbol{\delta}_{h\,j} & = & \boldsymbol{\delta}_{h\,k} & \text{si} \quad h \leq i \\ \boldsymbol{\delta}_{j\,l} & = & \boldsymbol{\delta}_{k\,l} & \Longrightarrow & \boldsymbol{\delta}_{j\,m} & = & \boldsymbol{\delta}_{k\,m} & \text{si} \quad m \geq l \end{array}$$

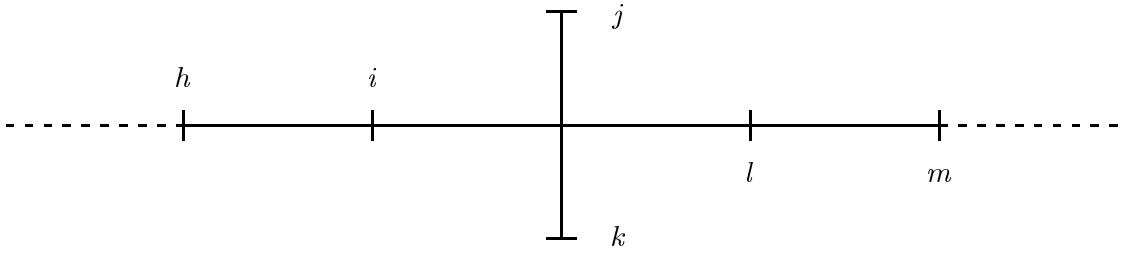
Las propiedades que relacionan las matrices de Robinson, las cuadripolares y las pirámides son las siguientes:

1. Si $\boldsymbol{\Delta}$ es de Robinson (salvo permutaciones), entonces I se puede representar mediante una pirámide indexada en sentido amplio y recíprocamente (Diday [35]).

2. Si $\mathbf{\Delta}$ es de Robinson y cuadripolar, entonces es Robinson fuerte (Critchley[18]).

El resultado (1) generaliza la biyección entre ultramétricas y dendrogramas. El resultado (2) caracteriza las matrices de Robinson que pueden ser representadas mediante un árbol aditivo. Cuando una matriz de Robinsos es cuadripolar, si i equidista de j y k, entonces todos los predecesores de i equidistan de j y k.

Figure 1.3: Ordenación cronológica definida por una matriz Robinson fuerte



La figura 1.3 visualiza esta propiedad, donde la ordenación

$$h < i < j = k < l < m$$

significa que j y k aparecen simultáneamente en el tiempo.

La siguiente matriz sobre $I = \{a, b, c, d\}$

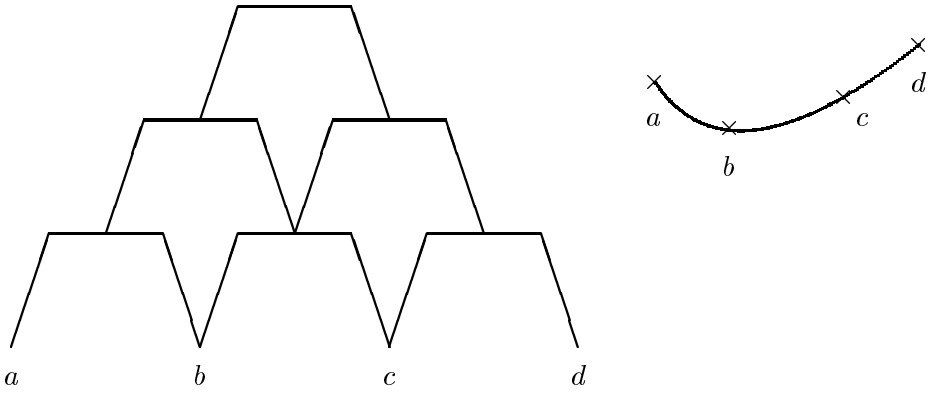
$$\mathbf{\Delta}_R = \begin{pmatrix} 0 & 1 & 2 & 3 \\ & 0 & 1 & 2 \\ & & 0 & 1 \\ & & & 0 \end{pmatrix}$$

es de Robinson, y define la pirámide

$$P = \{\{a\},\ \{b\},\ \{c\},\ \{d\},\ \{a,\ b\},\ \{b,\ c\},\ \{c,\ d\},\ \{a,\ b,\ c\},\ \{b,\ c,\ d\},\ I\}$$

y su representación viene dada en la figura 1.4. Obsérvese que $\mathbf{\Delta}_R$ no es ultramétrica.

Figure 1.4: Representación piramidal de cuatro objetos correspondiente a la matriz de Robinson Δ_R. A la derecha aparece una posible ordenación cronológica



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

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

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

1.5 Un Teorema fundamental y primeras consecuencias

Sea $\boldsymbol{\Delta} = (\boldsymbol{\delta}_{ij})$ una matriz $n \times n$ de disimilaridades sobre un conjunto finito U, con n elementos. Consideremos la matriz

$$A = (a_{ij}) \qquad a_{ij} = -\frac{1}{2}\,\boldsymbol{\delta}_{ij}{}^2$$

y la matriz

$$B = H \cdot A \cdot H \tag{1.29}$$

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.

El siguiente teorema es fundamental para todo lo que sigue

TEOREMA 1.5.1 *Sea $\boldsymbol{\Delta}$ una matriz $n \times n$ de disimilaridades sobre U, y B definida como en (1.29). $\boldsymbol{\Delta}$ es euclídea si, y sólo si B es semidefinida positiva. Entonces U puede ser representado por $x_1, \ldots, x_n \in \mathbf{R}^p$, siendo $p = rang\,(B)$, de modo que*

$$\boldsymbol{\delta}_{ij}{}^2 = \|x_i - x_j\|^2 \qquad \forall\, i,\, j \in U$$

(indicando por $\|\cdot\|$ la norma euclídea usual).

La demostración puede encontrarse en Mardia et al. [78], Cuadras [20], Seber [96]. Será muy importante tomar como solución la habitual del Análisis de Coordenadas Principales (Torgerson [99], Gower [49]). Consideremos la descomposición espectral de B

$$B = V \cdot \Lambda \cdot V' = X \cdot X'$$

donde X es la matriz $n \times p$ consistente en las p columnas no nulas de $V \cdot \Lambda^{1/2}$. Las filas de X constituyen la configuración euclídea deseada.

Se verifican las siguientes propiedades:

a) Las columnas de son los vectores propios de B, así que podemos escribir la configuración:

		λ_1	λ_2	$\cdots$	λ_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

$$(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0)$$

El elemento i–ésimo del conjunto U viene representado por el punto $x'_i = (x_{i\,1}, \ldots, x_{i\,p}) \in \mathbf{R}^p$.

b) Los *datos* de la matriz X son centrados, es decir, se anulan las medias de las columnas:

$$\begin{aligned} X' \cdot \mathbf{1}_n &= 0 \\ \overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{i\,j} &= 0 \end{aligned}$$

La *varianzas* de cada una de las columnas de X son proporcionales a los valores propios de B:

$$s^2{}_j = \frac{1}{n} \sum_{i=1}^{n} x_{i\,j}^2 = \frac{1}{n} \lambda_j$$

c) Las *variables* (columnas) de X son *incorrelacionadas*:

$$s_{j\,j'} = \frac{1}{n} \sum_{i=1}^{n} x_{i\,j} x_{i\,j'} = 0$$

d) Propiedad de optimalidad: Entre todas las proyecciones de U sobre $\mathbf{R}^k$, donde $k \leq p$, la proyección sobre las k primeras coordenadas principales es óptima en el sentido de que si $x_1(k), \ldots, x_n(k)$ representan tales coordenadas, y $y_1(k), \ldots, y_n(k)$ otras coordenadas, entonces

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

La proporción de la *variabilidad geométrica* de U que es explicada por estas k primeras coordenadas es

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

Tales propiedades nos dicen que la representación euclídea de U en dimensión reducida goza de excelentes propiedades: es centrada, la ortogonalidad de los ejes puede interpretarse como incorrelación, y la resolución en dimensión k es máxima. Haremos uso extensivo de estas propiedades, aunque la elección de las k coordenadas pueda cambiar en algunos problemas de predicción.

En el caso no euclídeo, el comportamiento de la matriz de distancias se refleja en el siguiente resultado:

TEOREMA 1.5.2 *Sea* $\boldsymbol{\Delta}$ *una matriz* $n \times n$ *de disimilaridades sobre el conjunto finito* U*, y* B *como en 1.5.1 Supongamos que* B *tiene* $p > 0$ *valores propios positivos y* $q > 0$ *valores propios negativos. Entonces existen* $z_1, \ldots, z_n \in \mathbf{R}^p \oplus \boldsymbol{i}\, \mathbf{R}^q$*, con* $\boldsymbol{i} = \sqrt{-1}$*, es decir,*

$$z_j = (x_j, \boldsymbol{i}\, y_j), \text{ con } x_j \in \mathbf{R}^p \text{ y } y_j \in \mathbf{R}^q \qquad (j = 1, \ldots, n)$$

verificando que

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

Véase una demostración en Cuadras [20].

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

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

siendo

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0 > \mu_1 \geq \mu_2 \geq \cdots \geq \mu_q$$

Obsérvese que $\mathbf{1}_n$ es vector propio de B de valor propio 0.

En este trabajo vamos a hacer una aplicación sistemática de los teoremas (1.5.1) y (1.5.2), pero no en el sentido de utilizarlos como técnica de representación de datos a lo largo de ejes principales, sino como método analítico de establecer propiedades de conjuntos y estructuras a través del estudio de las coordenadas principales.

La representación de conjuntos, sea a través de coordenadas principales, sea a través de dendrogramas, es y ha sido muy frecuente en las aplicaciones. Esta dualidad de representación impulsó a diversos especialistas a relacionarlas entre sí. Gower [51] conjeturó que toda distancia ultramétrica $u_{i\,j}$ 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 [58], y desde entonces se han obtenido diversos resultados en esta línea.

1.5.1 Representación euclídea de árboles ultramétricos

En este apartado se sintetizan algunas propiedades de la representación euclídea de disimilaridades ultramétricas. Sea $\boldsymbol{\Delta} = (\boldsymbol{\delta}_{i\,j})$ una matriz ultramétrica sobre un conjunto finito U de n elementos.

Proposición 1.5.1 *Supongamos que $\boldsymbol{\delta}_{i\,j} > 0$ para $i \neq j$. Entonces $\boldsymbol{\Delta}$ es euclídea $(n-1)$–dimensional.*

Véase Holman [58], Gower y Banfield [52], Cuadras y Carmona [21].

Proposición 1.5.2 *Sea $h_1 = \min\{\boldsymbol{\delta}_{i\,j} : \boldsymbol{\delta}_{i\,j} > 0\}$. Entonces el mínimo valor propio de la matriz B definida en el teorema (1.5.1) es $\lambda_1 = \frac{1}{2}h_1^2$.*

Véase Cuadras [20].

Proposición 1.5.3 *Existe una partición del conjunto U*

$$U = U_0 + U_1 + \cdots + U_r$$

tal que U_0 está formado por elementos aislados, y cada U_j, para $j = 1, \ldots, r$, es un cluster maximal de elementos equidistantes con distancia común h_j. Si μ_0 es el mayor valor propio de B, entonces

$$\mu_0 > \lambda_r^2 = h_r^2 \geq \cdots \geq \lambda_1^2 = \frac{1}{2}h_1^2$$

donde $\lambda_r \geq \cdots \geq \lambda_1$ son valores propios de B. Además, la matriz X descrita en el teorema (1.5.1) 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 Cuadras y Oller [22]

Proposición 1.5.4 *U puede representarse perfectamente en dimensión 1. La representación unidimensional significa que existe una transformación monótona $D = f(\boldsymbol{\Delta})$, es decir, $d_{i,\,j} = f(\boldsymbol{\delta}_{i\,j})$ verificando que*

$$d = At$$

donde

$$\begin{aligned} d &= (d_{1\,2}, d_{1\,3}, \ldots, d_{1\,n}, d_{2\,3}, \ldots, d_{n-1\,n})' \\ t &= (t_1, \ldots, t_{n-1})' \end{aligned}$$

con $t_i \geq 0$, *y siendo A la matriz cuyas primeras filas son:*

$$\begin{array}{ccccc} 1 & 0 & 0 & \ldots & \ldots \\ 1 & 1 & 0 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & 1 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{array}$$

Véase Critchley [17].

1.5.2 Otras representaciones euclídeas

Proposición 1.5.5 *Sea* $\boldsymbol{\Delta}$ *una matriz cuadripolar sobre U.*

Entonces $\boldsymbol{\Delta}^{(\alpha)} = (\boldsymbol{\delta}_{ij}{}^{\alpha})$ *es euclídea, siendo* $\alpha = (1/2)^k$, *para* $k = 1, 2, \ldots$, *en dimensión* $n-1$.

Proposición 1.5.6 *Sea* $\boldsymbol{\Delta}$ *una matriz de Robinson sobre U.*

Entonces $\boldsymbol{\Delta}^{(\alpha)} = (\boldsymbol{\delta}_{ij}{}^{\alpha})$ *es euclídea, siendo* $\alpha = (1/2)^k$, *para* $k \geq k_0$, *en dimensión* $n-1$.

El teorema de Holman (proposición 1.5.1) viene a decirnos 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 (1.5.3) sirve para clarificar la relación entre ambos tipos de representaciones. La proposición (1.5.4) afirma que una transformación monótona de $\boldsymbol{\Delta}$ permite una ordenación euclídea unidimensional algo especial, que puede ser utilizada como medio de representar el eje horizontal del dendrograma.

Sin embargo, las distancias cuadripolares y de Robinson no son euclídeas en general. Los resultados (1.5.5) y (1.5.6) ligan ambas con la propiedad euclídea, pero poco más se sabe al respecto.

La finalidad de este trabajo es aplicar la técnica del análisis de componentes principales al estudio de modelos de regresión y análisis discriminante, cuando tales modelos pueden ser estudiados a través de distancias. El estudio se llevará a cabo de forma similar al análisis de una matriz ultramétrica $\boldsymbol{\Delta}$ en el sentido de la proposición (1.5.3).

1.6 Predicción basada en distancias

Sea Y una variable dependiente, Ξ un conjunto de variables independientes, posiblemente de tipo mixto, es decir, conteniendo variables continuas, binarias, y cualitativas.

Supongamos que la observación de Y sobre un conjunto U de n indivíduos permite obtener una matriz de datos D a partir de la cual construimos una matriz $n \times n$ de distancias $\mathbf{\Delta}$. Sea X la matriz de coordenadas principales obtenida a partir de $\mathbf{\Delta}$ (teorema 1.5.1). Vamos a considerar tres tipos de problemas:

1. Predecir una variable contínua Y como una función de regresión de Ξ, siendo Ξ un conjunto mixto de variables.
2. Predecir Y como una función de regresión no lineal de Ξ, siendo Ξ un conjunto de variables contínuas.
3. Predecir Y, discreta con g estados, como un problema de clasificación siendo Ξ un conjunto mixto de variables.

El esquema de la predicción basada en distancias es:

$$\left.\begin{array}{llll} U & \xrightarrow{\Xi} & D & \longrightarrow & X \\ \boxed{\text{n+1}} & \xrightarrow{\Xi} & x & & \\ \boxed{\text{n+1}} & \xrightarrow{Y} & y & & \end{array}\right\} \quad y(n+1) = f(y, x, X)$$

siendo:

$\boxed{\text{n+1}}$	un nuevo individuo
x	las observaciones Ξ sobre $\boxed{\text{n+1}}$
y	las observaciones de Y sobre Y
X	las coordenadas principales sobre U
y(n+1)	la predicción de Y para $\boxed{\text{n+1}}$

La formulación general de este problema ha sido presentada por Cuadras en [24].

1.6.1 Predicción con variables mixtas

Sea y el vector $n \times 1$ el vector de observaciones de Y sobre U. Utilizamos el modelo de regresión

$$y = \mu\, \mathbf{1}_n + X_k \cdot \beta_k + e \tag{1.30}$$

donde X_k es una matriz $n \times k$ resultante de elegir $k \leq n-1$ coordenadas principales (columnas de X) según un criterio conveniente. β_k es un vector $k \times 1$ de parámetros. Este modelo ha sido estudiado por Cuadras y Arenas [25], probando que:

$$\begin{aligned} \widehat{\beta}_0 &= \overline{y} \\ \widehat{\beta}_k &= {\Lambda_k}^{-1} \cdot {X_k}' \cdot y \\ \widehat{y}\,(n+1) &= {x_k}' \cdot {\Lambda_k}^{-1} \cdot {X_k}' \cdot y \end{aligned}$$

Λ_k es la matriz diagonal $k \times k$ con los k valores propios de B (ver teorema (1.5.1)) que corresponden a los vectores seleccionados en X_k,

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

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

1.6.2 Regresión 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 $(\xi_{i\,1}, \ldots, \xi_{ip})$, $(\xi_{j\,1}, \ldots, \xi_{j\,p})$ observaciones sobre un par (i, j) de elementos de U. Cuadras [25] demuestra que adoptando la distancia $\boldsymbol{\delta}_{i\,j}$ definida por

$$\delta^2_{i\,j} = \sum_{h=1}^{p} |\xi_{i\,h} - \xi_{j\,h}|$$

y aplicando el modelo (1.30), se consigue una buena predicción de Y sin necesidad de conocer f.

En los capítulos 2 y 3 de esta memoria presentamos soluciones a algunos problemas algebraicos planteados por el estudio de esta distancia.

1.6.3 Análisis discriminante

Si Y tiene g estados, que podemos indicar $\pi_1, \ldots, \pi_g$, la predicción sobre un nuevo individuo, al que debemos asignar un valor de Y, es decir, uno de los g estados, es equivalente al problema del Análisis Discriminante:

Dada la partición de la población U en las g subpoblaciones definidas por los estados de Y

$$U = U_1 \cup U_2 \cup \ldots \cup U_g$$

donde U_k es el conjunto de los n_k individuos para los que la predicción cierta es π_k, clasificar el nuevo individuo $\boxed{\text{n+1}}$ en una de las g subpoblaciones.

Cuadras [24] estudia una regla de clasificación que parte de las g funciones discriminantes

$$f_k(\boxed{\text{n+1}}) = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{\delta}_i{}^2(k) - \frac{1}{{n_k}^2} \sum_{i=1}^{n_k} \sum_{j=i}^{n_k} \boldsymbol{\delta}_{ij}{}^2(k)$$

donde $\boldsymbol{\Delta}_k = (\boldsymbol{\delta}_{ij}(k))$ es la matriz de distancias de la subpoblación U_k, y $\boldsymbol{\delta}_i(k), \quad (i = 1, \ldots, n_k)$ las distancias de $\boxed{\text{n+1}}$ a los n_k individuos de esta subpoblación.

Este método de discriminación goza de buenas propiedades:

- Equivale al discriminador lineal clásico en el caso de que Ξ sea un conjunto de variables sólo contínuas, y tomando como δ_{ij} la distancia de Mahalanobis en U.
- 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.

El capítulo 5 de esta memoria es una contribución al estudio de este método.

Chapter 2

Modelo de regresión basado en distancias

2.1 Definición del modelo

Sea Y una variable respuesta contínua y $\Xi_1, \Xi_2, \ldots, \Xi_p$ variables regresoras contínuas, discretas o mixtas. Supongamos inicialmente que las variables son contínuas, y sea

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

un conjunto de n individuos o unidades experimentales, sobre los cuales observamos las variables. El modelo de regresión múltiple clásico es (Seber [96]):

$$y_i = \mu + \xi_{i\,1}\beta_1 + \ldots + \xi_{i\,p}\beta_p + e_i \qquad (i = 1, \ldots, n) \tag{2.1}$$

que escribiremos en notación matricial:

$$\boldsymbol{y} = \mu\,\mathbf{1} + \boldsymbol{\xi} \cdot \beta + e \tag{2.2}$$

donde $\boldsymbol{y}$ es el vector $n \times 1$ de observaciones de la variable Y, $\boldsymbol{\xi}$ es la matriz $n \times p$ que contiene en la fila i el vector ξ_i (de dimensión $1 \times p$), con las observaciones de las variables Ξ correspondientes al individuo i. El escalar μ y el vector β de dimensión $p \times 1$, son los parámetros del modelo, y el vector e, de dimensión $n \times 1$, contiene los errores aleatorios.

Sin embargo, el modelo 2.2 puede ser cuestionable en los siguientes casos:

1. Las variables regresoras son mixtas
2. La relación entre Y y $\Xi_1, \ldots, \Xi_p$ es no lineal

Poco se conoce acerca del planteamiento más adecuado en el caso (1). A menudo se resuelve cuantificando las variables cualitativas, pero el modelo depende de la cuantificación elegida.

El caso (2) exige conocer la función no lineal en el modelo de regresión, es decir:

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

A veces, tal conocimiento proviene de la naturaleza física del problema, pero otras veces la función f se desconoce. Si $p = 1$, lo más cómodo consiste en utilizar un modelo polinómico, pero si $p > 1$ este enfoque es más complicado.

Un nuevo enfoque al problema de definir un modelo de regresión y predecir los valores de Y sobre nuevos individuos, ha sido propuesto por Cuadras [24] y estudiado por Cuadras y Arenas [25].

Las observaciones sobre $U = \{1, 2, \ldots, n\}$ de las variables regresoras permiten, mediante una elección adecuada de una función de distancia, encontrar una matriz de distancias $\boldsymbol{\Delta} = (\boldsymbol{\delta}_{i\,j})$ de orden $n \times n$. Supongamos que la distancia es euclídea. Entonces, aplicando el teorema (1.5.1), existe

una matriz X de dimensión $n \times p$ tal que $B = X \cdot X'$, siendo $p = \text{rang}\,(B)$. Elegimos la llamada solución de coordenadas principales, por lo que podemos contar con las propiedades descritas en la sección 5 del capítulo 1.

Introducimos entonces el p–modelo, es decir, un modelo total de dimensión máxima p como sigue:

$$\boldsymbol{Y} = \mu\mathbf{1} + X \cdot \beta + e \tag{2.3}$$

Obsérvese que ahora X no contiene *observaciones sobre variables regresoras*, sino las coordenadas principales obtenidas de $\boldsymbol{\Delta}$.

2.1.1 Propiedades generales del modelo global

Las estimaciones LS (mínimos cuadrados) de μ y β vienen dadas por

$$\widehat{\mu} = \overline{y} \qquad \widehat{\beta} = \Lambda^{-1} \cdot X' \cdot \boldsymbol{y} \tag{2.4}$$

siendo $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_p)$ la matriz diagonal de valores propios de B.

El coeficiente de determinación, es decir, el cuadrado del coeficiente de correlación múltiple entre Y y X viene dado por

$$R^2 = \frac{\boldsymbol{y}' \cdot X' \cdot \Lambda^{-1} \cdot X' \cdot \boldsymbol{y}}{\sum (y_i - \overline{y})^2} \tag{2.5}$$

Supongamos ahora que $\boxed{\text{n+1}}$ es un nuevo individuo del que conocemos las observaciones sobre las variables independientes Ξ. Tales observaciones permiten calcular las distancias entre $\boxed{\text{n+1}}$ y cada uno de los individuos de I:

$$\boldsymbol{\delta}_{n+1,\,i} = \boldsymbol{\delta}(\boxed{\text{n+1}}, \boxed{\text{i}}) \qquad i \in I$$

A partir de ellas podemos hacer una predicción empleando el siguiente resultado (Gower [50]), que relaciona el vector $d = (\boldsymbol{\delta}^2_{n+1,\,1}, \ldots, \boldsymbol{\delta}^2_{n+1,\,n})'$ de los cuadrados de estas distancias con el vector $x_{n+1} = (x_{n+1,\,1}, \ldots, x_{n+1,\,p})$ de las coordenadas principales atribuibles al nuevo individuo.

Proposición 2.1.1

$$x'_{n+1} = \frac{1}{2}\Lambda^{-1} \cdot X' \cdot (b - d) \tag{2.6}$$

siendo b el vector de dimensión $n \times 1$ que contiene los elementos diagonales de la matriz B.

Demostración:

$$\begin{aligned} \boldsymbol{\delta}^2_{n+1,i} &= (x_{n+1} - x_i) \cdot (x_{n+1} - x_i)' \\ &= x_{n+1} \cdot x'_{n+1} + x_i \cdot x'_i - 2x_{n+1} \cdot x'_i \end{aligned} \tag{2.7}$$

Sumando para i de 1 a n, y teniendo en cuenta que las columnas de X suman 0

$$\sum_{i=1}^{n} \boldsymbol{\delta}^2_{n+1,i} = n x_{n+1} \cdot x'_{n+1} + \mathrm{tr}B$$

Sustituyendo en (2.7)

$$2x_i \cdot x'_{n+1} = \frac{1}{n} \left(\sum_{i=1}^{n} \boldsymbol{\delta}^2_{n+1,i} - \mathrm{tr}B \right) + b_{i,i} - \boldsymbol{\delta}^2_{n+1,i}$$

Superponiendo estas n ecuaciones en forma matricial

$$2X \cdot x'_{n+1} = \frac{1}{n} \left(\sum_{i=1}^{n} \boldsymbol{\delta}^2_{n+1,i} - \mathrm{tr}B \right) \mathbf{1}_n + (b - d)$$

Finalmente, el resultado se obtiene multiplicando a la izquierda por X', dado que $X' \cdot \mathbf{1}_n = 0$. □

El modelo (2.3) depende de la distancia $\boldsymbol{\delta}_{i\,j}$ elegida, contiene el modelo de regresión clásica como caso particular, pero es especialmente interesante en los casos mixto y no lineal, con tal de elegir una distancia adecuada.

2.1.2 Regresión lineal clásica

Si las variables $\Xi = (\Xi_1, \ldots, \Xi_p)$ son continuas, el modelo de regresión lineal clásico es

$$\boldsymbol{y} = \mu \mathbf{1} + \boldsymbol{\xi} \cdot \gamma + e \tag{2.8}$$

donde $\boldsymbol{\xi}$ es la matriz de dimensión $n \times p$ formada por los n vectores $\xi_1, \ldots, \xi_n$ (de dimensión $1 \times p$) de observaciones de las variables Ξ correspondientes a los n individuos,

Elijamos ahora la distancia euclídea al cuadrado

$$\boldsymbol{\delta}^2_{i\,j} = (\xi_i - \xi_j)' \cdot (\xi_i - \xi_j) = \xi'_i \xi_i + \xi'_j \xi_j - 2\xi'_i \xi_j$$

La matriz $\boldsymbol{\Delta}^{(2)} = (\boldsymbol{\delta}^2_{i\,j})$ verifica

$$\boldsymbol{\Delta}^{(2)} = S_F + S_C - 2\boldsymbol{\xi} \cdot \boldsymbol{\xi}'$$

donde S_F tiene las filas iguales y S_C tiene las columnas iguales. Como $HS_F = S_C H = 0$, la matriz B (Teorema 5.1) es

$$B = (H\boldsymbol{\xi}) \cdot (\boldsymbol{\xi}' H) = X \cdot X'$$

Asípues, al aplicar el modelo de regresión basado en distancias a este caso, obtenemos

$$\boldsymbol{y} = \mu \mathbf{1} + X \cdot \gamma + e \tag{2.9}$$

que es *equivalente* al anterior. En efecto, 2.9 es el modelo 2.8 expresado en la forma centrada y ortogonal, es decir, donde las variables se han centrado en la media y se ha llevado a cabo una transformación lineal que las convierte en ortogonales.

Sea ahora ξ el vector de observaciones sobre un nuevo individuo. Entonces la distancia euclídea al cuadrado del individuo $\boxed{\text{n+1}}$ al individuo $\boxed{\text{i}}$ del conjunto de referencia I es:

$$\boldsymbol{\delta}_i^2 = \|\xi - \xi_i\|^2 = \|x - x_i\|^2 = x' \cdot x + b_{i\,i} - 2x' \cdot x_i$$

Siendo x el vector de observaciones de $\boxed{\text{n+1}}$ respecto a 2.9. La ecuación anterior nos permite escribir

$$(b - d)' = 2 \cdot x' \cdot X' - x' \cdot x \cdot \mathbf{1}'$$

con lo cual, la predicción para $\boxed{\text{n+1}}$ es

$$\widehat{y}^{n+1} = \overline{y} + x' \cdot X' \cdot X \cdot \Lambda^{-2} \cdot X' \cdot y = \overline{y} + x' \cdot \Lambda^{-1} \cdot X' \cdot y \tag{2.10}$$

que es la misma formula que obtendríamos utilizando regresión clásica, bajo el modelo ortogonal centrado.

2.1.3 Regresión con variables cualitativas

Supongamos que $\Xi = (\Xi_1, \ldots, \Xi_p)$ son variables cualitativas, y que Ξ_r posee q_r estados excluyentes. Una medida bastante utilizada de similaridad entre dos individuos i, j es $m_{i\,j}$, el número de estados presentes simultáneamente en i y en j. Como $m_{i\,j} \leq p$, una medida de distancia viene dada por

$$\boldsymbol{\delta}_i^2 = 2\,(p - m_{i\,j})$$

Por otra parte, codificando los estados de Ξ_r como q_r variables binarias 0/1, es evidente que $\boldsymbol{\delta}_i^2$ es una distancia euclidea. Luego el modelo DB proporciona los mismos resultados que el modelo clásico si tratamos las p variables como $\sum_{r=1}^{p} q_r$ variables binarias 0/1.

2.1.4 Regresión con variables mixtas

Supongamos que el conjunto de variables regresoras consta de p_1 variables continuas, p_2 variables dicotómicas, y p_3 variables cualitativas. Se puede construir un modelo de regresión basado en la distancia $\boldsymbol{\delta}_{ij}{}^2 = 2\,(1 - s_{ij})$, siendo

$$s_{ij} = \frac{\displaystyle\sum_{k=1}^{p_1} \left(1 - \frac{|x_{ik} - x_{jk}|}{R_k}\right) + a + m_{ij}}{p_1 + (p_2 - d) + p_3} \tag{2.11}$$

donde m_{ij} es el número de estados presentes simultáneamente para las p_1 variables cualitativas, a y d representan el número de coincidencias 1/1 y 0/0, respectivamente, para las p_2 variables dicotómicas, y R_k es el rango de la k–ésima variable continua. La expresión 2.11 ha sido propuesta por Gower [51], y su utilización ha dado muy buenos resultados [25].

2.1.5 Regresión no lineal

El modelo de regresión no lineal

$$y_i = f(x_{i\,1}, \ldots, x_{ip}) + e_i \qquad i = 1, \ldots, n \tag{2.12}$$

ha sido poco estudiado porque se pierden muchas de las propiedades geométricas y estadísticas que permiten el elegante tratamiento del modelo lineal (ver secc 2.1, cap 1). Sin embargo, recientemente se ha prestado mayor atención al tema, (véanse, por ejemplo, los libros de Ratkowsky [91] y Seber and Wild [97]).

El modelo 2.3 basado en distancias permite también abordar con éxito algunos problemas que en principio requerirían el modelo (2.12). El método consiste en utilizar como distancia al cuadrado entre dos observaciones de las variables regresoras la distancia *valor absoluto* definida como

$$\boldsymbol{\delta}_{ij}^2 = |x_{i\,1} - x_{j\,1}| + \cdots + |x_{ip} - x_{jp}| \tag{2.13}$$

Según veremos en el teorema (2.2.1), $\boldsymbol{\Delta} = (\boldsymbol{\delta}_{ij})$ es una matriz de distancias euclídeas (entendiendo este concepto como existencia de una configuración de puntos en un espacio euclídeo cuya matriz de interdistancias es $\boldsymbol{\Delta}$), aunque la función $\boldsymbol{\delta}_{ij}{}^2$ de las variables x no es una forma cuadrática.

Este modelo de regresión tiene tres importantes ventajas:

1. No es necesario conocer una función f.
2. Es posible calcular un coeficiente de determinación R^2 que mida el grado de ajuste al modelo.

3. No presenta los delicados problemas numéricos de convergencia, propios de los ajustes a modelos no lineales.

Ejemplo 2.1.1 *Afifi y Azen [1] (pp. 187–188) consideran el siguiente modelo de regresión:*

$$H = \alpha(1 - \exp(-\beta t)) \tag{2.14}$$

que relaciona la concentración de hormona H con el tiempo t. La estimación LS de los parámetros α y β del modelo 2.14, con los datos de la tabla 1, es:

$$\widehat{\alpha} = 0.1758 \qquad \widehat{\beta} = 0.1531$$

Para estos mismos datos, utilizando el modelo DB, se obtiene un coeficiente de determinación $R^2 = 0.9971$. La siguiente tabla detalla las predicciones $\widehat{H}$ de los valores de H correspondientes a los valores observados empleando los dos modelos.

t	H	$\widehat{H}$, según el modelo 2.14	$\widehat{H}$, según el modelo DB
0.0	0.000	0.000	0.004
1.0	0.025	0.025	0.022
1.5	0.035	0.036	0.033
2.0	0.045	0.046	0.044
2.5	0.055	0.056	0.055
3.0	0.065	0.065	0.065
3.5	0.075	0.073	0.074
4.0	0.082	0.081	0.083
4.5	0.088	0.088	0.090
5.0	0.094	0.094	0.095
5.5	0.100	0.100	0.100
6.0	0.105	0.106	0.103
6.5	0.110	0.111	0.107
7.0	0.115	0.116	0.115
7.5	0.120	0.120	0.121
8.0	0.125	0.124	0.126

2.2 Estudio de la distancia Valor Absoluto

Supongamos que p variables continuas $X_1, \ldots, X_p$ permiten obtener las observaciones $x_i = (x_{i\,1}, \ldots, x_{1\,p})$ sobre cada individuo i del conjunto $U = \{0, \ldots, n\}$.

Como hemos visto anteriormente, la regresión no lineal a través del modelo DB se basa de un modo bastante eficiente en la distancia Valor Absoluto.

$$\boldsymbol{\delta}_{ij}^2 = |x_{i\,1} - x_{j\,1}| + \cdots + |x_{i\,p} - x_{j\,p}| \tag{2.15}$$

En las secciones siguientes proponemos un estudio de esta distancia que proporcione una justificación teórica de su buen comportamiento en la predicción según el modelo DB.

En primer lugar, se demuestra la siguiente propiedad básica

TEOREMA 2.2.1 *La matriz de distancias* $\boldsymbol{\Delta}^{(2)} = (\boldsymbol{\delta}_{ij}^2)$ *es euclídea.*

Demostración: La matriz $\boldsymbol{\Delta}^{(2)}$ de distancias al cuadrado es la suma de p matrices

$$\boldsymbol{\Delta}^{(2)} = \boldsymbol{\Delta}(1)^{(2)} + \cdots + \boldsymbol{\Delta}(p)^{(2)}$$

donde cada sumando contiene información de una de las variables

$$\boldsymbol{\Delta}(k)^{(2)} = (\boldsymbol{\delta}_{ij}^2(k)) = (|x_{i\,k} - x_{j\,k}|)$$

Veamos que cada una de las matrices $\boldsymbol{\Delta}(k)^{(2)}$ es euclídea: Podemos suponer, sin pérdida de generalidad, que para una variable k fija, tenemos las desigualdades:

$$x_{0\,k} \leq x_{1\,k} \leq \ldots \leq x_{n\,k}$$

(el caso general se obtiene reordenando los individuos).

En estas condiciones, se observa por cálculo directo que la matriz de distancias euclideas entre los puntos

$$\begin{array}{lccccccccc}
0: & (& 0 & , & 0 & , & 0 & , \ldots , & 0 &) \\
1: & (& \sqrt{x_1 - x_0} & , & 0 & , & 0 & , \ldots , & 0 &) \\
2: & (& \sqrt{x_1 - x_0} & , & \sqrt{x_2 - x_1} & , & 0 & , \ldots , & 0 &) \\
 & & \ldots & & \ldots & & \ldots & & \ldots & \\
i: & (& \sqrt{x_1 - x_0} & , & \ldots & , & \sqrt{x_i - x_{i-1}} & , \ldots , & 0 &) \\
 & & \ldots & & \ldots & & \ldots & & \ldots & \\
n: & (& \sqrt{x_1 - x_0} & , & \ldots & , & \sqrt{x_i - x_{i-1}} & , \ldots , & \sqrt{x_n - x_{n-1}} &)
\end{array} \tag{2.16}$$

coincide con la matriz $\boldsymbol{\Delta}(k)^{(2)}$. (Se ha omitido el subíndice común k en las x, por claridad de notación). Por tanto, la matriz

$$B(k) = H \cdot (-\frac{1}{2}\boldsymbol{\Delta}(k)^{(2)}) \cdot H$$

es definida positiva.

Finalmente,

$$B = H \cdot (-\frac{1}{2}\boldsymbol{\Delta}^{(2)}) \cdot H$$

es una matriz definida positiva, por ser suma de matrices de esta clase.□

2.2.1 El caso unidimensional equidistante

Las dificultades que presenta el estudio de la distancia 2.15 aconsejan abordar primero el caso de un conjunto unidimensional de puntos equidistantes, para los que tomaremos coordenadas con valores enteros.

Consideremos el conjunto $U = \{0, 1, \ldots, n\}$, y definamos sobre U la distancia

$$\boldsymbol{\delta}^2_{i\,j} = |i - j| \tag{2.17}$$

Como la obtención directa de las coordenadas principales a partir de la matriz de distancias $\boldsymbol{\Delta}^{(2)}$, por diagonalización de

$$B = H \cdot (-\frac{1}{2}\boldsymbol{\Delta}^{(2)}) \cdot H$$

donde H es la matriz de centrado de dimensión $(n + 1, n + 1)$, es algo complicada, las deduciremos indirectamente por transformación de una matriz de coordenadas euclídeas centradas X, cuya matriz de interdistancias (euclídeas) sea igual a $\boldsymbol{\Delta}^2$. Podremos encontrar tal configuración ya que la matriz de distancias asociada a U por la distancia valor absoluto es euclídea.

El procedimiento consiste en calcular las componentes principales (vectores propios de la matriz de covarianzas) de $X_1, \ldots, X_n$, y transformar X con la matriz ortogonal de vectores propios obtenida de este cálculo. Es decir, si C es la matriz de covarianzas de X, entonces las coordenadas principales vienen dadas por

$$Y = X \cdot V \tag{2.18}$$

siendo $C = V \cdot \Lambda \cdot V'$ la descomposición espectral de C, puesto que

$$\begin{aligned} Y \cdot Y' &= X \cdot X' = B \\ Y' \cdot Y &= V' \cdot X' \cdot X \cdot V = V' \cdot C \cdot V = \Lambda \end{aligned}$$

(véase Mardia [78, Teorema 14.3.1]).

Partimos de las coordenadas euclídeas proporcionadas por las variables convencionales $X_1, \ldots, X_n$ siguientes

	Variables				
Individuos	X_1	X_2	X_3	$\ldots$	X_n
0	$-n$	$-(n-1)$	$-(n-2)$	$\ldots$	-1
1	1	$-(n-1)$	$-(n-2)$	$\ldots$	-1
2	1	2	$-(n-3)$	$\ldots$	-1
3	1	2	3	$\ldots$	-1
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$n-1$	1	2	3	$\ldots$	-1
n	1	2	3	$\ldots$	n

(2.19)

Proposición 2.2.1 *Estas variables definen una representación euclídea del conjunto U con la distancia 2.17 (salvo un factor irrelevante de $n+1$)*

Demostración: Se ha obtenido la matriz (2.19) multiplicando a la izquierda

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 1 & 1 & 0 & & 0 & 0 \\ 1 & 1 & 1 & 1 & & 0 & 0 \\ \vdots & \vdots & & & \ddots & & \vdots \\ 1 & 1 & 1 & & & 1 & 0 \\ 1 & 1 & 1 & 1 & \ldots & 1 & 1 \end{pmatrix}$$

por la matriz de centrado de dimensión $(n+1, n+1)$

$$H = I_{n+1} - \frac{1}{n+1} \mathbf{1}_{n+1} \cdot \mathbf{1}'_{n+1}$$

(con un factor $(n+1)$ a fin de obtener coordenadas enteras, más manejables)

Se puede comprobar directamente que: las filas de A constituyen una configuración euclídea que corresponde a los puntos de U con la distancia valor absoluto.

Finalmente, el producto a la izquierda por H conserva las distancias euclídeas entre filas. □

Por construcción, las medias de las variables (columnas) de X son nulas: $\overline{X}_i = 0, \qquad i = 1, \ldots, n.$

La covarianza entre X_i y X_j es $c_{i\,j} = \frac{1}{n+1} X_i' \cdot X_j$. Para $i < j$, las coordenadas i y j son:

$$\begin{array}{ccc} & X_i & X_j \\ i & \left\{ \begin{array}{cc} -(n+1-i) & -(n+1-j) \\ \vdots & \vdots \\ & \\ \vdots & \vdots \\ -(n+1-i) & -(n+1-j) \end{array} \right. \\ j-i & \left\{ \begin{array}{cc} i & -(n+1-j) \\ \vdots & \vdots \\ & \\ \vdots & \vdots \\ i & -(n+1-j) \end{array} \right. \\ n-j & \left\{ \begin{array}{cc} i & j \\ \vdots & \vdots \\ & \\ \vdots & \vdots \\ i & j \end{array} \right. \end{array}$$

Luego:

$$\begin{aligned} X_i' \cdot X_j &= i\,(n+1-i)\,(n+1-j) - (j-i)\,i\,(n+1-j) + (n+1-j)\,i\,j \\ &= i\,(n+1)\,(n+1-j) \end{aligned}$$

Con lo cual:

$$\begin{aligned} c_{i\,j} &= i\,(n+1-j) \qquad i \leq j \\ c_{j\,i} &= c_{i\,j} \end{aligned} \tag{2.20}$$

o, equivalentemente

$$c_{i\,j} = (n+1)\,\min\{i,\,j\} - i\,j \qquad (1 \leq i,\,j \leq n)$$

Por ejemplo, para $n = 4$ es

$$C = \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 6 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

Observemos que la parte superior de C y la parte inferior son iguales salvo simetrías y permutaciones. En general, para un conjunto de $n+1$ individuos, la matriz C (de dimensión n) cumple la siguiente propiedad:

$$C_{i,\,j} = C_{n+1-j,\,n+1-i} \tag{2.21}$$

En la sección siguiente vamos a estudiar con más detalle las matrices que presentan esta estructura, y en particular, sus valores y vectores propios.

2.3 Estructura de las matrices centro–simétricas

2.3.1 Definición

Dada una matriz C de dimensión (n, n), designaremos por $C^\sharp$ la *anti–transpuesta* de C, es decir, el resultado de efectuar la operación de transposición respecto la antidiagonal

$$C^\sharp{}_{i,\,j} = C_{n+1-j,\,n+1-i}$$

Esta operación puede describirse mediante la acción sobre C de la matriz de permutación $W\,(n)$, definida como la matriz de dimensión (n, n), que tiene todos los elementos de la antidiagonal iguales a 1 y todos los demás nulos.

$$W_{i\,j} = \delta_{i,\,n+1-j}$$

$W(n)$ representa la permutación de longitud máxima *(permutación total)* sobre un conjunto de n elementos:

$$(1, \ldots, n) \mapsto (n, \ldots, 1)$$

W es idempotente, simétrica (y, evidentemente, $W^\sharp = W$). En lo sucesivo omitimos el orden n en la notación de la matriz W, siempre que resulte claro del contexto.

El resultado de la operación de W sobre una matriz C es

$$W \cdot C \cdot W = (W')^\sharp = \left(W^\sharp\right)'$$

que equivale a aplicar la permutación total a filas y columnas de C, o bien a transponer sucesivamente respecto las dos diagonales de C.

Definición 2.3.1 *Una matriz cuadrada C es* centro–simétrica *si $W \cdot C \cdot W = C$, o equivalentemente, si conmuta con W. Los elementos de una matriz centro–simétrica C de dimensión (n, n) verifican que $C_{i,\,j} = C_{n+1-i,\,n+1-j}$*

Este tipo de matrices aparecen también en el estudio de Ecuaciones Integrales en Física Matemática. Algunas propiedades elementales pueden encontrarse en Cord y Sylvester [14] y en Good [48].

Un caso particular notable de matrices centro–simétricas es el de las matrices de Toeplitz, que se definen como aquellas matrices en las que todos los elementos de las diagonales equidistantes de la diagonal principal son iguales.

Las matrices centro–simétricas tienen la siguiente estructura en cajas

Si n es par (sea $n = 2\,q$),

$$C = \begin{pmatrix} M & N \cdot W \\ W \cdot N & W \cdot M \cdot W \end{pmatrix} \tag{2.22}$$

Si n es impar (sea $n = 2\,q + 1$),

$$C = \begin{pmatrix} M & W \cdot u & N \cdot W \\ v' \cdot W & p & v' \\ W \cdot N & u & W \cdot M \cdot W \end{pmatrix} \tag{2.23}$$

donde M y N son matrices (q, q), u y v son vectores $(q, 1)$, p es un escalar, y las W que aparecen en las cajas representan la matriz $W(q)$.

En general, una matriz centro–simétrica no es simétrica. De las expresiones en cajas 2.22 y 2.23 se deduce que la condición de simetría para una matriz centro–simétrica C es

- Si $n = 2\,q$, $\quad C$ es simétrica $\iff$ M y N son simétricas
- Si $n = 2\,q + 1$, C es simétrica $\iff$ M y N son simétricas, y $u = v$

La matriz de covarianzas de la sección anterior, para un conjunto de $n + 1$ puntos equidistantes, es una matriz simétrica y centro–simétrica de dimensión (n, n). Sus cajas componentes son

- Si $n = 2\,q$,

$$\begin{aligned} M_{i\,j} &= (2\,q + 1)\,\min\{i,\, j\} - i\,j \\ N_{i\,j} &= i\,j \end{aligned} \tag{2.24}$$

- Si $n = 2\,q + 1$

$$\begin{aligned} M_{i\,j} &= (2\,q + 2)\,\min\{i,\, j\} - i\,j \\ N_{i\,j} &= i\,j \\ u_i &= (q+1)^2 - (q+1)\,i \\ p &= (q+1)^2 \end{aligned} \tag{2.25}$$

2.3.2 Valores y vectores propios

De la definición anterior se sigue que los subespacios propios de W son invariantes por la acción de de una matriz centro–simétrica C, propiedad que emplearemos en la descripción de los vectores propios de C

Definición 2.3.2 *Daremos el nombre de* espacio impar *al subespacio propio de W correspondiente al valor propio $+1$*

$$F = \{x \in \mathbf{R}^n | W \cdot x = x\}$$

y de espacio par *al subespacio propio de W correspondiente al valor propio -1*

$$G = \{x \in \mathbf{R}^n | W \cdot x = -x\}$$

El espacio $\mathbf{R}^n$ descompone en suma ortogonal de F y G

$$\dim F = \begin{cases} q & \text{si} \quad n = 2\,q \\ q+1 & \text{si} \quad n = 2\,q+1 \end{cases} \tag{2.26}$$

$$\dim G = q \qquad \text{(en ambos casos)} \tag{2.27}$$

Los vectores de F tienen la estructura en cajas

$$\begin{pmatrix} x \\ W \cdot x \end{pmatrix} \qquad \text{si} \quad n = 2\,q$$

$$\begin{pmatrix} x \\ y \\ W \cdot x \end{pmatrix} \qquad \text{si} \quad n = 2\,q+1$$

donde x es un vector columna de dimensión $(q, 1)$, y y es un escalar.

Los elementos de G tienen la estructura en cajas

$$\begin{pmatrix} x \\ -W \cdot x \end{pmatrix} \qquad \text{si} \quad n = 2\,q$$

$$\begin{pmatrix} x \\ 0 \\ -W \cdot x \end{pmatrix} \qquad \text{si} \quad n = 2\,q+1$$

Proposición 2.3.1

1) Si x es un vector propio de una matriz centrosimétrica C con valor propio λ, entonces $x = f + g$, siendo $f \in F$, $g \in G$ vectores propios de C con valor propio λ. En particular, si λ es un valor propio simple, entonces $x \in F$ o $x \in G$

2) Si C es positiva (en el sentido de tener todos sus elementos positivos), entonces el valor propio máximo es positivo y simple, y el vector propio correspondiente pertenece a F,

Demostración: Si $C \cdot x = \lambda\, x$, consideramos la descomposición $x = f + g$, siendo $f \in F$ y $g \in G$. Entonces $C \cdot x = C \cdot f + C \cdot g = \lambda\, f + \lambda\, g$, por tanto, $C \cdot f = \lambda\, f$, y $C \cdot g = \lambda\, g$.

La segunda parte resulta del teorema de Perron: el valor propio máximo es positivo y simple, y podemos tomar todas las componentes de su vector propio positivas, lo que es incompatible con la pertenencia a G. □

Teniendo en cuenta la primera parte de esta propiedad, en muchos casos de interés, y en concreto, en el caso de la matriz de covarianzas de los puntos equidistantes, que tiene n valores propios distintos, según veremos en el Capítulo 3, el cálculo de vectores propios de una matriz centro–simétrica se reduce a un problema análogo en dimensión mitad.

En efecto, si $n = 2\,q$, la ecuación de los vectores propios contenidos en F

$$\begin{aligned} C \cdot \begin{pmatrix} x \\ W \cdot x \end{pmatrix} &= \begin{pmatrix} M & N \cdot W \\ W \cdot N & W \cdot M \cdot W \end{pmatrix} \cdot \begin{pmatrix} x \\ W \cdot x \end{pmatrix} \\ &= \begin{pmatrix} (M+N) \cdot x \\ W \cdot (M+N) \cdot x \end{pmatrix} \\ &= \lambda \cdot \begin{pmatrix} x \\ W \cdot x \end{pmatrix} \end{aligned}$$

equivale al problema de vectores propios de la matriz $M + N$ de dimensión (q, q). Análogamente, la ecuación de los vectores propios contenidos en G equivale al problema de vectores propios de la matriz $M - N$ de dimensión (q, q).

Si $n = 2\,q + 1$, la ecuación para los vectores propios contenidos en G es también equivalente al problema de vectores propios de la matriz $M - N$ de dimensión (q, q), mientras que para los vectores propios contenidos en F, se obtiene

$$\begin{aligned} C \cdot \begin{pmatrix} x \\ y \\ W \cdot x \end{pmatrix} &= \begin{pmatrix} M & W \cdot u & N \cdot W \\ v' \cdot W & p & v' \\ W \cdot N & u & W \cdot M \cdot W \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ W \cdot x \end{pmatrix} \\ &= \begin{pmatrix} (M+N) \cdot x & + & W \cdot uy \\ 2\,v' \cdot W \cdot x & + & p\,y \\ W \cdot (M+N) \cdot x & + & u\,y \end{pmatrix} \\ &= \lambda \cdot \begin{pmatrix} x \\ y \\ W \cdot x \end{pmatrix} \end{aligned}$$

que equivale al problema de vectores propios para la matriz de dimensión $(q + 1, q + 1)$

$$\begin{pmatrix} M + N & W \cdot u \\ 2\,v' \cdot W & p \end{pmatrix}$$

2.3.3 Las matrices B y $\widetilde{B}$

La propiedad de centro–simetría de la matriz de covarianzas C de la sección anterior permite aplicar estas propiedades al estudio de sus valores y vectores

propios

La matriz $M + N$, que aparece en el cálculo de los vectores propios de C contenidos en F es, según (2.24) y (2.25), un múltiplo de la matriz

$$B(q) = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 2 & \dots & 2 \\ 1 & 2 & 3 & \dots & 3 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 2 & 3 & \dots & q \end{pmatrix}$$

de componentes

$$B(q)_{i\,j} = \min\{i,\, j\} \qquad 1 \leq i,\, j \leq q$$

$$M + N = \begin{cases} (2\,q + 1)\,B(q) & \text{si} \quad n = 2\,q \\ (2\,q + 2)\,B(q) & \text{si} \quad n = 2\,q + 1 \end{cases}$$

En el caso $n = 2\,q$, la matriz $M - N$ que aparece en el cálculo de los vectores propios de C contenidos en G es una matriz de dimensión $(q,\, q)$, que denotaremos $\widetilde{B}(q)$, con elementos

$$\widetilde{B}(q)_{i\,j} = (2\,q + 1)\ \min\{i,\, j\} - 2\,i\,j$$

y, con una notación más compacta

$$\widetilde{B}(q) = (2\,q + 1)\,B(q) - 2\,b(q) \cdot b(q)'$$

donde $b(q)$ es el vector de dimensión $(q,\, 1)$

$$b(q) = (1, 2, \dots, q)'$$

En el caso $n = 2\,q + 1$, la matriz $M - N$ es igual al doble de la matriz

$$C(q) = (q + 1)\,B(q) - b(q) \cdot b(q)'$$

(es decir, una matriz como la misma C, pero de dimensión (q, q)).

Finalmente, el problema de valores y vectores propios de dimensión $q+1$ que aparece en el cálculo de los vectores propios de C contenidos en F en el caso $n = 2\,q+1$, también puede expresarse en función de la matriz $B(q+1)$. Dado que $W \cdot u = W \cdot v = (q + 1)\,b(q)$, este problema queda

$$\begin{pmatrix} 2\,(q+1)\,B(q) & (q+1)\,b(q) \\ 2\,(q+1)\,b(q)' & (q+1)^2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

que, tomando $x_{q+1} = y/2$, y teniendo en cuenta la estructura de la matriz B, es igual a la ecuación

$$2\,(q+1)\,B(q+1)\cdot x = \lambda\, D(q+1)\cdot x$$

donde $x = (x_1, \ldots, x_{q+1})$, y $D(q+1)$ es la matriz diagonal de dimensión $(q+1, q+1)$ cuyos n primeros elementos son iguales a 1, y el elemento $q+1$ es igual a 2.

Obsérvese, además, que se verifica la igualdad

$$B(q) + \widetilde{B}(q) = 2\,C(q)$$

En resumen, se manifiesta que las tres familias de matrices $C(q)$, $B(q)$ y $\widetilde{B}(q)$ están inseparablemente relacionadas, no solamente por el hecho de aparecer las B y $\widetilde{B}$ como cajas de dimensión mitad al estudiar la matriz C, sino también cuando se comparan en igual dimensión.

A continuación debemos abordar el estudio de la estructura de vectores propios de estas matrices. Por claridad de exposición, hemos agrupado todos los cálculos y demostraciones en el capítulo 3, y en la sección siguiente empleamos los resultados para esclarecer la estructura de las coordenadas principales de la distancia Valor Absoluto.

2.4 Coordenadas principales de la distancia Valor Absoluto

En esta sección se aplican los resultados obtenidos en el capítulo 3 sobre las matrices B, $\widetilde{B}$ y C a la descripción de las coordenadas principales para la distancia valor absoluto.

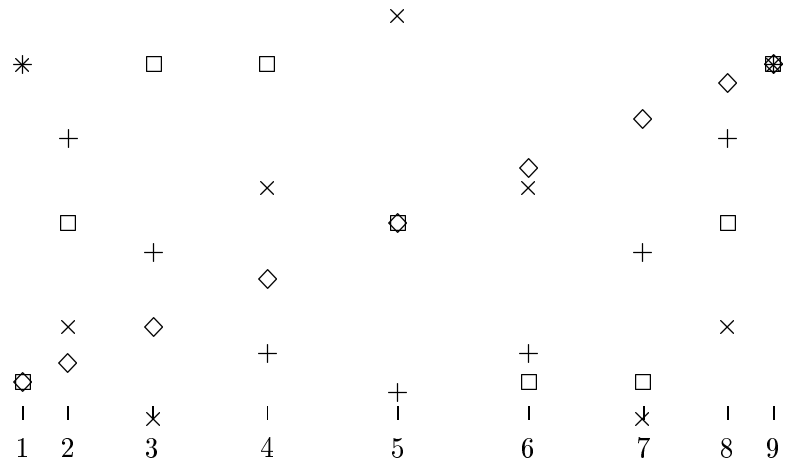
La conclusión a que se llega, principal resultado de este capítulo, es la interpretación de las coordenadas principales de la distancia Valor Absoluto como funciones polinómicas de grado creciente con el índice del correspondiente eje principal.

Este hecho proporciona una explicación razonable al buen comportamiento de la regresión basada en la distancia Valor Absoluto, pues permite interpretar este modelo como una proyección del vector de observaciones sobre ejes principales que definen dimensiones lineales, cuadráticas, cúbicas, etc.

Ejemplo: Para $n+1 = 9$ obtenemos las siguientes 4 primeras coordenadas principales, asimilables a dimensiones de tipo lineal, cuadrática, cúbica, y cuártica.

	Coordenada			
	1	2	3	4
Autovalor:	8.290	2.137	1.000	0.605
Individuo				
1	–1.337	0.648	–0.408	0.2809
2	–1.176	0.345	0.000	–0.1833
3	–0.873	–0.120	0.408	–0.3446
4	–0.464	–0.528	0.408	0.0637
5	0.000	–0.689	0.000	0.3667
6	0.464	–0.528	–0.408	0.0637
7	0.873	–0.120	–0.408	–0.3446
8	1.176	0.345	0.000	–0.1833
9	1.337	0.648	0.408	0.2809
Dimensión	Lineal ($\diamond$)	Cuadrática ($+$)	Cúbica ($\square$)	Cuártica ($\times$)

El gráfico siguiente corresponde a estas coordenadas principales, y pone de manifiesto la interpretación de cada coordenada. (Se ha escalado independientemente cada coordenada, para mejor visualización)



Como advertíamos en la sección 2.2.1, en lugar de hallar las coordenadas principales sobre el conjunto de $n+1$ puntos equidistantes sobre una recta, dotado de la distancia Valor absoluto, $\boldsymbol{\delta}(x_i, x_j) = \sqrt{|x_i - x_j|}$ por aplicación directa del teorema 1.5.1, practicamos un análisis de componentes principales sobre la matriz de datos 2.19

$$X_{i\,j} = \begin{cases} -(n+1)+j & \text{si} \quad i < j \\ j & \text{si} \quad i \geq j \end{cases} \qquad \left(\begin{array}{l} i = 0, \ldots, n \\ j = 1, \ldots, n \end{array} \right)$$

y, una vez obtenida la descomposición espectral de la matriz de covarianzas $C = V \cdot \Lambda \cdot V'$, calculamos la matriz Y de coordenadas principales por

$$Y = X \cdot V$$

TEOREMA 2.4.1 *Los elementos de la columna j de Y, para $j = 1, \ldots, n$, vienen dados por*

$$Y_{i\,j} = a_j - b_j\, T_j(z_i) \qquad i = 1, \ldots, n+1 \tag{2.28}$$

donde T_j es el j–ésimo polinomio de Tchebytchev de primera especie, a_j y b_j son constantes, y los $z_i \in [-1, 1]$ son los $n+1$ ceros del polinomio $T_{n+1}(z)$, es decir,

$$z_i = \cos\left(\frac{2\,i-1}{2\,n+2}\pi\right) \qquad i = 1, \ldots, n+1$$

Demostración: Empleando la expresión (3.20) para la matriz V de vectores propios

$$\begin{aligned} Y_{i\,j} &= \sum_{k=1}^{n} X_{i\,k}\, V_{k\,j} = \\ &= \frac{2}{\sqrt{2\,n+2}} \sum_{k=1}^{n} k \cdot \sin\left(\frac{k\,j}{n+1}\pi\right) - \\ &\quad - \sqrt{2\,n+2} \sum_{k=i}^{n} \sin\left(\frac{k\,j}{n+1}\pi\right) \end{aligned}$$

Ahora, empleando la identidad

$$\sum_{k=1}^{n} \sin(k\,\alpha) = -\frac{1}{2}\left[\frac{\cos\left(\left(n+\frac{1}{2}\right)\alpha\right) - \cos\left(\frac{\alpha}{2}\right)}{\sin\left(\frac{\alpha}{2}\right)}\right]$$

en el segundo sumando, obtenemos

$$\sum_{k=i}^{n} \sin\left(\frac{k\,j}{n+1}\pi\right) = \frac{1}{2}\left[\frac{\cos\left(\left(i-\frac{1}{2}\right)\,\left(\frac{j}{n+1}\pi\right)\right) - \cos\left(\left(n+\frac{1}{2}\right)\,\left(\frac{j}{n+1}\pi\right)\right)}{\sin\left(\frac{j}{2\,n+2}\pi\right)}\right]$$

Agrupando como a_j y b_j los términos que no dependen de i, y teniendo en cuenta la definición de los polinomios de Tchebytchev de primera especie,

$$T_j(z) = \cos(j\,\theta) \qquad \text{siendo } z = \cos(\theta)$$

obtenemos el enunciado. □

Chapter 3

Estructura de una clase paramétrica de matrices

3.1 Propiedades elementales

En este capítulo vamos a estudiar las matrices C, B y $\widetilde{B}$ introducidas en el capítulo 2, que supondremos de dimensión $(n,\, n)$ salvo indicación contraria. Veremos que aparecen como casos particulares de una clase paramétrica de matrices $F_n(a)$.

Los elementos de estas matrices son

$$\begin{aligned} B_{i\,j} &= \min\{i,j\} \qquad i,j = 1,\ldots,n \\ \widetilde{B}_{i\,j} &= (2\,n+1)\cdot\min\{i,j\} - 2\,i\,j \qquad i,j = 1,\ldots,n \\ C_{i\,j} &= (n+1)\cdot\min\{i,j\} - i\,j \qquad i,j = 1,\ldots,n \end{aligned} \tag{3.1}$$

Se verifican las siguientes relaciones entre las matrices B, $\widetilde{B}$ y C:

$$\begin{aligned} C &= (n+1)\,B - b\cdot b' \\ \widetilde{B} &= (2\,n+1)\,B - 2\,b\cdot b' \\ C &= \frac{1}{2}[B+\widetilde{B}] \end{aligned} \tag{3.2}$$

siendo $b = (1,2,\ldots,n)'$. De hecho, dado que b coincide con la última columna de B, vemos que la estructura de B es fundamental para el estudio de las otras dos matrices.

Otra propiedad relevante de B es su descomposición *à la* Cholesky:

$$B = M\cdot M' = N^2 \tag{3.3}$$

siendo

$$M = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & & & \ddots & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \qquad N = \begin{pmatrix} 0 & \ldots & 0 & 0 & 1 \\ 0 & \ldots & 0 & 1 & 1 \\ 0 & \ldots & 1 & 1 & 1 \\ \vdots & & & & \vdots \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

En particular, B es definida positiva y $\det(B) = 1$. Igualmente C es definida positiva, al ser una matriz de covarianzas, y según veremos en el apartado siguiente, también $\widetilde{B}$ tiene esta propiedad.

Los valores propios de B son conocidos (Véase Frank [43])

$$\lambda_i = \frac{1}{2}\left[1 - \cos\left(\frac{2\,i-1}{2\,n+1}\pi\right)\right]^{-1} \qquad i = 1,\ldots,n \tag{3.4}$$

mientras que los vectores propios no parecen ser bien conocidos. Según Graybill ([53], pp. 186–187), B pertenece a una familia de matrices cuya estructura característica "es difícil de evaluar en general".

Las relaciones 3.2 y las propiedades vistas en el capítulo anterior indican que los vectores propios de B, $\widetilde{B}$ y C deben estar muy relacionados, como confirmamos en lo que sigue.

Introducimos la siguiente familia paramétrica de matrices (dependientes del parámetro a) [1]

$$F_n(a) = \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & -1 & a \end{pmatrix} \tag{3.5}$$

Es inmediato comprobar que

$$\det F_n = n\,(a-1) + 1$$

También por cálculo directo se muestra que $F_n(1) = B^{-1}$, de modo que

$$F_n(a) = B^{-1} + (a-1) \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \cdot (0, \dots, 0, 1)$$

expresión que permite el cálculo de la inversa de $F_n(a)$

$$F_n^{-1}(a) = \frac{1}{n\,(a-1)+1}\, G_n(a)$$

donde

$$G_n(a) = (n\,(a-1)+1)\, B - (a-1)\, b \cdot b' \tag{3.6}$$

La importancia de $G_n(a)$ se debe a que permite generar los tres tipos de matrices anteriores. Comparando con las relaciones 3.2 se observa que

$$\begin{aligned} B &= G_n\,(1) \\ C &= G_n\,(2) \\ \widetilde{B} &= G_n\,(3) \end{aligned} \tag{3.7}$$

Ejemplo: Para $n = 4$ obtenemos

$$G_n(1) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix} \quad G_n(2) = \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 6 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix} \quad G_n(3) = \begin{pmatrix} 7 & 5 & 3 & 1 \\ 5 & 10 & 6 & 2 \\ 3 & 6 & 9 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

[1] Obsérvese que $F_n(2)$ es una matriz de Toeplitz.

3.2 Vectores propios

3.2.1 Valores y vectores propios de $F_n(a)$

Los valores y vectores propios de $F_n(a)$ se expresan por medio de los polinomios de Tchebytchev de segunda especie:

$$\begin{aligned} U_0(\xi) &= 1 \\ U_1(\xi) &= 2\,\xi \\ U_{k+2}(\xi) &= 2\,\xi\,U_{k+1}(\xi) - U_k(\xi) \qquad (k \geq 0) \end{aligned}$$

TEOREMA 3.2.1 *Sea μ un valor propio de $F_n(a)$ con vector propio $v = (v_1, \ldots, v_n)$. Entonces:*

- *$\xi = 1 - \mu/2$ es una raiz del polinomio*

$$q_n(\xi) = U_n(\xi) + (a-2)\,U_{n-1}(\xi) \tag{3.8}$$

- *Las componentes de v son*

$$v_i = \frac{2\,\sin(i\,\theta)}{\sqrt{2\,n+1-u_{2n}(\xi)}} \qquad i = 1, \ldots, n \tag{3.9}$$

donde θ se define por $\xi = \cos\theta$.

Demostración: Sea $v = (v_1, v_2, \ldots, v_n)$ vector propio de $F_n(a)$ con valor propio μ. Entonces

$$\left.\begin{array}{rcrcrcl} & & 2\,v_1 & - & v_2 & = & \mu\,v_1 \\ -v_1 & + & 2\,v_2 & - & v_3 & = & \mu\,v_2 \\ \vdots & & \vdots & & \vdots & & \vdots \\ -v_{n-2} & + & 2\,v_{n-1} & - & v_n & = & \mu\,v_{n-1} \\ -v_{n-1} & + & 2\,v_n & & & = & \mu\,v_n \end{array}\right\}$$

Pongamos $\xi = 1 - \mu/2$, es decir, $\mu = 2\,(1-\xi)$. Entonces

$$\left.\begin{array}{rcrcl} v_2 &= 2\,\xi\,v_1 & & & \\ v_3 &= 2\,\xi\,v_2 & - & v_1 & \\ \vdots & \vdots & & \vdots & \\ v_n &= 2\,\xi\,v_{n-1} & - & v_{n-2} & \\ (2-a)\,v_n &= 2\,\xi\,v_n & - & v_{n-1} & \end{array}\right\} \tag{3.10}$$

Teniendo en cuenta la fórmula de recurrencia de los polinomios de Tchebytchev de segunda especie, vemos que haciendo $v_1 = 1$ se cumple que $v_i = U_{i-1}(\xi)$. En general, las componentes del vector propio de valor propio μ serán:

$$v_i = U_{i-1}(\xi)\, v_1 \qquad i = 1, \ldots, n-1 \tag{3.11}$$

Por otra parte, el polinomio característico de $F_n(a)$ en función de ξ es

$$q_n(\xi) = \det(F_n(a) - 2\,(1-\xi)\, I_n)$$

Para $a = 2$, obtenemos de nuevo los polinomios de Tchebytchev

$$q_1(\xi) = 2\,\xi, \qquad q_2(\xi) = 2\,\xi^2 + 1, \quad \ldots \quad q_n(\xi) = U_n(\xi)$$

Sumando la columna $(0, 0, \ldots, a-2)'$ a la última columna de $F_n(2)$ obtenemos $F_n(a)$, y desarrollando el determinante obtenemos el polinomio característico de $F_n(a)$

$$q_n(\xi) = U_n(\xi) + (a-2)\, U_{n-1}(\xi) \tag{3.12}$$

Seguidamente calculamos v_1 imponiendo la condición de módulo 1 para el vector v

$$1 = \sum_{i=1}^{n} v_i^2 = v_1^2 \sum_{i=0}^{n-1} U_i^2(\xi)$$

Según la identidad de Christoffel–Darboux se tiene que

$$\sum_{i=0}^{n-1} U_i^2(\xi) = \frac{1}{2}\,[U_n'(\xi)\, U_{n-1}(\xi) - U_n(\xi)\, U_{n-1}(\xi)]$$

Evaluamos esta expresión empleando la representación trigonométrica de los polinomios u_n

$$U_n(\xi) = \frac{\sin\,(n+1)\,\theta}{\sin\,\theta} \qquad \text{siendo } \xi = \cos\theta$$

Teniendo en cuenta que

$$U_n'(\xi) = \frac{(n+1)\, U_{n-1}(\xi) - n\,\xi\, U_{n-1}(\xi)}{1-\xi^2}$$

(véase, por ejemplo [7], pág. 116), y que

$$\frac{d\, U_n(\xi)}{d\,\theta} = -\sin(\theta)\, U_n(\xi)$$

se concluye que

$$\begin{aligned} U_n'(\xi)\, U_{n-1}(\xi) &= \\ &= \frac{1}{\sin^2(\theta)} \left[\cos(\theta)\, U_n(\xi)\, U_{n-1}(\xi) - (n+1)\, \cos((n+1)\,\theta)\, U_{n-1}(\xi)\right] \\ U_n(\xi)\, U_{n-1}'(\xi) &= \\ &= \frac{1}{\sin^2(\theta)} \left[\cos(\theta)\, U_n(\xi)\, U_{n-1}(\xi) - n\, \cos(n\,\theta)\, U_n(\xi)\right] \end{aligned}$$

Utilizando la identidad $\sin(x)\, \cos(y) = 1/2\, [\sin(x+y)+\sin(x-y)]$ se llega a

$$\sum_{i=0}^{n-1} U_i^2(\xi) = \frac{1}{4\, \sin^2(\theta)}\, (2\, n + 1 - U_{2n}(\xi))$$

Luego

$$v_1 = \frac{\pm 2\, \sin(\theta)}{\sqrt{2\, n + 1 - U_{2\, n}(\xi)}}$$

Según 3.11, las componentes del vector propio $v = (v_1, \ldots, v_n)$ son (salvo el signo)

$$v_i = \frac{2\, \sin(i\, \theta)}{\sqrt{2\, n + 1 - u_{2\, n}(\xi)}} \qquad i = 1, \ldots, n \tag{3.13}$$

El valor propio correspondiente a v es

$$\mu = 2\, (1 - \cos(\theta)) \tag{3.14}$$

Siendo θ tal que

$$u_n(\cos(\theta)) = (2 - a)\, U_{n-1}(\cos(\theta)) \tag{3.15}$$

□

Nota 3.2.1 *Como*

$$F_n^{-1}(a) = [\det F_n(a)]^{-1}\, G_n(a)$$

el valor propio λ *de* $G_n(a)$ *correspondiente al valor propio* μ *de* $F_n(a)$ *es*

$$\lambda = \frac{(a-1)\, n + 1}{\mu} \tag{3.16}$$

3.2.2 Vectores propios de B, C y $\widetilde{B}$

Los valores y vectores propios de las matrices B, C y $\widetilde{B}$ se obtienen como consecuencia de los cálculos anteriores

Matriz B $(a = 1)$

Sustituyendo $a = 1$ en 3.15 resulta $\sin((n+1)\,\theta) - \sin(n\,\theta) = 0$, que según la identidad $\sin(x) - \sin(y) = 2\,\sin((x-y)/2)\,\cos((x+y)/2)$, equivale a

$$\sin\left(\frac{\theta}{2}\right)\,\cos\left(\frac{(2\,n+1)\,\theta}{2}\right) = 0$$

Las soluciones de esta ecuación son

$$\theta_j = \frac{2\,j-1}{2\,n+1}\,\pi \qquad j = 1, \ldots, n$$

y cumplen que $U_{2\,n}(\theta_i) = 0$.

Luego las componentes del vector propio $v_j = (v_{1\,j}, \ldots, v_{n\,j})'$ son

$$v_{i\,j} = \frac{2}{\sqrt{2\,n+1}}\,\sin\left(\frac{i\,(2\,j-1)}{2\,n+1}\,\pi\right) \qquad i = 1, \ldots, n \tag{3.17}$$

y su valor propio es

$$\mu_j = 2\,\left(1 - \cos\frac{2\,j-1}{2\,n+1}\,\pi\right)$$

De 3.16 resulta el correspondiente valor propio de B

$$\lambda_j = \frac{1}{2}\left[1 - \cos\left(\frac{2\,j-1}{2\,n+1}\,\pi\right)\right]^{-1} \tag{3.18}$$

en coincidencia con el resultado 3.4.

Obsérvese que los λ_j en (3.18) quedan ordenados en sucesión decreciente. Si, en cambio, tomamos el orden opuesto, vemos que

Proposición 3.2.1 *Ordenando los valores y vectores propios de B de forma que la sucesión de valores propios sea creciente, las componentes del vector $v_j = (v_{1\,j}, \ldots, v_{n\,j})'$ vienen dadas por*

$$v_{i\,j} = (-1)^{i+j+1}\,\frac{2}{\sqrt{2\,n+1}}\,\sin\left(\frac{2\,i\,j}{2\,n+1}\pi\right) \qquad i = 1, \ldots, n \tag{3.19}$$

y el correspondiente valor propio es

$$\lambda_j = \frac{1}{4}\,\sec^2\left(\frac{j}{2\,n+1}\pi\right)$$

En particular, la matriz $V = (v_1, \ldots, v_n)$ es (ortogonal) simétrica.

Demostración: Tomemos $j' = n + 1 - j$. Entonces

$$\begin{aligned}\sin\left(\frac{i\,(2\,j-1)}{2\,n+1}\,\pi\right) &= \sin\left(i\,\pi - \frac{2\,i\,j'}{2\,n+1}\,\pi\right)\\ &= -\cos(i\,\pi)\,\sin\left(\frac{2\,i\,j'}{2\,n+1}\,\pi\right)\\ &= (-1)^{i+1}\,\sin\left(\frac{2\,i\,j'}{2\,n+1}\pi\right)\end{aligned}$$

Finalmente, multiplicando cada vector v_j por el factor constante $(-1)^j$, operación que conserva la estructura de vectores propios, se obtiene el enunciado. Un cálculo análogo proporciona la fórmula para los λ_j. □

Matriz C ($a = 2$)

Ahora $q_n(\xi) = U_n(\xi)$, luego las raíces características son los ceros del polinomio $U_n(\xi)$, es decir,

$$\theta_j = \frac{j}{n+1}\,\pi \qquad j = 1, \ldots, n$$

Puesto que para todos estos valores propios se verifica que $U_{2\,n}(\xi_j) = -1$, las componentes del vector propio $v_j = (v_{1\,j}, \ldots, v_{n\,j})$ son

$$v_{i\,j} = \frac{2}{\sqrt{2\,n+2}}\,\sin\left(\frac{i\,j}{n+1}\,\pi\right) \qquad i = 1, \ldots, n \tag{3.20}$$

y su valor propio es

$$\mu_j = 2\,\left[1 - \cos\left(\frac{j}{n+1}\,\pi\right)\right]$$

El valor propio correspondiente de C se deduce de 3.16

$$\lambda_j = \frac{n+1}{2}\,\left[1 - \cos\left(\frac{j}{n+1}\,\pi\right)\right]^{-1}$$

Vemos que la matriz de vectores propios de C es también simétrica. Además, estamos en condiciones de asignar cada vector propio al subespacio par o impar según la notación empleada en el capítulo 2.

Proposición 3.2.2 *El vector propio v_j de C dado por (3.20) pertenece al subespacio par de $\mathbf{R}^n$ si j es par, y al espacio impar si j es impar*

Demostración: Empleando la fórmula (3.20), es inmediato ver que

$$v_{n+1-i,\,j} = (-1)^{j+1}\, v_{i\,j}$$

que equivale al enunciado, teniendo en cuenta la descripción dada en el capítulo 2 de los subespacios par e impar. □

Matriz $\widetilde{B}$ ($a = 3$)

En este caso, 3.15 equivale a $\sin((n+1)\,\theta) + \sin(\theta) = 0$. De la identidad $\sin(x) + \sin(y) = 2\,\sin((x+y)/2)\,\cos((x-y)/2)$ resulta la ecuación

$$\sin\left(\frac{(2\,n+1)\,\theta}{2}\right)\,\cos\left(\frac{\theta}{2}\right) = 0$$

cuya solución es

$$\theta_j = \frac{2\,j}{2\,n+1}\,\pi \qquad j = 1, \ldots, n$$

que también anula $U_{2\,n}(\xi)$. Luego las componentes de $v_j = (v_{1\,j}, \ldots, v_{n\,j})$ son

$$v_{i\,j} = \frac{2}{\sqrt{2\,n+1}}\,\sin\left[\frac{2\,i\,j}{2\,n+1}\,\pi\right] \qquad i = 1, \ldots, n \tag{3.21}$$

Su valor propio es

$$\mu_j = 2\,\left[1 - \cos\left(\frac{2\,j}{2\,n+1}\,\pi\right)\right]$$

Finalmente, el correspondiente valor propio de $\widetilde{B}$ es

$$\lambda_j = \frac{2\,n+1}{2}\,\left[1 - \cos\left(\frac{2\,j}{2\,n+1}\,\pi\right)\right]^{-1}$$

Una vez más, la matriz de vectores propios obtenida es una matriz (ortogonal) simétrica. Además, comparando la expresión (3.21) con la (3.19), vemos que las dos matrices de vectores propios difieren únicamente en signos alternados.

3.3 Estructura de los vectores propios de B

3.3.1 Introducción

En esta sección vamos a ver que, para un conjunto amplio de valores de n, los vectores propios de B se pueden generar de un modo bastante simple a partir del primero.

Ejemplo 3.3.1 *Supongamos $n = 5$. Para visualizar las relaciones existentes entre las columnas de la matriz V de vectores propios de B empleemos $(1, 2, 3, 4, 5)'$ para simbolizar las componentes del primer vector propio*

$$(0.1699,\ 0.3260,\ 0.4557,\ 0.5485,\ 0.5969)'$$

Con esta notación, la matriz V tiene el siguiente aspecto:

$$V = \begin{array}{c} \begin{array}{ccccc} v_1 & v_2 & v_3 & v_4 & v_5 \end{array} \\ \left(\begin{array}{rrrrr} 1 & 5 & 3 & -4 & 2 \\ 2 & 1 & 5 & 3 & -4 \\ 3 & -4 & 2 & 1 & 5 \\ 4 & -2 & -1 & -5 & -3 \\ 5 & 3 & -4 & 2 & 1 \end{array} \right) \end{array}$$

Se observa que los demás vectores propios se obtienen permutando las componentes de v_1, salvo el signo. Además, tomando la matriz signo–permutable

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

se comprueba fácilmente que

$$\begin{aligned} Q^5 &= I \quad (= Q^0) \\ v_j &= Q^{j-1}\, v_1 \qquad j = 1, \ldots, 4 \end{aligned}$$

Es decir, los vectores propios de B son generados a partir del primero por la acción de potencias de la matriz Q.

3.3.2 Permutación de componentes del primer vector propio

En este apartado se muestra como se pueden obtener las componentes de cualquier vector propio de B a partir de las del primero.

Proposición 3.3.1 *Sea* $v_1 = (v_{1\,1}, \ldots, v_{n\,1})'$ *el primer vector propio. Sea* $v_{i\,j} \neq 0$ *una componente no nula de otro vector propio* $v_j = (v_{1\,j}, \ldots, v_{n\,j})'$. *Entonces, existe un entero* k, $(1 \leq k \leq n)$, *para el cual se verifica que*

$$|v_{i\,j}| = |v_{k\,1}| \tag{3.22}$$

Demostración: Sea h el único entero congruente con $i \cdot j$ módulo $2\,n+1$ contenido en el intervalo $[0, 2\,n]$. El caso $h = 0$ queda excluido por la hipótesis, ya que equivale a $v_{i\,j} = 0$. En los restantes casos tratamos separadamente las dos posibilidades

- Si $h \in [1, n]$, tomamos $k = h$. Como $i \cdot j = \alpha(2\,n+1) + k$, para algún $\alpha \geq 0$, teniendo en cuenta 3.19, resulta $v_{i\,j} = (-1)^{i+j+k+1}\, v_{k\,1}$.
- Si $h \in [n+1, 2\,n]$, tomamos $k = 2\,n + 1 - h$, con lo que $i \cdot j = \alpha(2\,n+1) - k$, y análogamente al caso anterior, obtenemos $v_{i\,j} = (-1)^{i+j+k}\, v_{k\,1}$.

□

Proposición 3.3.2 *El primer vector propio* $v_1 = (v_{1\,1}, \ldots, v_{n\,1})'$ *verifica las desigualdades*

$$0 < v_{1\,1} < \ldots < v_{n\,1} \tag{3.23}$$

y, si $v_j = (v_{1\,j}, \ldots, v_{n\,j})'$ *es otro vector propio con todas sus componentes no nulas, entonces la sucesión de los valores absolutos*

$$\{|v_{1\,j}|, \ldots, |v_{n\,j}|\}$$

es una permutación de

$$\{|v_{1\,1}|, \ldots, |v_{n\,1}|\}$$

Demostración: Como todos los elementos de la matriz B son positivos, podemos tomar todas las componentes de v_1 positivas (teorema de Perron). Sea $w_1 = (w_{1\,1}, \ldots, w_{n\,1})' = B \cdot v_1$. Entonces

$$w_{i\,1} - w_{(i-1)\,1} = \sum_{k=i}^{n} v_{k\,1} > 0 \tag{3.24}$$

y, dado que $w_1 = \lambda_1 v_1$, se siguen las desigualdades 3.23.

Para probar la segunda afirmación, teniendo en cuenta la proposición 3.3.1, es suficiente ver que si v_j no tiene componentes nulas, entonces los valores absolutos de todas las componentes son distintos.

Supongamos que, por el contrario, existe un par $i_1 < i_2$ de índices tales que $|v_{i_1\,j}| = |v_{i_2\,j}|$. Esto equivale a que $|\sin(i_1\,\varphi)| = |\sin(i_2\,\varphi)|$ (donde $\varphi = \frac{2\,j}{2\,n+1}\pi$).

Es decir, para algún $k > 0$ se verifica una de las igualdades

$$\begin{aligned} i_1\,\varphi &= k\,\pi + i_2\,\varphi \\ i_1\,\varphi &= k\,\pi - i_2\,\varphi \end{aligned}$$

Si la primera es cierta, $(i_2 - i_1)\varphi = k\,\pi$, y la componente $v_{i_0\,j}$ de este vector (siendo $i_0 = i_2 - i_1$) es nula, en contradicción con la hipótesis.

Si la segunda es cierta, llegamos a la misma conclusión, tomando $i_0 = i_2 + i_1$ si esta suma es ≤ 0, o bien $i_0 = 2\,n + 1 - (i_2 + i_1)$ si la suma es > 0. □

3.3.3 Vectores propios con componentes nulas

Seguidamente vamos a caracterizar los casos en que se presentan ceros como componentes de los vectores propios. En las proposiciones 3.3.3 y 3.3.4 se emplea la expresión 3.19 para los vectores propios.

Proposición 3.3.3 *Supongamos que $v_{i\,j} = 0$. Entonces*

1. *Necesariamente $i \geq 3$ y $j \geq 3$*
2. *La dimensión n es de la forma*

$$n = \dot{p} + \frac{(p-1)}{2} \tag{3.25}$$

   para algún número primo p.

3. *p divide a i ó a j.*

Demostración: Deducimos 1 del hecho que $v_{i\,j} = 0 \iff 2\,i\,j = 2\,k\,(2\,n+1)$ para algún entero k, $(0 < k < n/2)$. El mínimo valor del producto $i\,j$ que cumple la condición es $i\,j = 2\,n + 1$, luego uno de los dos factores es mayor que 2.

Puesto que en la igualdad $2\,i\,j = 2\,k\,(2\,n+1)$ el entero k está comprendido estrictamente entre 0 y $n/2$, y tanto i como j son $\leq n$, necesariamente $2\,n+1$ es no primo.

Sea $p > 2$ un factor primo de n. Como $(2\,n+1)/p$ también es impar, podemos escribir $2\,n+1 = p\,(2\,r+1)$, para un entero r, $0 < r < n/p$. Esta igualdad equivale a la relación 2. Por último, las igualdades anteriores hacen evidente 3. □

Tabla 4.1 Algunos valores de n con componentes nulas

p	$n = \dot{p} + (p-1)/2$				
3	4	7	10	13	16
5	7	12	17	22	27
7	10	17	24	31	38
11	16	27	38	49	60
13	19	32	45	58	71

Proposición 3.3.4 *Supongamos que n cumple la condición 3.25, de modo que es posible la existencia de elementos $v_{i\,j}$ nulos. Entonces*

1. *$v_{i\,j} \neq 0$, siempre que se cumpla una de las condiciones siguientes*

$$i \leq 2, \qquad j \leq 2, \qquad i = n, \qquad j = n$$

2. *$v_{i\,j} = 0$ si $i \cdot j$ es múltiplo de $2\,n+1$*

3. *Si $v_{i\,j} = 0$, entonces se verifican las igualdades*

$$\begin{array}{lcccl} v_{i+k,\,j} & = & -v_{i-k,\,j} & = & \pm v_{k\,j} \\ v_{n-i-k,\,j} & = & v_{n-i+k+1,\,j} & = & \pm v_{2\,k+1,\,j} \end{array}$$

   para todos los $k \geq 0$ para los que las expresiones en los subíndices estén comprendidas en el intervalo $[1, n]$.

4. *Si $v_{i\,j} = 0$ siendo $i = n/2+1$ ó $i = (n-1)/2$ (según la paridad de n), entonces $\lambda_j = 1$ es valor propio de B.*

Demostración: La primera parte de (1) ya se ha visto en la proposición 3.3.3. Puesto que $B = N^2$ (3.3), si $B\,v_j = \lambda_j\,v_j$, entonces $N^{-1}\,v_j = \lambda^{-1/2}\,v_j$. Teniendo en cuenta que

$$N^{-1} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & -1 & 1 \\ 0 & 0 & 0 & \dots & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 \\ \vdots & \vdots & & & & \vdots & \vdots \\ 0 & -1 & 1 & \dots & 0 & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 & 0 \end{pmatrix}$$

resulta que las componentes de v_j verifican

$$\begin{aligned} v_{n-i+1,\,j} - v_{n-i,\,j} &= \lambda^{-1/2}\, v_{i\,j} \\ v_{1\,j} &= \lambda^{-1/2}\, v_{n\,j} \end{aligned} \qquad (3.26)$$

y, al ser no nulo $v_{1\,j}$, también $v_{n\,j}$.

(2) es evidente a partir de 3.19

Para (3), de la anulación de $v_{i\,j}$ se deducen las igualdades:

$$\left.\begin{array}{rcl} \sin\left(\frac{2\,(i+k)\,j}{2\,n+1}\,\pi\right) & = & \cos\left(\frac{2\,i\,j}{2\,n+1}\,\pi\right)\cdot\sin\left(\frac{2\,k\,j}{2\,n+1}\,\pi\right) \\ \sin\left(\frac{2\,(i-k)\,j}{2\,n+1}\,\pi\right) & = & -\cos\left(\frac{2\,i\,j}{2\,n+1}\,\pi\right)\cdot\sin\left(\frac{2\,k\,j}{2\,n+1}\,\pi\right) \end{array}\right\}$$

que equivalen a $v_{i+k,\,j} = -v_{i-k,\,j} = \pm v_{k\,j}$

Análogamente se obtiene que $v_{n-i-k,\,j} = v_{n-i+k+1,\,j}$, de las identidades

$$\begin{aligned} 2\,(n-i-k)\,j &= -2\,i\,j + (2\,n+1)\,j - (2\,k+1)\,j \\ 2\,(n-i+k+1)\,j &= -2\,i\,j + (2\,n+1)\,j + (2\,k+1)\,j \end{aligned}$$

Para (4), supongamos que $v_{i\,j} = 0$, y que $k = n/2 + 1$. La ecuación $k-1$ de (3.26) es

$$v_{k\,j} - v_{k-1,\,j} = \lambda^{-1/2}\, v_{k-1,\,j}$$

por tanto, $\lambda = \lambda^{-1/2} = -1$. De forma análoga, si $k = (n-1)/2$, la ecuación $k+1$ implica que $\lambda = 1$. □

3.3.4 Generación de los vectores propios a partir del primero

En este apartado volvemos a emplear la fórmula (3.17)

$$v_{i\,j} = \sin\left(\frac{i\,(2\,j-1)}{2\,n+1}\,\pi\right) \qquad i = 1, \ldots, n$$

para los vectores propios de B, pero para claridad de notación en los razonamientos que siguen, omitiremos el factor $2/\sqrt{2\,n+1}$.

Estamos interesados en hallar la expresión de la transformación lineal $v_j \longrightarrow v_{j'}$ tal que

$$v_{1\,j} = \pm\, v_{2\,j'} \qquad (3.27)$$

donde v_j y $v_{j'}$ son dos vectores propios de B. Entonces

$$\sin\left(\frac{(2\,j-1)}{2\,n+1}\,\pi\right) = \pm \sin\left(\frac{2\,(2\,j'-1)}{2\,n+1}\,\pi\right)$$

con lo que tenemos dos posibilidades a) $v_{1\,j} = \quad v_{2\,j'}$. En este caso resulta la igualdad

$$\frac{(2\,j-1)}{2\,n+1}\pi = \pi - \frac{2\,(2\,j'-1)}{2\,n+1}\pi \Rightarrow j' = \frac{n-j+2}{2}$$

válida si n y j son ambos pares o impares ($n+j$ par).

b) $v_{1\,j} = -v_{2\,j'}$. En este caso resulta la igualdad

$$\frac{(2\,j-1)}{2\,n+1}\pi + \pi = \frac{2\,(2\,j'-1)}{2\,n+1}\pi \Rightarrow j' = \frac{n-j+2}{2}$$

válida si n es par y j impar, ó bien si n es impar y j es par ($n+j$ impar)

Hemos obtenido la transformación $j \longrightarrow j'$, es decir la correspondencia entre los índices de los vectores relacionados por (3.27).

Seguidamente estudiamos las transformaciones $i \longrightarrow i'$ tales que

$$v_{i\,j} = v_{i'\,j'}$$

en los dos casos expuestos.

a) Sea $j' = (n-j+2)/2 \iff v_{1\,j} = v_{2\,j'}$. n y j tienen la misma paridad, $n+j$ es par, y $2\,j'-1 = n-j+1$

a.1) Si $2\,i \leq n$, tomemos $i' = 2\,i$

$$\sin\left(\frac{i\,(2\,j-1)}{2\,n+1}\,\pi\right) = \pm\,\sin\left(\frac{2\,i\,(n-j+1)}{2\,n+1}\,\pi\right)$$

De la identidad

$$i\,(2\,j-1) - i\,(2\,n+1) = -2\,i\,(n-j+1)$$

deducimos

$$\begin{aligned}\sin\left(\frac{i\,(2\,j-1)}{2\,n+1}\,\pi\right) &= \sin\left(i\,\pi - \frac{2\,i\,(n-j+1)}{2\,n+1}\,\pi\right) = \\ &= -\cos\left(i\,\pi\right)\cdot\sin\left(\frac{2\,i\,(n-j+1)}{2\,n+1}\,\pi\right) \\ &= (-1)^{k+1}\,\sin\left(\frac{2\,i\,(n-j+1)}{2\,n+1}\,\pi\right)\end{aligned}$$

Luego obtenemos

$$v_{i'\,j'} = (-1)^{i+1}\,v_{i\,j} \qquad 1 \leq i \leq (n-1)/2 \tag{3.28}$$

a.2) Si $n < 2\,i \leq 2\,n$, tomemos $i' = 2\,(n-k)+1$. De la identidad

$$i\,(2\,j-1) = (2\,(n-i)+1)\,(n-j+1) - (n-i-j+1)(2\,n+1)$$

deducimos

$$\sin\left(\frac{i\,(2\,j-1)}{2\,n+1}\,\pi\right) = \\ = \sin\left(\frac{(2\,(n-i)+1)\,(n-j+1)}{2\,n+1}\,\pi\right)\cdot\cos\left((n-i-j+1)\,\pi\right)$$

Como $n-1$ es par, obtenemos

$$v_{i'\,j'} = (-1)^{i+1}\,v_{i\,j} \qquad (n-1)/2 \le i \le n \tag{3.29}$$

b) Sea $j' = (n+j+1)/2 \iff v_{1\,j} = -v_{2\,j'}$. n y j tienen paridad opuesta, $n+j$ es impar, y $2\,j'-1 = n+j$

b.1) Si $2\,i \le n$, de la identidad

$$i\,(2\,j-1) = 2\,i\,(n+j) - j\,(2\,n+1)$$

se deduce

$$\sin\left(\frac{i\,(2\,j-1)}{2\,n+1}\,\pi\right) = \sin\left(\frac{2\,i\,(n+j)}{2\,n+1}\,\pi\right)\cdot\cos\left(i\,\pi\right)$$

Y finalmente

$$v_{i'\,j'} = (-1)^{i}\,v_{i\,j} \qquad 1 \le i \le (n-1)/2 \tag{3.30}$$

b.2) Si $n \le 2\,i \le 2\,n$

De la identidad

$$i\,(2\,j-1) = -(2\,(n-i)+1)(n+j) + (n-i+j)(2\,n+1)$$

deducimos

$$\sin\left(\frac{i\,(2\,j-1)}{2\,n+1}\,\pi\right) = -\sin\left(\frac{(2\,(n-i)+1)(n+j)}{2\,n+1}\,\pi\right)\cdot\cos\left((n+j-i)\,\pi\right)$$

Como $n+j$ es impar, concluimos que

$$v_{i'\,j'} = (-1)^{i}\,v_{i\,j} \qquad (n-1)/2 \le i \le n \tag{3.31}$$

Cuadro 3.3.1 Resumen del cambio $(i,j) \longrightarrow (i',j')$ (condicionado al punto de partida (3.27))

$n+j$ impar $\quad j' = \frac{n+j+1}{2}$ $\quad \left\{ \begin{array}{ll} 1 \le i \le (n-1)/2 & i' = 2\,i \\ (n-1)/2 < i \le n & i' = 2\,(n-i)+1 \end{array} \right\} \quad v_{i'\,j'} = (-1)^{i}\,v_{i\,j}$

$n+j$ par $\quad j' = \frac{n-j+2}{2}$ $\quad \left\{ \begin{array}{ll} 1 \le i \le (n-1)/2 & i' = 2\,i \\ (n-1)/2 < i \le n & i' = 2\,(n-i)+1 \end{array} \right\} \quad v_{i'\,j'} = (-1)^{i+1}\,v_{i\,j}$

Nota 3.3.1 *Como $i' = 2\,i$ si $2\,i < n$, y $i = n$ se transforma en $i' = 1$, resulta que cuando n es una potencia de 2 las formulas anteriores sólo generan permutaciones diferentes para $i \leq n/2$. En efecto, después de $n/2$ pasos, $i = 1$ se convierte en $i' = n$, y en la siguiente etapa, se transforma en $i = 1$, volviendo a aparecer el primer vector propio en lugar de una nueva permutación.*

Este inconveniente se evita fácilmente modificando la igualdad (3.27) empleada como punto de partida en el sentido de imponer que la transformación $v_j \longrightarrow v_{j'}$ verifique

$$v_{1\,j} = \pm\, v_{3\,i'} \tag{3.32}$$

es decir, exigimos que la primera componente pase a la tercera al cambiar el índice del vector propio.

Siguiendo un procedimiento que no detallamos aquí, análogo al caso ya descrito, obtendríamos el cambio $i \mapsto i'$ compatible con (3.32). Obsérvese que en el caso de n potencia de 3 nos encontraríamos con el mismo problema, esto es, $i = 1$ pasaría a $i' = 1$ en menos de $n - 1$ pasos.

Podemos asíenunciar la

Proposición 3.3.5 *Los cambios de componentes $(i, j) \mapsto (i', j')$ detallados en el cuadro 3.3.1. son válidos para todo n y generan permutaciones distintas de las componentes del primer vector propio, excepto en los casos*

- *n es una potencia de 2. En este caso bastaría imponer la condición (3.32) y efectuar los cambios necesarios*
- *$n = \dot{p} + (p-1)/2$ donde p es un número primo*

Estamos ahora en condiciones de enunciar el principal resultado de esta sección, que ya adelantábamos en el capítulo 2. Este resultado (teorema (3.3.1)) había sido conjeturado por Cuadras en 1990 ([26]) para $n \neq \dot{(3)} + 1$.

Empleamos en el enunciado la siguiente terminologia: Diremos que una matriz Q de orden $n \times n$ es *signo–permutable* (o matriz de permutación con signo) si cada fila y cada columna de Q contiene exactamente un elemento igual a 1 o a -1, mientras que los restantes elementos son nulos.

Es inmediato probar que una matriz signo–permutable Q verifica que:

$$Q' = Q^{-1} \qquad Q^{n-1} = Q' \qquad Q^n = I \tag{3.33}$$

(véase ejemplo en la sección 3.3.1)

TEOREMA 3.3.1 *Supongamos que $n \neq \dot{p} + (p-1)/2$, donde p es un número primo. Sea v_1 el primer vector propio de B. Existe entonces una matriz signo–permutable Q tal que*

$$v_j = Q^{j-1}\, v_1 \qquad j = 2, \ldots, n$$

son los restantes $n-1$ vectores propios de B.

Demostración: Teniendo en cuenta las fórmulas (3.28) a (3.31), si tomamos la matriz $Q = (q_{i\,i'})$ definida por

$$q_{i\,i'} = \begin{cases} (-1)^i & \text{si } i' = 2\,i \leq n-1 \text{ ó } i' = 2\,(n-i)+1 \leq n \\ 0 & \text{en caso contrario} \end{cases}$$

Entonces Q es signo–permutable, y pemite expresar las fórmulas (3.28) a (3.31) como

$$v_{j'} = \pm\, Q \cdot v_j$$

donde el signo depende de la paridad de j y n. Ahora, suponiendo que $n \neq 2^m$ (para todo m), basta observar que v_j y $v_{j'}$ son ambos vectores propios de B, que el paso $v_j \mapsto v_{j'}$ se consigue siempre con la misma Q, y finalmente, que $Q^n = I$.

En caso que $n = 2^m$ para algún m, buscaríamos otra matriz Q de manera análoga, reflejando el cambio de componentes $(i,j) \mapsto (i',j')$ tomando (3.32) como punto de partida. □

Chapter 4

Algunas generalizaciones

4.1 Caso general discreto

4.1.1 Introducción

En esta sección se plantea la generalización del estudio realizado en el Capítulo 2 para un conjunto de puntos equidistantes al caso de un conjunto de puntos unidimensional arbitrario.

Veremos que, aunque no se llega a un resultado tan preciso como el teorema (2.4.1), que describe cada coordenada principal k–ésima del conjunto de puntos con la distancia valor absoluto como un polinomio de grado k, se obtiene el teorema (4.9), que permite una interpretación cualitativa de las coordenadas principales análoga al caso equidistante.

4.1.2 Vectores propios

Consideremos un conjunto de $n+1$ puntos $U = (x_0, \ldots, x_n)$ en $\mathbf{R}$. Podemos suponer, sin pérdida de generalidad, que $x_0 < x_1 < \cdots < x_n$. Nos proponemos obtener las coordenadas principales del conjunto U con la distancia valor absoluto.

$$\boldsymbol{\delta}^2_{i\,j} = \boldsymbol{\delta}^2\,(x_i,\, x_j) = |x_i - x_j| \qquad (i,\, j = 0, \ldots,\, n)$$

Como en el caso de puntos equidistantes, como alternativa al cálculo directo por diagonalización de

$$B = H \cdot (-\frac{1}{2}\boldsymbol{\Delta}^{\,(2)}) \cdot H$$

donde H es la matriz de centrado de dimensión $(n+1,\, n+1)$, y $\boldsymbol{\Delta}^{\,(2)} = (\boldsymbol{\delta}^2_{i\,j})$ es la matriz de cuadrados de distancias, partimos de una configuración euclídea conveniente X, cuya matriz de interdistancias (euclídeas) sea igual a $\boldsymbol{\Delta}^2$, y estudiamos las componentes principales (vectores propios de la matriz de covarianzas) de esta configuración. Finalmente las coordenadas principales resultan transformando X con la matriz ortogonal de vectores propios obtenida de este cálculo.

Elegimos como configuración euclídea inicial, como en la demostración del teorema (2.2.1), la definida por la matriz de dimensión $(n+1, n)$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & \ldots & 0 \\ g_1 & 0 & 0 & 0 & \ldots & 0 \\ g_1 & g_2 & 0 & 0 & \ldots & 0 \\ g_1 & g_2 & g_3 & 0 & & 0 \\ g_1 & g_2 & g_3 & g_4 & & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ g_1 & g_2 & g_3 & \ldots & & g_n \end{pmatrix}$$

donde $g_i = \sqrt{x_i - x_{i-1}}$, $\qquad (i = 1, \ldots, n)$, que podemos escribir como

$$\begin{pmatrix} 0 & 0 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 1 & 0 & & 0 \\ 1 & 1 & 1 & 1 & & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & & 1 \end{pmatrix} \cdot G$$

donde $G = \operatorname{diag}(g_1, \ldots, g_n)$.

Multiplicando a la izquierda por la matriz de centrado H (con un factor adicional $(n+1)$) obtenemos la matriz X

$$X_{i\,j} = \begin{cases} (-(n+1) + j)\, g_j & \text{si} \quad i < j \\ j\, g_j & \text{si} \quad i \geq j \end{cases} \qquad \begin{pmatrix} i = 0, \ldots, n \\ j = 1, \ldots, n \end{pmatrix} \tag{4.1}$$

Calculamos la matriz (n, n) de covarianzas de X

$$\Sigma = \frac{1}{n+1}\, X' \cdot X = \frac{1}{n+1}\, G \cdot\, C \cdot\, G \tag{4.2}$$

donde, como en el capítulo 2,

$$\begin{aligned} C &= (n+1)\, B - b \cdot b' \\ B_{i\,j} &= \min\{i,\, j\} \\ b &= (1, \ldots, n)' \end{aligned}$$

Planteamos el problema de valores propios

$$\Sigma \cdot x = \lambda\, x$$

que equivale al problema

$$\Sigma^{-1} \cdot x = \mu\, x \qquad \text{siendo } \mu = 1/\lambda$$

Teniendo en cuenta que

$$\Sigma^{-1} = G^{-1} \cdot F(2) \cdot G^{-1}$$

(donde $F(2)$ es la matriz de Toeplitz estudiada en el Capítulo 3), y definiendo $v = G^{-1} \cdot x$, llegamos al problema de valores propios generalizado

$$F(2) \cdot v = \mu\, G^2 \cdot v \tag{4.3}$$

que podemos resolver mediante la misma técnica empleada en la demostración del teorema (3.2.1): fijado un valor propio μ, se aplica recurrencia al sistema de ecuaciones

$$\left.\begin{array}{rcrcrcl} & & 2\,v_1 & - & v_2 & = & \mu\, g_1^2\, v_1 \\ -v_1 & + & 2\,v_2 & - & v_3 & = & \mu\, g_2^2\, v_2 \\ \vdots & & \vdots & & \vdots & & \vdots \\ -v_{n-2} & + & 2\,v_{n-1} & - & v_n & = & \mu\, g_{n-1}^2\, v_{n-1} \\ -v_{n-1} & + & 2\,v_n & & & = & \mu\, g_n^2\, v_n \end{array}\right\} \tag{4.4}$$

para obtener el vector $v = (v_1, \ldots, v_n)'$. Definiendo $\xi = 1 - \mu/2$ resulta, como en (3.2.1), que

$$v_i = p_{i-1}\,(\xi)\, v_1 \tag{4.5}$$

donde los $\{p_i(\xi)\}$ son la familia de polinomios definidos por la fórmula de recurrencia

$$p_{i+2} = \left[2\, g_{i+2}^2\, \xi + 2\,(1 - g_{i+2}^2)\right]\, p_{i+1} - p_i \tag{4.6}$$

Por otro lado, calculamos el polinomio característico de (4.3)

$$\Delta_n = \det\left[F(2) - \mu\, G^2\right]$$

Desarrollando el determinante por la última fila, queda

$$\Delta_n = \left[2\, g_n^2\, \xi + 2\,(1 - g_n^2)\right]\, \Delta_{n-1} - \Delta_{n-2}$$

que es la misma relación de recurrencia de los $\{p_i(\xi)\}$. En consecuencia, los valores propios vienen dados por

$$\mu = 2\,(1 - \xi)$$

donde los ξ son ceros del polinomio p_n.

Los primeros polinomios son

$$\begin{array}{rcl} p_0 & = & 1 \\ p_1 & = & 2\, g_1^2\, \xi + 2\,(1 - g_1^2) \\ p_2 & = & (4\, g_1^2\, g_2^2)\, \xi^2 + 4\,(g_1^2 + g_2^2 - g_1^2\, g_2^2) + 3 \end{array}$$

Como se ha visto en (3.2.1), cuando los coeficientes g_i son todos iguales a 1, esta familia de polinomios es la de los polinomios U_n de Tchebytchev, ortogonales en el intervalo $[-1,\, 1]$.

En el caso general, estos polinomios no coinciden con ninguna familia de polinomios ortogonales clásicos, y de hecho no es aparente que sean ortogonales respecto a alguna función peso en algún intervalo.

En cambio, vemos a continuación, que empleando otras técnicas, podemos llegar a obtener información cualitativa satisfactoria sobre las coordenadas principales.

4.1.3 Cálculo de las coordenadas principales

En primer lugar obtenemos unas identidades útiles a partir del sistema de ecuaciones (4.4): Sumando las n ecuaciones, y empleando la notación

$$s = \sum_{k=1}^{n} v_k$$

obtenemos

$$-(s - v_n) + 2\,s - (s - v_1) = \mu \sum_{k=1}^{n} g_k^2\, v_k$$

es decir

$$v_1 + v_n = \mu \sum_{k=1}^{n} g_k^2\, v_k$$

Más en general, sumando desde la ecuación i–ésima hasta la n–ésima, y empleando la notación

$$s_i = \sum_{k=i}^{n} v_k$$

obtenemos

$$-(s_i - v_n + v_{i-1}) + 2\,s_i - (s_i - v_i)$$

de donde resulta

$$v_i - v_{i-1} + v_n = \mu \sum_{k=i}^{n} g_k^2\, v_k \qquad i = 2, \ldots, n \tag{4.7}$$

Otra identidad se obtiene multiplicando cada ecuación del sistema (4.4) por su número de orden y sumando los resultados

$$\left.\begin{array}{rcrcrcl}
 & & 2\,v_1 & - & v_2 & = & \mu\, g_1^2\, v_1 \\
-2\,v_1 & + & 4\,v_2 & - & 2\,v_3 & = & \mu\, 2\, g_2^2\, v_2 \\
-3\,v_2 & + & 6\,v_3 & - & 3\,v_4 & = & \mu\, 3\, g_3^2\, v_3 \\
\vdots & & \vdots & & \vdots & & \vdots \\
-(n-2)\,v_{n-3} & + & 2\,(n-2)\,v_{n-2} & - & (n-2)\,v_{n-1} & = & \mu\,(n-2)\, g_{n-2}^2\, v_{n-2} \\
-(n-1)\,v_{n-2} & + & 2\,(n-1)\,v_{n-1} & - & (n-1)\,v_n & = & \mu\,(n-1)\, g_{n-1}^2\, v_{n-1} \\
-n\,v_{n-1} & + & 2\,n\,v_n & & & = & \mu\, n\, g_n^2\, v_n
\end{array}\right\}$$

resultando

$$(n+1)\,v_n = \mu \sum_{k=1}^{n} k\, g_k^2\, v_k \tag{4.8}$$

Ahora, dado un vector v, solución de (4.3), recuperamos el correspondiente vector propio de Σ

$$x = G \cdot v$$

y calculamos la correspondiente coordenada principal

$$y = (y_0, y_1, \ldots, y_n)'$$

efectuando el producto

$$y = X \cdot x$$

Sustituyendo X de (4.1) obtenemos

$$y_i = \begin{cases} \displaystyle\sum_{k=1}^{n} k\, g_k^2\, v_k - (n+1) \sum_{k=i+1}^{n} g_k^2 v_k & \text{para} \quad 0 \le i \le n-1 \\ \displaystyle\sum_{k=1}^{n} k\, g_k^2\, v_k & \text{para} \quad i = n \end{cases}$$

Teniendo en cuenta las identidades (4.7) y (4.8)

$$y_i = \begin{cases} -(n+1)\, v_1/\mu & \text{para} \quad i = 0 \\ -(n+1)\,(v_{i+1} - v_i)/\mu & \text{para} \quad 1 \le i \le n-1 \\ (n+1)\, v_n/\mu & \text{para} \quad i = n \end{cases}$$

Podemos resumir estas igualdades, tomando convencionalmente $v_0 = v_{n+1} = 0$, resultando

$$y_i = (n+1)\,\lambda\,(v_i - v_{i+1}) \qquad 0 \le i \le n \tag{4.9}$$

4.1.4 Propiedades de las coordenadas principales

Despues de los cálculos realizados estamos ya en condiciones de describir cualitativamente las coordenadas principales. Partimos del siguiente

Lema 4.1.1

$$y_i - y_{i-1} = (n+1)\, g_i^2\, v_i \qquad 1 \le i \le n$$

En particular, la sucesión de primeras diferencias de la coordenada principal j–ésima tiene los mismos signos que el vector propio j–ésimo, (ordenando los vectores propios v según orden decreciente en λ).

Demostración:

$$\begin{aligned} y_i - y_{i-1} &= \\ &= (n+1)\,(v_i - v_{i+1} - v_{i-1} + v_i)/\mu \\ &= (n+1)\,(-v_{i-1} + 2\,v_i - v_{i+1})/\mu \\ &= (n+1)\,g_i^2\,v_i \end{aligned}$$

□

Para determinar los signos de los vectores propios v aplicamos el siguiente teorema (vease Gantmacher [45, vol. II, pag. 101])

TEOREMA 4.1.1 *(Gantmacher)*
1. Una matriz oscilatoria A de dimensión (n, n) tiene n valores propios positivos distintos

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$$

2. El vector propio de A correspondiente al mayor valor propio λ_1 tiene todas sus componentes no nulas de igual signo.

El vector propio de A correspondiente al segundo valor propio λ_2 tiene exactamente una variación de signo en sus componentes.

En general, el vector propio de A correspondiente al valor propio λ_k tiene exactamente $k-1$ variaciones de signo en sus componentes, $(k = 1, \ldots, n)$.

□

Este teorema es aplicable a las matrices oscilatorias, definidas por

Definición 4.1.1
1. Una matriz de dimensión (m, n) es totalmente no negativa *si todos sus menores de todos los órdenes son no negativos, y es* totalmente positiva *si todos sus menores de todos los órdenes son positivos.*
2. Una matriz cuadrada A es oscilatoria *si es totalmente no negativa, y existe un entero $q > 0$ tal que A^q es totalmente positiva*

Quizás convenga aclarar que el nombre de *matriz oscilatoria* se debe a la aparición de matrices de este tipo en los estudios de Gantmacher sobre pequeñas oscilaciones de sistemas mecánicos.

Nos proponemos mostrar que la matriz de covarianzas Σ calculada en (4.2) es un matriz oscilatoria, y a través del lema (4.1.1), describir el comportamiento de las coordenadas principales del conjunto U.

Aunque la definición propiamente dicha es más bien intratable en nuestro caso, encontramos un criterio, también debido a Gantmacher (véase [45, vol. II, pag. 100], y la referencia citada en dicha obra), que simplifica las cosas

Proposición 4.1.1 *(Gantmacher) Una matriz cuadrada A totalmente no negativa es oscilatoria si y sólo si*

a) *A es no singular (i.e. $\det A > 0$)*

b) *Todos los elementos de la diagonal principal de A, y de las dos diagonales paralelas a la diagonal principal y contiguas a ésta son no nulos (i.e. $a_{ij} > 0$ si $|i-j| \leq 1$)*

□

Las condiciones **a)** y **b)** se cumplen claramente en el caso de Σ, por lo que será suficiente probar el siguiente enunciado

Proposición 4.1.2 *Σ es totalmente no negativa*

Para demostrar este enunciado no será necesario (!) considerar individualmente los 2^{2n} menores de Σ, sino que en su lugar, estudiaremos los menores de su inversa

$$\Sigma^{-1} = G^{-1} \cdot F(2) \cdot G^{-1}$$

que pueden ser descritos fácilmente, por ser ésta tridiagonal.

Emplearemos la conocida relación siguiente entre los menores de una matriz (n, n) A y los de su inversa $B = A^{-1}$ (véase, por ejemplo, Gantmacher [45, vol. I, pag. 21]):

$$B \begin{pmatrix} i_1 & i_2 & \dots & i_p \\ k_1 & k_2 & \dots & k_p \end{pmatrix} = \tag{4.10}$$

$$= \frac{(-1)^{\sum_{\nu=1}^{p} i_\nu + \sum_{\nu=1}^{p} k_\nu} A \begin{pmatrix} k'_1 & k'_2 & \dots & k'_{n-p} \\ i'_1 & i'_2 & \dots & i'_{n-p} \end{pmatrix}}{A \begin{pmatrix} 1 & 2 & \dots & n \\ 1 & 2 & \dots & n \end{pmatrix}}$$

Donde p es un entero $(1 \leq p \leq n)$, $(i_1, \dots, i_p)$ y $(k_1, \dots, k_p)$ son dos p–índices crecientes $(1 \leq i_1 < \dots < i_p \leq n$ y $1 \leq k_1 < \dots < k_p \leq n)$, $(i'_1, \dots, i'_{n-p})$ es el $(n-p)$–índice creciente complementario de $(i_1, \dots, i_p)$, y $(k'_1, \dots, k'_{n-p})$ es el $(n-p)$–índice creciente complementario de $(k_1, \dots, k_p)$.

La notación

$$B \begin{pmatrix} i_1 & i_2 & \dots & i_p \\ k_1 & k_2 & \dots & k_p \end{pmatrix}$$

representa el menor de la matriz B formado por las filas $(i_1, \dots, i_p)$ y las columnas $(k_1, \dots, k_p)$.

En la demostración de la proposición (4.1.2) haremos uso del siguiente detalle de la fórmula (4.10): la paridad de

$$\sum_{\nu=1}^{p} i_\nu + \sum_{\nu=1}^{p} k_\nu$$

es la misma que la de

$$\sum_{\nu=1}^{n-p} i'_\nu + \sum_{\nu=1}^{n-p} k'_\nu$$

pues su suma es $n\,(n+1)$, que es par.

También necesitamos la descripción de los menores de una matriz tridiagonal, que tomamos también de Gantmacher [45, vol. II, pág. 95]

Lema 4.1.2 *(Gantmacher)*

Sea A una matriz tridiagonal, y sean $I = (i_1, \dots, i_p)$ y $K = (k_1, \dots, k_p)$ dos p–índices crecientes.

Supongamos que existe un ν $(1 \leq \nu \leq p)$ tal que $i_\nu \neq k_\nu$ y que para todos los $\sigma \neq \nu$ se verifica que $i_\sigma = k_\sigma$.

Entonces

$$A \begin{pmatrix} I \\ K \end{pmatrix} = A \begin{pmatrix} i_1 & \dots & i_{\nu-1} \\ k_1 & \dots & k_{\nu-1} \end{pmatrix} A \begin{pmatrix} i_\nu \\ k_\nu \end{pmatrix} A \begin{pmatrix} i_{\nu+1} & \dots & i_p \\ k_{\nu+1} & \dots & k_p \end{pmatrix}$$

□

A partir de este enunciado se deduce que todo menor de una matriz tridiagonal es producto de menores principales y elementos no diagonales de la matriz. Formulamos este resultado en forma ligeramente más precisa, de forma que en el caso de la matriz Σ nos permita seguir la pista del signo de cada menor.

Lema 4.1.3 *Sea A una matriz tridiagonal. Supongamos que todos los menores principales y todos los elementos de las dos diagonales inferior y superior a la diagonal principal son no nulos.*

Sean $I = (i_1, \dots, i_p)$ y $K = (k_1, \dots, k_p)$ dos p–índices crecientes.

El menor $A\begin{pmatrix} I \\ K \end{pmatrix}$ *es no nulo solamente si el intervalo* $[1,\, p] \subset \mathbf{Z}$ *es reunión disjunta de tres subconjuntos* h_0, h_+, h_- *(con alguno de ellos posiblemente vacío) tales que*

$$\begin{array}{rcll} i_\mu & = & k_\mu & \text{si} \quad \mu \in h_0 \\ i_\mu & = & k_\mu + 1 & \text{si} \quad \mu \in h_+ \\ i_\mu & = & k_\mu - 1 & \text{si} \quad \mu \in h_- \end{array}$$

y en tal caso, $A\begin{pmatrix} I \\ K \end{pmatrix}$ *es igual al producto de los menores principales correspondientes a los índices de* h_0 *por los elementos no diagonales situados en las posiciones* $(k_\mu + 1,\, k_\mu)$*, para* $\mu \in h_+$*, por los elementos no diagonales situados en las posiciones* $(k_\mu - 1,\, k_\mu)$*, para* $\mu \in h_-$*.*

□

Finalemente estamos en condiciones de demostrar la proposición (4.1.2)

Demostración: (de la proposición (4.1.2))

En primer lugar observamos que $\sum^{-1}$ es tridiagonal, con todos los elementos de la diagonal positivos y los elementos de las dos diagonales paralelas negativos

En segundo lugar veamos que todos los menores principales de $\sum^{-1}$ son positivos:

El menor principal correspondiente al p–índice creciente $K = (k_1, k_2, \ldots, k_p)$,

$$\sum{}^{-1}(K) = \sum{}^{-1}\begin{pmatrix} k_1 & k_2 & \ldots & k_p \\ k_1 & k_2 & \ldots & k_p \end{pmatrix}$$

es el producto de los menores $(G(K))^2 F(K)$, según resulta de aplicar la fórmula de Binet–Cauchy, teniendo en cuenta que G es una matriz diagonal. Por ello es suficiente considerar los menores principales $F(K)$ de la matriz de Toeplitz F (que tiene los elementos de la diagonal principal iguales a 2 y los de las diagonales contiguas iguales a -1, siendo nulos todos los demás elementos).

El menor $F(K)$ tiene una estructura muy simple: es el determinante de una matriz E de dimensión (p, p) cuyos elementos no nulos se agrupan en cajas sobre la diagonal

$$E = \operatorname{diag}(F_{d_1}, \ldots, F_{d_q})$$

donde cada F_{d_i} es una matriz F de dimensión d_i ($\sum_{i=1}^{q} d_i = p$) (Se entiende que la matriz F_1 de dimensión (1, 1) es el escalar (2)).

Las dimensiones d_i vienen determinados por la contigüidad del p–índice K: si $k_{i+1} \neq k_i + 1$, entonces los elementos $e_{i,\,i+1}$ y $e_{i+1,\,i}$ son nulos, y tenemos una caja de dimensión (1, 1). En caso contrario estos elementos son iguales a -1. Por ejemplo, si $K = (1, 2, 4, 6, 7, 8)$ se obtiene

$$E = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & & & & \\ & & 2 & & & \\ & & & 2 & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

Puesto que el determinante de F_d, según se ha calculado en el capítulo 3, es igual a $d + 1$, concluimos que todos los menores principales son positivos.

Finalmente, a partir del lema (4.1.3), y empleando la notación definida en su enunciado, un menor de Σ^{-1} es negativo si, y sólo si el número de elementos de la reunión $h_+ \cup h_-$ es impar.

Trasladando a los menores de Σ, según la fórmula (4.10), vemos que para todos los menores no nulos el signo resultante es positivo. En efecto, dado un menor

$$\Sigma^{-1} \begin{pmatrix} i_1 & i_2 & \dots & i_p \\ k_1 & k_2 & \dots & k_p \end{pmatrix}$$

el signo de la potencia de (-1) en (4.10) es positivo o negativo según la paridad de

$$\sum_{\nu=1}^{p} i_\nu + \sum_{\nu=1}^{p} k_\nu =$$

$$= \sum_{\nu \in h_0} i_\nu + \sum_{\nu \in h_0} k_\nu + \sum_{\nu \in h_+} i_\nu + \sum_{\nu \in h_+} k_\nu + \sum_{\nu \in h_-} i_\nu + \sum_{\nu \in h_-} k_\nu$$

$$= 2 \left(\sum_{\nu=1}^{p} i_\nu \right) + \sharp\,(h_+) - \sharp\,(h_-)$$

donde el símbolo $\sharp(\cdot)$ representa "número de elementos de". Esta suma tiene la paridad de $\sharp(h_+) - \sharp(h_-)$, que coincide con la de $\sharp(h_+) + \sharp(h_-)$, de modo que el factor potencia de (-1) es negativo si, y sólo si el menor es negativo, y el signo resultante es positivo en todos los casos. □

Resulta que la matriz Σ es oscilatoria, y combinando el teorema de Gantmacher con el lema (4.1.1) tenemos la descripción de las coordenadas principales a la que nos proponíamos llegar:

La primera coordenada es una función monótona (creciente o decreciente, según el signo que asignemos convencionalmente a su primera componente), por lo que puede interpretarse cualitativamente como una *dimensión lineal*.

La segunda coordenada sigue una dirección de crecimiento hasta alcanzar el cambio de signo en su vector de primeras diferencias, punto en que se invierte la dirección de crecimiento, por lo que puede interpretarse como una *dimensión cuadrática*.

Sucesivamente, la tercera coordenada presenta dos inversiones en su dirección de crecimiento, lo que permite su interpretación como una *dimensión cúbica*, y en general, la coordenada k–ésima se puede interpretar como una dimensión de grado k.

Ejemplo: Para el conjunto de puntos (1, 3, 7, 15, 31, 63, 127, 255) obtenemos las siguientes 4 primeras coordenadas principales, asimilables a dimensiones de tipo lineal, cuadrática, cúbica, y cuártica.

	Coordenada			
	1	2	3	4
Autovalor:	223.	55.6	24.5	11.4
Individuo				
1	–3.88	1.68	–1.17	–0.95
2	–3.85	1.62	–1.07	–0.78
3	–3.71	1.38	–0.71	–0.18
4	–3.30	0.71	0.26	1.16
5	–2.24	–0.84	2.01	2.20
6	0.20	–3.45	2.90	–1.87
7	5.01	–4.70	–2.91	0.46
8	11.76	3.61	0.69	–0.05
Dimensión	Lineal	Cuadrática	Cúbica	Cuártica

4.2 Extensión al caso continuo

4.2.1 Introducción

Hemos visto que la distancia valor absoluto puede ser estudiada con relativa dificultad para el caso de puntos equidistantes en el Capítulo 2 y para el caso más general en la sección precedente.

Como el análisis de Coordenadas principales es una técnica que se aplica sobre un conjunto *finito* de puntos, a los que se asocia una configuración euclídea finita, en dimensión también finita, parece que la generalización al caso numerable o contínuo es un problema de difícil solución.

Afortunadamente, algunos resultados de la teoría de estadísticos de bondad de ajuste, basados en procesos estocásticos, pueden adaptarse al problema de definir unas coordenadas principales para el caso continuo.

Anderson y Darling [2] consideran el estadístico de Cramér–von Mises

$$W_n^2 = \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 \psi(F)\, dF(x) \tag{4.11}$$

donde F_n es la función distribución empírica de una muestra $x_1, \ldots, x_n$ de una variable aleatoria X con función de distribución F, y ψ es una función peso.

El cambio de variable $u = F(x)$ traslada el problema al caso de la distribución uniforme en $[0, 1]$, con lo que puede escribirse

$$W_n^2 = n \int_0^1 [G_n(u) - u]^2\, \psi(u)\, du \tag{4.12}$$

Anderson y Darling [2] y [3] obtienen la distribución asintótica de W_n^2 partiendo del proceso estocástico

$$\mathbf{Y}_n(u) = \sqrt{n}\,[G_n(u) - u] \qquad 0 \leq u \leq 1 \tag{4.13}$$

descomponiéndolo en serie de Fourier

$$\mathbf{Y}_n \cong \sum_{j=1}^{\infty} Z_j\, f_j \tag{4.14}$$

donde los términos de la sucesión $\{Z_j\}_{j \in \mathbf{N}}$ son variables aleatorias incorrelacionadas, con $\mathbf{E}(Z_j) = 0$ y $\mathbf{V}(Z_j) = \lambda_j < \infty$, y la notación $X \cong Y$ significa "X, Y tienen igual distribución".

$\{f_j\}_{j \in \mathbf{N}}$ es una sucesión de funciones ortogonales respecto al producto interno en el espacio de funciones de cuadrado integrable $\mathcal{L}^2([0, 1])$

$$\langle f_i,\, f_j \rangle = \int_0^1 f_i(t)\, f_j(t)\, dt = \delta_{i\,j}$$

Los variables aleatorias de (4.14) se obtienen por

$$Z_j = \langle \mathbf{Y}_n, f_j \rangle \tag{4.15}$$

y si escribimos

$$\mathbf{Y}_n \cong \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, Z_j^\star\, f_j \tag{4.16}$$

donde ahora $\mathbf{V}(Z_j^\star) = 1$, por la identidad de Parseval

$$\|\mathbf{Y}_n\|^2 = \int_0^1 \mathbf{Y}_n(t)^2\, dt = \sum_{j=1}^{\infty} \lambda_j\, Z_j^{\star 2} = W_n^2 \tag{4.17}$$

Así, W_n^2 se puede escribir como la suma, ponderada por los λ_j, de las variables incorrelacionadas $Z_j^\star$, de media 0 y varianza 1.

Un caso concreto de la expansión (4.14) ha sido estudiada por Durbin y Knott [38] utilizando la base ortonormal

$$f_j(t) = \sqrt{2}\, \sin(j\, \pi\, t) \qquad 0 \leq t \leq 1 \tag{4.18}$$

obteniendo la siguiente descomposición

$$\mathbf{Y}_n = \sum_{j=1}^{\infty} \frac{f_j}{j\, \pi}\, Z_j \tag{4.19}$$

donde las variables Z_j son normales independientes $N(0,1)$, que se obtienen aplicando (4.15). El estadístico de Cramér–von Mises puede escribirse como

$$W_n^2 = \sum_{j=1}^{\infty} \frac{Z_j^2}{j^2\, \pi^2} \tag{4.20}$$

Los autores estudian esta descomposición interpretando $Z_1,\ Z_2, \ldots$ como componentes principales.

Un proceso estocástico de la forma

$$\sqrt{n}\, [F_n(x) - F(x)] \qquad -\infty < x < +\infty \tag{4.21}$$

donde $F_n(x)$ es la función distribución empírica de una muestra $x_1, \ldots, x_n$, y $F(x)$ es una función de distribución (continua) recibe el nombre de *proceso empírico*.

La teoría de procesos empíricos tiene múltiples aplicaciones en Estadística (pruebas de bondad de ajuste, estadísticos de rangos, estadísticos con norma L^1, estadísticos no paramétricos, por citar algunos ejemplos). Una extensa exposición del tema puede consultarse en Shorack y Wellner [98]. Utilizaremos a continuación algunas técnicas descritas en esta obra.

4.2.2 Coordenadas Principales de la distribución uniforme

Vamos representar los valores de una variable aleatoria U con distribución uniforme en $[0,1]$ mediante una sucesión de valores de variables aleatorias, que según veremos, pueden recibir apropiadamente el nombre de *Coordenadas Principales* respecto la distancia valor absoluto, de modo análogo al estudio realizado en el capítulo 2 con un conjunto finito de puntos equidistantes.

Introducimos el proceso estocástico

$$\mathbf{U} = \{\mathbf{U}_t; 0 \leq t \leq 1\}$$

donde para cada t, la variable aleatoria $\mathbf{U}_t$, indicador del intervalo $[t,1] \subset [0,1]$, sigue la distribución de Bernoulli con parámetro t

$$\mathbf{U}_t(x) = \begin{cases} 0 & \text{si} \quad x < t \quad \text{con probabilidad} \quad t \\ 1 & \text{si} \quad x \geq t \quad \text{con probabilidad} \quad 1-t \end{cases}$$

Como justificación heurística de esta definición, se puede imaginar $\mathbf{U}$ como una matriz *continua* análoga a la matriz

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 1 & 0 & & 0 & 0 \\ 1 & 1 & 1 & 1 & & 0 & 0 \\ \vdots & \vdots & & & \ddots & & \vdots \\ 1 & 1 & 1 & & & 1 & 0 \\ 1 & 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix}$$

que era nuestro punto de partida para el estudio de la distancia valor absoluto en el caso de un conjunto de $n+1$ puntos equidistantes.

Proposición 4.2.1 *Se puede reconstruir la variable U a partir de $\mathbf{U}$:*

$$U = \int_0^1 \mathbf{U}_t \, dt \tag{4.22}$$

Demostración: Para un $x \in [0,1]$ dado, la función $\mathbf{U}_t(x)$ de $t \in [0,1]$ es la función indicadora del intervalo $[0,x]$, de modo que

$$\int_0^1 \mathbf{U}_t(x) \, dt = x$$

□

Puede interpretarse (4.22) como una descomposición *continua* de la variable U en suma de indicadores, de modo que a cada valor x de la variable aleatoria U le corresponde la *fila* (trayectoria) x_t de $\mathbf{U}$.

Proposición 4.2.2 *La distancia euclídea entre dos trayectorias x_t, y_t de $\mathbf{U}$, definida como*

$$\int_0^1 (x_t - y_t)^2 \, dt$$

es igual a la distancia valor absoluto $|x-y|$ entre los correspondientes puntos en $[0, 1]$.

Demostración: Supongamos, por ejemplo, $x < y$. Entonces

$$\int_0^1 (x_t - y_t)^2 \, dt = \int_x^y 1^2 \, dt = y - x$$

□

La analogía con el caso discreto permite pensar $\mathbf{U}$ como la configuración euclídea convencional cuya matriz de distancias coincide con la de los puntos dados con la función distancia valor absoluto, que nos servía como punto de partida para el cálculo de las coordenadas principales.

Seguiremos a continuación un programa paralelo, aprovechando las técnicas descritas en Shorack y Wellner [98, pág. 205] para calcular una descomposición numerable del proceso $\mathbf{U}$.

$$\mathbf{U} \cong \sum_{j=1}^{\infty} Z_j \, f_j \tag{4.23}$$

donde las Z_j son variables aleatorias con varianza 1, y las f_j son un sistema completo de funciones ortonormales en $\mathcal{L}^2([0, 1])$. (Por comodidad de cálculo, no centramos el proceso, sino que aplicaremos esta operación al final, sobre las Z_j resultantes).

Proposición 4.2.3 *La función de covarianza para $\mathbf{U}$ es*

$$K(s, t) = \mathrm{Cov}\,(\mathbf{U}_s, \mathbf{U}_t) = \min(s, \, t) - s\, t \qquad 0 \leq s, \, t \leq 1 \tag{4.24}$$

Demostración:

$$\begin{aligned} K(s, t) &= \mathbf{E}\,(\mathbf{U}_s\, \mathbf{U}_t) - \mathbf{E}(\mathbf{U}_s)\, \mathbf{E}(\mathbf{U}_t) \\ &= (1 - \max\{s, \, t\}) - (1 - s)\,(1 - t) \\ &= s + t - \max\{s, \, t\} - s\, t \end{aligned}$$

□

Este núcleo es bien conocido, pues aparece como función de covarianza del *puente Browniano*, y también en Mecánica, como función de Green del problema de Sturm–Liouville de la cuerda vibrante con extremos fijos.

Es un núcleo continuo, simétrico y definido positivo, por lo que, según el teorema de Mercer (véase, por ejemplo, Courant and Hilbert [15, vol. I chap 3]), la descomposición

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j \, f_j(s) \, f_j(t)$$

converge absoluta y uniformemente en s y t, siendo los λ_j y f_j los valores propios y funciones propias del núcleo K, es decir, tales que

$$\int_0^1 f_j(s) \, K(s,t) \, dt = \lambda_j \, f_j(t)$$

Las f_j (normalizadas a 1) forman un sistema ortonormal completo en $\mathcal{L}^2([0,1])$.

Encontramos esta descomposición calculada en Shorack and Wellner [98, pág. 214]

$$\begin{aligned} \lambda_j &= \frac{1}{(j\,\pi)^2} \\ f_j(t) &= \sqrt{2}\,\sin(j\,\pi\,t) \qquad 0 \le t \le 1 \\ & \qquad (j \in \mathbf{N}) \end{aligned} \tag{4.25}$$

Ahora, recordando que en el caso finito se obtiene cada eje principal como producto de la matriz euclídea inicial por el correspondiente vector propio de la matriz de covarianzas, la operación en el caso presente es

$$Z_j = \int_0^1 \mathbf{U}_t \, f_j(t) \, dt$$

Esta operación coincide con el cálculo de las *componentes principales* del proceso $\mathbf{U}$, en la nomenclatura de Shorack y Wellner, y puesto que la traza

$$\int_0^1 K(t,t) \, dt = \int_0^1 (t - t^2) \, dt = \frac{1}{6} < \infty$$

y valen las hipótesis del teorema de Mercer para el núcleo $K(s,t)$ es aplicable el teorema de Kac y Siegert (véase Shorack y Wellner [98, pág. 210]) que permite afirmar la descomposición (4.23).

Proposición 4.2.4
1. Las variables aleatorias Z_j de la fórmula (4.23) vienen dadas en función de U por

$$Z_j = \frac{\sqrt{2}}{j\,\pi}\,(1-\cos(j\,\pi\,U)) \tag{4.26}$$

2. Las variables Z_j tienen momentos de todos los órdenes, y en particular

$$\mathbf{E}\,(Z_j) = \frac{\sqrt{2}}{j\,\pi} \qquad \mathbf{V}\,(Z_j) = \frac{1}{j^2\,\pi^2}$$

Demostración:

1. Dado $x \in [0,1]$

$$\begin{aligned} Z_j(x) &= \int_0^1 \mathbf{U}_t(x)\, f_j(t)\, dt \\ &= \int_0^x f_j(t)\, dt \\ &= \int_0^x \sqrt{2}\,\sin(j\,\pi\,t)\, dt \\ &= \frac{\sqrt{2}}{j\,\pi}\,(1-\cos(j\,\pi\,x)) \end{aligned}$$

2. Se calcula la esperanza de Z_j a partir de la relación funcional con la variable U

$$\mathbf{E}\,(Z_j) = \int_0^1 \frac{\sqrt{2}}{j\,\pi}\,(1-\cos(j\,\pi\,x))\, dx = \frac{\sqrt{2}}{j\,\pi}$$

Los momentos centrales de orden superior se calculan fácilmente de igual forma. □

Designaremos C_j la variable aleatoria centrada correspondiente a Z_j

$$C_j = -\frac{\sqrt{2}}{j\,\pi}\,\cos(j\,\pi\,U) \tag{4.27}$$

y $C_j^\star$ la variable tipificada.

$$C_j^\star = -\sqrt{2}\,\cos(j\,\pi\,U) \tag{4.28}$$

El siguiente teorema, objetivo de este apartado, es la que justifica, por analogía con la definición usual en Estadística Multivariante (véase el teorema (1.5.1) del Capítulo 1), denominar a las C_j *Ejes principales* para la variable U respecto la distancia Valor Absoluto, y *Coordenadas principales* de un $x \in [0,1]$ en esta representación a la sucesión de valores

$$\{C_j(x)\}_{j\in\mathbf{N}} = (-\sqrt{2}\cos(\pi\, x), -\sqrt{2}\cos(2\,\pi\, x), -\sqrt{2}\cos(3\,\pi\, x), \dots) \tag{4.29}$$

Compárese, además, esta expresión con la obtenida en el teorema (2.4.1) del capítulo 2, donde para el caso discreto equidistante, se llega a idéntico resultado, dado que

$$C_j(x) = -\sqrt{2}\, T_j\,(z)$$

siendo T_j el j–ésimo polinomio de Tchebychev de primera especie, y $z = \cos(\pi\, x)$.

TEOREMA 4.2.1

1. Las variables C_j son incorrelacionadas.

2. Las variables tipificadas $C_j^\star$ son igualmente distribuidas, siendo su función de distribución común (para todo $j \in \mathbf{N}$)

$$G(x) = \begin{cases} 0 & \text{si} \quad x < -\sqrt{2} \\ 1 - \dfrac{1}{\pi}\arccos\left(\dfrac{x}{\sqrt{2}}\right) & \text{si} \quad \sqrt{2} \le x < -\sqrt{2} \\ 1 & \text{si} \quad \sqrt{2} \le x \end{cases} \tag{4.30}$$

Son absolutamente continuas y su función densidad de probabilidad es

$$g(x) = \begin{cases} \dfrac{1}{\pi}\dfrac{1}{\sqrt{2-x^2}} & \text{si} \qquad \sqrt{2} < x < -\sqrt{2} \\ 0 & \text{en caso contrario} \end{cases} \tag{4.31}$$

3. La sucesión de varianzas $\{\mathbf{V}(C_j)\}_{j\in\mathbf{N}}$ es decreciente y sumable, siendo su suma igual a "la variabilidad total de la matriz de covarianzas", es decir, a la traza del núcleo $K(s,t)$

4. Dados x, $y \in [0,1]$, la distancia euclídea al cuadrado entre las sucesiones $\{C_j(x)\}_{j\in\mathbf{N}}$ y $\{C_j(y)\}_{j\in\mathbf{N}}$ es igual a la distancia valor absoluto $|x-y|$

$$\sum_{j=1}^{\infty}\left(C_j(x) - C_j(y)\right)^2 = |x-y| \tag{4.32}$$

Demostración:

1. Esta afirmación puede deducirse del teorema de Kac y Siegert mencionado más arriba, o bien obtenerse facilmente por cálculo directo.

2. Dado $x \in [-\sqrt{2}, +\sqrt{2}]$, calculamos la probabilidad $P(C_j^\star \leq x)$

La función $C_j^\star(t) = -\sqrt{2}\cos(j\,\pi\,t)$ es inyectiva en el intervalo $[0, 1/j]$, y el conjunto de los puntos $t \in [0, 1/j]$ para los que $C_j^\star(t) \leq x$ es el intervalo

$$I_1 = [0, z]$$

siendo

$$z = \frac{1}{j} - \frac{1}{j\,\pi}\arccos(\frac{x}{\sqrt{2}})$$

En general, en cada uno de los j intervalos

$$\left[\frac{i}{j}, \frac{i+1}{j}\right] \qquad 0 \leq i \leq j-1$$

el conjunto de los puntos que verifican la condición especificada es un intervalo I_i de igual longitud que I_1.

Por ejemplo, si $j \geq 2$ tenemos $I_2 = [2/j - z, 2/j]$, si $j \geq 3$, tenemos $I_3 = [2/j, 2/j + z]$, etc.

La probabilidad de la reunión de estos intervalos por la ley uniforme es la longitud total, igual a $j\,z$.

3. Hemos visto ya que la traza de K es igual a $1/6$. Ahora verificamos que

$$\sum_{j=1}^{\infty} \mathbf{V}(C_j) = \frac{1}{\pi^2}\sum_{j=1}^{\infty}\frac{1}{j^2} = \frac{1}{6}$$

4. El enunciado equivale a verificar la identidad

$$|x-y| = \frac{2}{\pi^2}\sum_{j=1}^{\infty}\frac{(\cos(j\,\pi\,x) - \cos(j\,\pi\,y))^2}{j^2}$$

que se obtiene calculando el desarrollo de la función $|x-y|$ en serie doble de Fourier en el cuadrado $[0,1]\times[0,1]$

$$|x-y| \quad = \quad \frac{1}{4}A_{0\,0} \; +$$

$$+ \quad \frac{1}{2}\sum_{m=1}^{\infty} A_{m\,0}\cos\, m\,\pi\, x + \frac{1}{2}\sum_{n=1}^{\infty} A_{0\,n}\cos\, n\,\pi\, y \; +$$

$$+ \quad \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} A_{m\,n}\,\cos\, m\,\pi\, x\;\cos\, n\,\pi\, y \tag{4.33}$$

donde los coeficientes $A_{m\,n}$, para $m,\, n \geq 0$ se calculan por

$$A_{m\,n} = 4\int_0^1\int_0^1 |x-y|\,\cos\, m\,\pi\, x\;\cos\, n\,\pi\, y\; dx\, dy$$

y son

$$A_{m\,n} = \begin{cases} \dfrac{4}{3} & \text{si} \quad m=0 \quad \text{y} \quad n=0 \\ \dfrac{4}{(m\,\pi)^2}\left[1+(-1)^m\right] & \text{si} \quad m>0 \quad \text{y} \quad n=0 \\ \dfrac{4}{(n\,\pi)^2}\left[1+(-1)^n\right] & \text{si} \quad m=0 \quad \text{y} \quad n>0 \\ -\dfrac{4}{(n\,\pi)^2}\,\delta_{m\,n} & \text{si} \quad m>0 \quad \text{y} \quad n>0 \end{cases}$$

Sustituyendo en (4.33), teniendo en cuenta que en los sumandos $(m,0)$ y $(0,n)$ los términos no nulos son los pares, empleando la identidad $\cos 2\,a = 2\cos^2 a - 1$ y la identidad $\displaystyle\sum_{m=1}^{\infty} 1/n^2 = \pi^2/6$ obtenemos

$$|x-y| \;=\; \frac{1}{3} + \frac{1}{2}\left(\frac{4}{\pi^2}\sum_{n=1}^{\infty}\frac{\cos^2\, n\,\pi\, x}{n^2} - \frac{1}{3}\right) +$$

$$+\frac{1}{2}\left(\frac{4}{\pi^2}\sum_{n=1}^{\infty}\frac{\cos^2\, n\,\pi\, y}{n^2} - \frac{1}{3}\right) -$$

$$-\frac{4}{\pi^2}\sum_{n=1}^{\infty}\frac{\cos\, n\,\pi\, x\ \cos\, n\,\pi\, y}{n^2}$$

$$=\ \frac{2}{\pi^2}\sum_{n=1}^{\infty}\frac{(\cos\, n\,\pi\, x-\cos\, n\,\pi\, y)^2}{n^2}$$

□

4.2.3 Regresión basada en distancias para variables aleatorias

Podemos plantear ahora el modelo de regresión basada en distancias para una variable aleatoria Y, que podemos suponer tipificada, sobre U (es decir, sobre la variable resultante de tipificar U)

Se trata de calcular, para un n dado, los coeficientes β_j en

$$Y=\sum_{j=1}^{n}\beta_j\, C_j^{\star}+e$$

que hagan máximo el porcentaje de variabilidad de Y explicado por las *variables regresoras* $C_j^{\star}$.

Se obtienen estos coeficientes por

$$\beta_j=\mathbf{E}\,(Y\, C_j^{\star})=-\int y\,\sqrt{2}\,\cos(j\,\pi\, x)\,dH \tag{4.34}$$

siendo H la distribución conjunta de (Y,U).

Al ser incorrelacionadas las variables regresoras, tenemos que el coeficiente de determinación es

$$R^2=\sum_{j=1}^{n}\beta_j^2$$

Aunque en general la distribución conjunta H es inaccesible, podemos utilizar este modelo para proponer un test de bondad de ajuste de la distribución de una Y a la distribución uniforme, y, por medio de un cambio de variable como en (4.12), de una distribución a otra.

4.2.4 Aplicación al estudio de bondad de ajuste

La distribución bivariante que proporciona máxima correlación entre sus marginales X, Y, con distribuciones F y G, respectivamente, es la cota de Fréchet

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

y dos variables aleatorias X, Y cuya distribución conjunta sea H^+ se relacionan por

$$F(X) = G(Y) \tag{4.35}$$

Suponiendo X, Y tipificadas, esta máxima correlación viene dada por ([56], [44])

$$\rho^+ = \int_0^1 F^{-1}(p)\, G^{-1}(p)\, dp \tag{4.36}$$

y puede interpretarse como una medida de ajuste entre F y G, y su cálculo para algunas distribuciones conocidas ha sido llevado a cabo por Cuadras y Fortiana [30]

Ahora, si una de las variables es la U, uniforme en $[0, 1]$, y la distribución de la otra es F, tomando como distribución conjunta la cota de Fréchet H^+ y aplicando la relación funcional (4.35) podremos calcular los coeficientes β_j de (4.34) por

$$\beta_j = -\int_0^1 F^{-1}(x)\, \sqrt{2}\, \cos(j\, \pi\, x)\, dx \tag{4.37}$$

Obsérvese que estos coeficientes son los del desarrollo de Fourier de la función F^{-1} en el intervalo $[0, 1]$ respecto al sistema ortonormal completo de funciones

$$\left\{ -\sqrt{2}\, \cos(j\, \pi\, x) \right\}_{j \in \mathbf{N}} \qquad x \in [0, 1]$$

y que la función F^{-1} es de cuadrado integrable siempre que la variable $X = F^{-1}(U)$ tenga segundo momento finito, puesto que

$$\int x^2\, dF(x) = \int_0^1 (F^{-1}(x))^2\, dx$$

condición que permite asegurar la convergencia de las integrales impropias (4.37).

Un caso particular importante es el estudio de una distribución empírica. Supongamos que $x_1 \leq \ldots \leq x_n$ es una muestra ordenada de una variable aleatoria X, y que se desea contrastar la hipótesis nula

$$H_0 : \quad \text{La distribución de } X \text{ es igual a } F$$

para una función de distribución F dada.

Si se cumple la hipótesis nula, la sucesión $y_1 \leq \ldots y_n$, donde $y_i = F(x_i)$, será una muestra de una distribución uniforme en $[0, 1]$.

Si G_n es la función de distribución empírica de los y_i, la discrepancia entre esta distribución y la uniforme puede proporcionar un criterio para decidir sobre H_0.

Los coeficientes β_j para este caso darán medidas de esta discrepancia según los sucesivos ejes principales, y se les puede dar una interpretación parecida a la que proponen Durbin y Knott [38], y Durbin, Knott y Taylor [39].

Puesto que en este caso

$$G_n^{-1}(t) = y_i \qquad \text{si } t_i < t \leq t_{i+1} \qquad (0 \leq i \leq n-1)$$

siendo $t_0 = 0$, y $t_i = i/n$ para $i > 0$, (suponiendo para simplificar la notación que los y_i no tienen repeticiones), la fórmula (4.37) se reduce a la suma finita

$$\beta_j = \frac{\sqrt{2}}{j\,\pi} \sum_{i=1}^{n} y_i \left(\sin \frac{(i-1)\,j\,\pi}{n} - \sin \frac{i\,j\,\pi}{n} \right)$$

Ejemplo 4.2.1 *Coeficientes teóricos para una distribución de probabilidad dada*

Si U_0 es la variable uniforme centrada $U_0 = \sqrt{3}\,(2\,U - 1)$, cuya función distribución de probabilidad es

$$F(x) = \begin{cases} 0 & \text{si} \quad x < -\sqrt{3} \\ \dfrac{1}{2\sqrt{3}}\,x + \dfrac{1}{2} & \text{si} \quad -\sqrt{3} \leq x < \sqrt{3} \\ 1 & \text{si} \quad \sqrt{3} \leq x \end{cases}$$

se pueden calcular explícitamente los coeficientes β_j, obteniéndose

$$\beta_j = \begin{cases} \dfrac{4\sqrt{6}}{j^2\,\pi^2} & \text{si } j \text{ es impar} \\ 0 & \text{si } j \text{ es par} \end{cases} \tag{4.38}$$

En general no es posible obtener, como en este caso, fórmulas explícitas para los coeficientes de Fourier β_j, pero siempre se puede recurrir a la integración numérica.

En la siguiente tabla se citan los primeros coeficientes β_j para algunos ejemplos de distribuciones conocidas. Se han tipificado las distribuciones para el cálculo, de forma que, por ejemplo, hay una sola fila para la distribución Normal, mientras que en otras distribuciones al tipificar no desa-

parecen todos los parámetros.

	β_1	β_2	β_3	β_4
Beta (2,2)	0.9786	0	0.1759	0
Beta (1,2)	0.9642	−0.1639	0.1634	−0.0655
Exponencial	0.8336	−0.3192	0.2513	−0.1679
$t\,(3)$	0.7822	0	0.3066	0
$t\,(20)$	0.9389	0	0.2530	0
Normal	0.9484	0	0.2407	0
F $(2,8)$	0.6604	−0.3366	0.2697	−0.2061
F $(8,5)$	0.4804	−0.2547	0.2249	−0.1730

Ejemplo 4.2.2 *Estudio de una distribución empírica*

Mediante muestreo artificial se obtuvo la siguiente muestra de tamaño $n = 20$

0.0162	0.0210	0.0614	0.0926	0.1088
0.1395	0.1711	0.2078	0.4481	0.4691
0.5119	0.6204	0.6679	0.7111	0.7842
0.7917	0.8531	0.8896	0.9661	0.9783

Se plantean dos hipótesis. Según la primera hipótesis, la muestra procede de una distribución uniforme (como es en realidad), y según la segunda hipótesis, la muestra procede de una distribución exponencial. Si es cierta esta segunda hipótesis, la transformación $y = 1 - \exp(-\widehat{\alpha}\, x)$ debería transformar la muestra en una uniforme, siendo $\widehat{\alpha}$ la estimación máximo verosímil del parámetro α.

El valor de ρ^+ calculado para la muestra original es

$$\rho_1^+ = 0.98559$$

y para la muestra transformada es

$$\rho_2^+ = 0.978477$$

Al ser $\rho_1^+ > \rho_2^+$, la hipótesis de distribución uniforme prevalece sobre la de distribución exponencial. Si efectuamos el cálculo de los coeficientes β_j, apreciamos una mejor resolución al comparar los valores calculados en cada caso con los teóricos para la distribución uniforme.

	Datos Originales	Datos Transformados	Coeficientes Teoricos
β_1	0.9930	0.9714	0.9927
β_2	−0.0040	0.2005	0
β_3	−0.0167	−0.0127	0.1103
β_4	−0.0573	−0.0481	0

Chapter 5

Análisis discriminante basado en distancias

5.1 Introducción

En este capítulo se desarrolla el problema del Análisis Discriminante desde el punto de vista de las distancias, en el sentido ya introducido en el Capítulo 1, sección 6.3.

Sea ω un individuo a clasificar en una de dos poblaciones posibles π_1, π_2 sobre la base de p variables $x_1, \ldots, x_p$, que pueden ser de cualquier tipo: contínuas, discretas, binarias, categóricas.

Como es bien sabido, el problema de asignar un individuo ω a una población π_i se resuelve mediante una función discriminante $f(x)$, donde x representa el vector de observaciones de las variables sobre ω. La regla general es

$$\text{Clasificar } x \text{ en } \begin{cases} \pi_1 & \text{si} \quad f(x) > 0 \\ \pi_2 & \text{si} \quad f(x) \leq 0 \end{cases} \tag{5.1}$$

El tema del Análisis Discriminante está ampliamente tratado en la literatura: Kendall [64], Anderson [4], Lachenbruch [71], Morrison [80], Mardia et al. [78], Seber [96], Krzanowski [70], Cuadras [20] , [24], y [28].

5.1.1 Notaciones

Utilizaremos las siguientes notaciones:

LDF para referirnos a la función discriminante lineal de Fisher [42]

$$L(x) = [x - \frac{1}{2}(\overline{x}_1 + \overline{x}_2)]' \cdot S^{-1} \cdot (\overline{x}_1 - \overline{x}_2) \tag{5.2}$$

donde S es la matriz de covarianzas común de las dos muestras (*pooled within groups*).

QDF para referirnos al discriminador cuadrático

$$\begin{aligned} Q(x) \quad = \quad & \frac{1}{2}\log\frac{|S_2|}{|S_1|} - \frac{1}{2}(\overline{x}_1' \cdot S_1 \cdot \overline{x}_1 - \overline{x}_2' \cdot S_1 \cdot \overline{x}_2) + \\ & + x' \cdot (S_1^{-1} \cdot \overline{x}_1 - S_2^{-1} \cdot \overline{x}_2) - \frac{1}{2} x' \cdot (S_1^{-1} - S_2^{-1}) \cdot x \end{aligned} \tag{5.3}$$

donde $\overline{x}_1$, $\overline{x}_2$ son las medias, y S_1 y S_2 son las matrices de covarianzas de muestras de tamaños n_1 y n_2, procedentes de las poblaciones π_1 y π_2, respectivamente. Obsérvese que QDF se convierte en LDF en el caso de igual matriz de covarianzas en las dos poblaciones (i.e. si se reemplaza S_1 y S_2 por S en 5.3).

LM para referirnos al Análisis Discriminante basado en el *location model*, El LM, studiado por Krzanowski [67], es aplicable cuando los vectores x de

observaciones de las variables constan de una componente x_b con k variables discretas binarias y una componente x_c con variables continuas normales.

Este modelo consiste en proponer para las variables continuas una distribución normal multivariante para cada una de las 2^k configuraciones posibles de las k variables binarias, con media (posiblemente) distinta, y matriz de covarianzas común. Se emplean las funciones discriminantes

$$[x - \frac{1}{2}(\mu_1{}^{(m)} + \mu_2{}^{(m)})] \cdot S^{-1} \cdot (\mu_1{}^{(m)} - \mu_2{}^{(m)})' - \log \frac{p_{2\,m}}{p_{1\,m}} \tag{5.4}$$

donde m recorre el conjunto de las 2^k configuraciones de las variables binarias, $\mu_i{}^{(m)}$ es la media de las variables continuas x_c para la población i $(i = 1, 2)$, en la configuración m, S es la matriz de covarianzas común, y $p_{i\,m}$ son las probabilidades de tener una observación en π_i $(i = 1, 2)$ en la configuración m. Los valores de $p_{i\,m}$ y $\mu_i{}^{(m)}$ se estiman empleando un modelo de regresión y un modelo log–lineal, respectivamente.

ML para referirnos a la regla de la máxima verosimilitud, basada en la función discriminante

$$V(x) = \log f_1(x) - \log f_2(x) \tag{5.5}$$

donde f_i es la densidad de probabilidad de x cuando se sabe que el individuo ω pertenece a π_i.

BR (regla de Bayes) se aplica cuando las probabilidades *a priori* q_1, q_2 de que ω pertenezca a π_1 o π_2 son conocidas, y se basa en la función discriminante

$$B(x) = V(x) + \log(q_1) - \log(q_2) \tag{5.6}$$

Según un resultado clásico, BR es admisible, es decir, no existe una regla de decisión mejor cuando las f_i y q_i son conocidas (Anderson [4, p. 144], Mardia et al. [78, p. 308].

Nos referiremos a la regla M cuando se asigna el individuo ω de coordenadas x_0 a la población más próxima, suponiendo conocidas las distancias $\boldsymbol{\delta}(x_0, \pi_i)$, $(i = 1, 2)$. Esta regla se considera introducida por Matusita ([75], [76], [77]).

Indicaremos por DB el método basado en una distancia entre individuos u observaciones. Introducido pos Cuadras [24], desarrollamos diversos aspectos y estudiamos sus propiedades en la sección siguiente.

Finalmente, haremos referencia a LR (Regresión logística), NN (*Nearest Neighbour*), cuyos principios están bien descritos en Lachenbruch [71] y Seber [96], y al llamado discriminador lineal euclídeo EDF (Marco et al. [73]), que está basado en una función discriminante análoga a la (5.2), pero emplando la métrica euclídea standard en lugar de la métrica de Mahalanobis.

5.1.2 Consideraciones sobre los métodos clásicos y el DB

En los casos en que las variables discriminadoras son continuas y se pueden suponer con distribución conjunta normal multivariante, son adecuados los métodos LDF y QDF, siendo aplicable el primero cuando se puede aceptar la hipótesis de homoscedasticidad. En dicho caso, LDF coincide con DB si se aplica como función distancia la dada por la métrica de Mahalanobis.

Cuando la normalidad no es aceptable, LDF es más robusto que QDF (Véase Seber [96, pp. 297–300] para una evaluación comparativa detallada). EDF es un buena elección si las variables son continuas y su número es grande en comparación con las dimensiones de las muestras.

Si el conjunto de variables discriminadoras contiene variables contínuas, binarias y cualitativas, es adecuado LM si se puede aceptar normalidad de las variables contínuas para cada configuración. Presenta el inconveniente de requerir que la mayoría de celdas (correspondientes a cada configuración de las variables discretas) no sean vacías y de requerir enormes recursos computacionales cuando las variables cualitativas presentan muchos estados, ya que debe desdoblarse cada una de estas variables en el número necesario de variables binarias, y el número de cálculos a realizar es de orden exponencial en el número total de variables binarias resultante.

LR permite tambien discriminación con variables mixtas, y es preferible a LDF en estos casos. (Véase Efron [40] y Press and Wilson [87]). Los argumentos en favor de este método se basan en que si los parámetros del modelo se ajustan por máxima verosimilitud, como es el caso más frecuente, proporciona una medida del ajuste del modelo a los datos, da una medida fiable de la significación de cada variable para la clasificación, y en general, tiene las ventajas propias de dicha técnica, notablemente la consistencia, mientras que LDF no mejora necesariamente las predicciones cuando crece el tamaño de la muestra. Obsérvese que LDF se basa en la regla de máxima verosimilitud sólo si se cumplen las hipótesis de continuidad, normalidad y homoscedasticidad para las variables discriminadoras.

LR presenta también el problema mencionado en el caso del LM de requerir un desdoblamiento interno de cada variable discreta en variables binarias, aunque en este caso el crecimiento de recursos computacionales con el número de variables es solamente polinómico, en lugar de exponencial, como en el LM. También el algoritmo IRLS (Iteratively Reweighted Least Squares) que se emplea usualmente en la estimación de parámetros (e.g. en el PROC LOGISTIC de SAS) puede presentar problemas de inestabilidad numérica (Véase Green [54]).

El método DB se basa en el hecho que en muchas ocasiones, el suponer que las variables x siguen una distribución de probabilidad $F(x)$ dada, es una hipótesis inadecuada o indemostrable, mientras que resulta natural asumir la

existencia de una función distancia $\boldsymbol{\delta}(\omega, \omega')$ que cuantifique el conocimiento que se tiene en cada problema concreto de clasificación de la *proximidad* o *similitud* entre dos observaciones ω, ω'.

Por ello, el método DB es todavía apropiado en problemas como la comparación de códigos (e.g. secuencias genéticas en DNA), o el reconocimiento de manuscritos o de voz, casos en los cuales incluso el suponer la existencia de una distribución de probabilidad para las variables del problema es una hipótesis arriesgada. Véase Valdés [100] y referencias citadas en dicho artículo.

DB es más robusto que los métodos del tipo NN, pues, como se muestra en la sección siguiente, la regla de decisión propuesta para clasificar un individuo ω equivale a minimizar la distancia de ω al individuo medio de cada población π_i en un espacio pseudo–euclídeo, lo que hace que sea menos sensible a la presencia de *outliers* que dichos métodos.

Adicionalmente, permite una estimación fácil de la tasa de error, y en caso de conocerse probabilidades de asignación *a priori*, pueden incorporarse al modelo.

Por último, los requerimientos computacionales de DB son sensiblemente menores que LM y LR. El número de *flops* crece cuadráticamente con el tamaño de la muestra, linealmente con el número de variables y, empleando las distancias usuales, no aumenta con el número de estados de las variables cualitativas, al no existir desdoblamientos internos en variables binarias. El tamaño de la memoria (real o virtual) necesaria crece sólo linealmente con el tamaño de la muestra (Véase implementación del algoritmo en el capítulo 6 de esta memoria).

5.2 El método DB de clasificación

5.2.1 Método DB para muestras

Como se ha mencionado en 1.6.3, si se dispone de un total de $n = n_1 + n_2$ observaciones, siendo n_k de π_k, $(k = 1, 2)$, las funciones discriminantes propuestas por Cuadras [24] para asignar una nueva observación ω son

$$\phi_k(\omega) = \frac{1}{n_k} \sum_{i=1}^{n_k} {\boldsymbol{\delta}_i}^2(k) - \frac{1}{{n_k}^2} \sum_{i=1}^{n_k} \sum_{j=i}^{n_k} {\boldsymbol{\delta}_{ij}}^2(k) \tag{5.7}$$

donde $\boldsymbol{\Delta}^{(2)}(k) = \left(\boldsymbol{\delta}^2_{ij}(k)\right)$ es la matriz de cuadrados de distancias de la subpoblación π_k, y $\boldsymbol{\delta}^2_i(k)$, $(i = 1, \ldots, n_k)$ los cuadrados de las distancias de ω a los n_k individuos de esta subpoblación.

Se asigna ω a la subpoblación k para la cual $\phi_k(\omega)$ es mínima.

Considerando la representación euclídea (o pseudo–euclídea) asociada a la configuración de interdistancias entre los n_k indivíduos de la muestra de π_k, según los teoremas 1.5.1 y 1.5.2 del capítulo 1, comprobamos a continuación que esta regla de asignación corresponde a un criterio de mínima distancia.

Sea $H(k)$ la matriz de centrado de dimensión (n_k, n_k), $B(k)$ y $X(k)$ como en los teoremas citados

$$B(k) = H(k) \cdot \left(-\frac{1}{2}\boldsymbol{\Delta}^{(2)}(k)\right) \cdot H(k) = X(k) \cdot X(k)'$$

Las filas $x_i(k)$ de $X(k)$ contienen los vectores que representan a los n_k individuos de la muestra de π_k en un espacio $\mathbf{R}^p \times i\,\mathbf{R}^q$, donde $i = \sqrt{-1}$, $p > 0$, $q \geq 0$, $p + q = \text{rang}\,(B(k)) \leq n_k - 1$.

El centroide

$$\overline{x}(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i(k)$$

es el vector nulo en esta representación, pero convendrá utilizarlo formalmente.

TEOREMA 5.2.1 *La función discriminante $\phi_k(\omega)$ definida en (5.7) es el cuadrado de la distancia entre el individuo a asignar y el centroide $\overline{x}(k)$*

$$\phi_k(\omega) = \boldsymbol{\delta}^2\ (\omega\,,\, \overline{x}(k))$$

Demostración: Sea x_0 el vector (fila) de $\mathbf{R}^p \times i\,\mathbf{R}^q$, correspondiente al individuo ω. (Se ha calculado explícitamente este vector para el caso euclídeo en la proposición 2.1.1; un resultado análogo vale para el caso general no euclídeo, pero no haremos uso de este resultado).

$$\begin{aligned}
&\| x_0 - \overline{x}(k) \|^2 = \\
&= \left(x_0 - \frac{1}{n_k} \sum_{i=1}^{n_k} x_i(k) \right) \cdot \left(x_0 - \frac{1}{n_k} \sum_{i=1}^{n_k} x_i(k) \right)' \\
&= x_0 \cdot x_0' - \frac{2}{n_k}\, x_0 \cdot \left(\sum_{i=1}^{n_k} x_i(k) \right)' + 0 \\
&= \frac{1}{n_k} \sum_{i=1}^{n_k} \left(x_0 \cdot x_0' - 2\, x_0 \cdot x_i' \right) \\
&= \frac{1}{n_k} \sum_{i=1}^{n_k} \left((x_0 - x_i) \cdot (x_0 - x_i)' - x_i \cdot x_i' \right) \\
&= \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{\delta}_i(k) - \frac{1}{n_k} \operatorname{tr}(B(k))
\end{aligned}$$

Finalmente, por la definición de $B(k)$,

$$-2\, b_{i\,j}(k) = \boldsymbol{\delta}^2_{i\,j}(k) - \frac{1}{n_k}\, s_i - \frac{1}{n_k}\, s_j + \frac{1}{n_k^2}\, D$$

donde s_i es la suma de los elementos de la fila i de la matriz $\boldsymbol{\Delta}^{(2)}(k)$ y $D(k)$ la suma de todos los elementos de $\boldsymbol{\Delta}^{(2)}(k)$. En particular

$$-2\, b_{i\,i}(k) = -\frac{2}{n_k}\, s_i + \frac{1}{n_k^2}\, D(k)$$

Por tanto

$$\operatorname{tr}(B(k)) = \frac{1}{2\, n_k}\, D(k)$$

Sustituyendo se llega al enunciado.

Obsérvese que en este cálculo no se ha supuesto que $\boldsymbol{\Delta}(k)$ sea euclídea (si no lo es, los cálculos con los x_i son entre números complejos). □

Como corolario del teorema (5.2.1) resulta que si la muestra consiste en grupos no solapantes (es decir, tales que la distancia de cada individuo al centroide del grupo a que pertenece es menor que la distancia de este individuo al centroide del otro grupo), entonces la clasificación es perfecta.

5.2.2 Estimación del error

El estimador *leave–one–out* de la probabilidad de clasificación errónea (Lachenbruch [71]) puede aplicarse con facilidad al método DB.

Se calcula esta estimación eliminando por turno cada individuo de la muestra y asignándolo a una u otra subpoblación según la regla de decisión obtenida a partir de los restantes $n-1$ individuos de la muestra.

El resultado de esta operación puede expresarse mediante una matriz C *de clasificación*, cuyo elemento $(r,\, s)$ es el número de individuos del grupo r que han sido asignados al grupo s por el algoritmo discriminante.

El estimador *leave–one–out* de la probabilidad de clasificación errónea es entonces

$$\frac{1}{n}\left(c_{1\,2}+c_{2\,1}\right)$$

Para el método DB, se recalcula la función discriminante de la subpoblación a que pertenece el individuo eliminado (sea, a titulo de ejemplo, la primera)

$$\phi_1(\boxed{\mathrm{i}}) = \frac{1}{n_1-1}a_i - \frac{1}{(n_1-1)^2}(D(1)-a_i) \tag{5.8}$$

donde a_i es la suma de distancias al cuadrado de $\boxed{\mathrm{i}}$ a los restantes individuos de la primera subpoblación y $D(1)$ se ha definido en la sección precedente. La segunda función discriminante es

$$\phi_2(\boxed{\mathrm{i}}) = \frac{1}{n_2}b_i - \frac{1}{{n_2}^2}D(2) \tag{5.9}$$

donde b_i es la suma de distancias al cuadrado de $\boxed{\mathrm{i}}$ a los individuos de la segunda subpoblación.

Compárese este cálculo con el equivalente para otros discriminadores: por ejemplo en LDF y QDF se debe recalcular la inversa de una matriz de covarianzas para cada uno de los n individuos de la muestra.

5.2.3 Método DB para variables aleatorias

Supongamos que las variables observadas siguen la distribución de probabilidad F_k para la subpoblación π_k $(k=1,\, 2)$. y que con respecto a una medida adecuada μ, la correspondiente densidad de probabilidad es f_k $(k=1,\, 2)$.

Sea ξ_0 el resultado de la observación sobre un individuo ω a asignar, y $\boldsymbol{\delta}$ una función distancia. La función discriminante que generaliza (5.7) es

$$\begin{aligned}\phi_k(\xi_0) &= \int \boldsymbol{\delta}^2\left(\xi_0,\xi\right) f_k(\xi)\, d\mu\left(\xi\right) - \frac{1}{2}\int \boldsymbol{\delta}^2\left(\xi,\eta\right) f_k(\xi)\, f_k(\eta)\, d\mu\left(\xi\right) d\mu\left(\eta\right) \\ &= \mathbf{H}_{k\,0} - \frac{1}{2}\mathbf{H}_k \end{aligned} \tag{5.10}$$

En la segunda igualdad, $\mathbf{H}_{k\,0}$ es el valor esperado en π_k de la función $\boldsymbol{\delta}^2(\xi_0, \xi)$ (de la variable aleatoria ξ),

$$\mathbf{H}_{k\,0} = \mathbf{E}_{\pi_k}\left[\boldsymbol{\delta}^2(\xi_0, \xi)\right]$$

y $\mathbf{H}_k$ es el valor esperado en $\pi_k \times \pi_k$ de la función $\boldsymbol{\delta}^2(\xi, \eta)$ de las dos variables aleatorias ξ, η, independientes y con igual distribución F_k

$$\mathbf{H}_k = \mathbf{E}_{\pi_k \times \pi_k}\left[\boldsymbol{\delta}^2(\xi, \eta)\right] \tag{5.11}$$

Nota 5.2.1 *La exposición del método DB se ha realizado para el caso de dos muestras o poblaciones por claridad de notación. Aplicando las modificaciones obvias en la notación, puede verse que es igualmente válido para más de dos muestras o poblaciones.*

5.2.4 Propiedades básicas del método DB

1) En primer lugar observamos que la expresión $\mathbf{H}_k$ en (5.10) es una medida de la dispersión en la población π_k y que $\mathbf{H}_{k\,0}$ es un promedio de las diferencias entre el individuo ω a asignar y los individuos de π_k. Considerando una población π_0 formada por el único individuo ω, tenemos que $H_0 = 0$, y podemos escribir

$$\phi_k(\omega) = \mathbf{H}_{k\,0} - \frac{1}{2}\left(\mathbf{H}_k + \mathbf{H}_0\right) \tag{5.12}$$

Es decir, la función discriminante es una diferencia de Jensen entre π_k y π_0.

2) Si en cada población se observan dos vectores aleatorios independientes x, y para los que se tienen distancias $\boldsymbol{\delta}_x^2(x_1,\, x_2)$ y $\boldsymbol{\delta}_y^2(y_1,\, y_2)$, segun Oller [84], una manera natural de definir una distancia entre los pares (x, y) que sea consistente con la independencia de las variables x, y es

$$\boldsymbol{\delta}^2\left((x_1, y_1),\, (x_2, y_2)\right) = \boldsymbol{\delta}_x^2(x_1,\, x_2) + \boldsymbol{\delta}_y^2(y_1,\, y_2)$$

Con esta distancia, vemos que, por la aditividad de la esperanza matemática, las funciones discriminantes $\phi_k(\omega)$ de (5.10), construidas teniendo en cuenta las dos variables x, y son la suma de las $\phi[x]_k$ y $\phi[y]_k$, construidas a partir de cada una de las variables por separado.

3) El método DB permite la consideración de probabilidades *a priori*: Si ω pertenece a π_k con probabilidad *a priori* igual a q_k, entonces las funciones discriminantes (5.10) deben sustituirse por

$$\phi_k(\omega) = \mathbf{H}_{k\,0} - \frac{1}{2}\,\mathbf{H}_k + \frac{1}{q_k} - 1$$

Vease Cuadras [28] para una justificación de esta fórmula y una discusión de la relación entre estas funciones discriminantes y las (5.6) obtenidas de la regla de decisión de Bayes.

5.2.5 Teorema de representación

La siguiente proposición, análoga al teorema (5.2.1) para muestras, permite interpretar el método DB como una regla de mínima distancia.

TEOREMA 5.2.2 *Supongamos que cada población π_k tiene una representación en un espacio vectorial m_k–dimensional E^{m_k} (euclídeo o pseudo–euclídeo), es decir, que existen funciones*

$$\psi_k : \pi_k \longrightarrow E^{m_k}$$

tal que para cada par $(x,\, y) \in \pi_k \times \pi_k$ *se verifica que*

$$\boldsymbol{\delta}^2(x,\, y) = \|\psi_k(x) - \psi_k(y)\|^2$$

Supongamos adicionalmente que existen los momentos

$$\begin{aligned} \boldsymbol{\mu}\ (k) &= \mathbf{E}_{\pi_k}(\psi_k(x)) \\ \boldsymbol{\mu}_2\,(k) &= \mathbf{E}_{\pi_k}(\psi_k(x)' \cdot \psi_k(x)) \end{aligned}$$

Entonces, la función discriminante $\phi_k(\omega)$ *para un individuo* ω *cuya observación es* x_0 *viene dada por*

$$\phi_k(\omega) = \|\psi_k(x_0) - \boldsymbol{\mu}(k)\|^2$$

Demostración:

$$\begin{aligned} \mathbf{H}_{k\,0} &= \mathbf{E}_{\pi_k}(\boldsymbol{\delta}^2(x,\, x_0)) \\ &= \mathbf{E}_{\pi_k}((s - s_0)' \cdot (s - s_0)) \\ &\qquad \text{siendo } s = \psi(x) \quad \text{y } s_0 = \psi(x_0) \\ &= \mathbf{E}_{\pi_k}(s' \cdot s - 2\, s_0' \cdot s + s_0' \cdot s_0) \\ &= \boldsymbol{\mu}_2(k) - 2\, s_0' \cdot \boldsymbol{\mu}(k) + s_0' \cdot s_0 \end{aligned}$$

Análogamente, poniendo $t = \psi(y)$, calculamos

$$\begin{aligned} \mathbf{H}_k &= \mathbf{E}_{\pi_k \times \pi_k}(\boldsymbol{\delta}^2(x,\, y)) \\ &= \mathbf{E}_{\pi_k \times \pi_k}((s - t)' \cdot (s - t)) \\ &= \mathbf{E}_{\pi_k \times \pi_k}(s' \cdot s - 2\, s' \cdot t + t' \cdot t) \\ &= 2\,[\boldsymbol{\mu}_2(k) - (\boldsymbol{\mu}(k))^2] \end{aligned}$$

con lo que al substituir en

$$\phi_k(\omega) = \mathbf{H}_{k\,0} - \frac{1}{2}\,\mathbf{H}_k$$

obtenemos el enunciado. □

5.3 Distancias entre individuos para el modelo DB

5.3.1 Aspectos generales

Según la definición de las funciones discriminantes del modelo DB, puede aplicarse dicho modelo siempre que se disponga de una función distancia entre individuos de cada subpoblación.

Cada individuo u observación queda especificado por un conjunto x de *coordenadas*, que pueden agruparse en numéricas continuas, binarias, cualitativas de tipo ordinal y cualitativas de tipo nominal (es decir, cuya codificación numérica es puramente convencional).

Se requieren entonces dos funciones distancia, una para cada subpoblación,

$$D_1(x_A, x_B) \qquad D_2(x_A, x_B)$$

de modo que $D_i(x_A, x_B)$ sea la distancia entre las observaciones x_A y x_B en el supuesto de pertenecer ambas a la subpoblación i. Las funciones D_1 y D_2 pueden coincidir, lo que redunda en una considerable simplificación de la estimación del error de clasificación, según veremos en el Capítulo 6 (6.3), al disponerse en este caso de una matriz de distancias global, pero esto no es una exigencia del modelo.

No es preciso que la función distancia proceda de un modelo probabilístico, ni que dé lugar a una configuración euclídea.

Por ello hay gran flexibilidad en la elección de funciones distancia, que pueden elegirse entre la multitud de las descritas en la literatura, o bien prepararse una *ad hoc*, según la información de que se disponga en cada caso concreto del significado de las variables y de sus relaciones.

Como ejemplo de distancias del primer tipo, tenemos las distancias entre variables continuas, como la distancia euclídea, las distancias de Minkowski

$$\boldsymbol{\delta}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

donde $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ son las coordenadas de dos observaciones x, y, y la distancia Valor Absoluto estudiada en los capítulos anteriores.

Si las coordenadas son sólo binarias, se dispone tambien de gran cantidad de distancias y coeficientes de similaridad descritos en la literatura, como por ejemplo, los de Jaccard, Sokal y Sneath, Kulezynski, etc.

Cuando las variables son de tipo mixto, una función distancia muy empleada es la de Gower, ya mencionada en el Capítulo 2 (2.11).

Por último, como ilustración de otras funciones distancia aplicables en contextos concretos, podemos mencionar la distancia de Levenshtein (véase

[100]), empleada para medir las diferencias entre dos cadenas de carácteres, que en su variante más simple puede describirse como el mínimo número de substituciones, inserciones y eliminaciones de carácteres que hay que efectuar sobre una cadena para obtener la otra. Un caso más general considera costes prefijados para insertar y eliminar cada carácter del alfabeto empleado, y para intercambiar cada pareja de carácteres.

5.3.2 Distancia basada en *efficient scores*

Si se dispone de un modelo probabilístico paramétrico para las observaciones x, definido mediante una función densidad $p(x,\theta)$ dependiente de los n parámetros $\theta = (\theta_1, \ldots, \theta_n)$, las funciones distancia entre individuos que resultan idóneas son las que se deducen del estudio de la Geometría Riemanniana de la variedad n–dimensional de los paramétros, cuyo tensor métrico es la *métrica de Rao*), que en las coordenadas θ se expresa por la matriz de información de Fisher

$$G_\theta = -\mathbf{E}\left(\frac{\partial^2 \log p\,(X,\theta)}{\partial\theta\cdot\partial\theta'}\right) = \mathbf{E}\left(\frac{\partial \log p\,(X,\theta)}{\partial\theta}\cdot\frac{\partial \log p\,(X,\theta)}{\partial\theta'}\right)$$

La distancia entre las observaciones x_1 y x_2 se calcula empleando los *efficient scores*, definidos como los vectores n–dimensionales

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

y se obtiene por

$$\boldsymbol{\delta}(x_1,x_2) = (z_1 - z_2)'\cdot G_\theta^{-1}\cdot(z_1 - z_2) \tag{5.13}$$

(véase Cuadras [23], Oller [84], Miñarro [79])

5.3.3 Condiciones para una distancia entre observaciones

Observando que si se emplea la distancia entre individuos basada en los *efficient scores* para la población π_k se cumple que

$$\begin{aligned}\mathbf{H}_k = \mathbf{E}_{\pi_k\times\pi_k}\ &\left[\boldsymbol{\delta}^2\,(Z_1,Z_2)\right] &&= \mathbf{E}_{\pi_k\times\pi_k}\ [(Z_1-Z_2)'\cdot G^{-1}\cdot(Z_1-Z_2)]\\ = 2\,\mathbf{E}_{\pi_k}\ &[Z'\cdot G^{-1}\cdot Z] &&= 2\,\mathbf{E}_{\pi_k}\ [\ \mathrm{tr}\,(G^{-1}\cdot Z\cdot Z')]\\ = 2\,\mathbf{E}_{\pi_k}\ &[\ \mathrm{tr}\,(G^{-1}\cdot G)] = 2\,n\end{aligned} \tag{5.14}$$

puede proponerse esta igualdad como condición de normalización en el caso de no disponer de un modelo paramétrico, siendo ahora n el número de variables.

Esta condición se extiende al caso de tener n variables x que se agrupan en dos o más conjuntos de variables, desconociéndose la estructura de dependencia entre ellas. Un ejemplo frecuente aparece si se tiene una mezcla de variables continuas, binarias y cualitativas.

La existencia de una distribución de probabilidad conjunta cuyas marginales son las de cada uno de los conjuntos es probada en Cuadras [29].

En dicho caso se puede proponer que la condición de normalización ha de aplicarse a cada conjunto de variables por separado, y finalmente obtener la distancia al cuadrado total como suma de las distancias al cuadrado componentes, una vez normalizadas (véase Cuadras [27], [31]).

Esto es una generalización natural del caso paramétrico, pues entonces la métrica de Rao tiene la forma

$$\begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix}$$

y se verifica (5.14) para G_1 y G_2 por separado, y la distancia global al cuadrado es la suma de los dos terminos correspondientes a G_1 y G_2.

5.4 Ejemplos con distribuciones conocidas

Como aplicación de (5.10), calculamos las funciones discriminantes para algunos casos de distribuciones clásicas, empleando como distancia entre individuos la calculada a partir de (5.13)

5.4.1 Distribución discreta finita genérica

Supongamos que la variable observada sigue una distribución discreta finita con m estados, cuyos parámetros son $(p_1, \ldots, p_m)$ en la población π_1, y $(q_1, \ldots, q_m)$ en la población π_2.

La función de probabilidad en π_1 es

$$f(x) = p_r \qquad \text{si } x = e_r \qquad (r = 1, \ldots, m)$$

donde

$$e_r = (0, \ldots, 0, \underset{\substack{\text{posición} \\ r}}{\overset{\uparrow}{1}}, 0, \ldots, 0)$$

Una expresión análoga, con q en lugar de p vale para π_2.

La distancia entre dos individuos $x = e_r$, $y = e_s$ en π_1 (véase Miñarro [79, pp. 66–67]) se calcula por

$$\boldsymbol{\delta}^2(x, y) = (1 - \delta_{r\,s}) \left(\frac{1}{p_r} + \frac{1}{p_s} \right)$$

Proposición 5.4.1 *Las funciones discriminantes para un individuo ω con valor observado $x_0 = e_r$ son*

$$\phi_1(\omega) = \frac{1 - p_r}{p_r} \qquad \phi_2(\omega) = \frac{1 - q_r}{q_r}$$

Demostración: Calculamos ϕ_1. Cambiando p por q resulta ϕ_2

$$\begin{aligned}
\phi_1(\omega) &= \\
&= \sum_{s=1}^{m} \left((1 - \delta_{r\,s}) \left(\frac{1}{p_r} + \frac{1}{p_s} \right) \right) p_s - \\
&\qquad - \frac{1}{2} \sum_{s=1}^{m} \sum_{t=1}^{m} \left((1 - \delta_{s\,t}) \left(\frac{1}{p_s} + \frac{1}{p_t} \right) \right) p_s p_t \\
&= \frac{1 - p_r}{p_r} + (m - 1) - \frac{1}{2} \left(\sum_{s=1}^{m} \sum_{t=1}^{m} (1 - \delta_{s\,t}) (p_s + p_t) \right) \\
&= \frac{1 - p_r}{p_r} + (m - 1) - \frac{1}{2} \left(\sum_{s=1}^{m} (1 - p_s) + \sum_{t=1}^{m} (1 - p_t) \right) \\
&= \frac{1 - p_r}{p_r}
\end{aligned}$$

□

En consecuencia, se asignará ω a la subpoblación π_1 si

$$\phi_1(\omega) < \phi_2(\omega) \iff p_r > q_r$$

5.4.2 Distribución multinomial

Sea $x = (x_1, \ldots, x_m)$ una variable aleatoria con distribución multinomial, de parámetros n y $p = (p_1, \ldots, p_m)$.

Los x_i son enteros positivos o nulos con $\sum_{i=1}^m x_i = n$, los parámetros p_i son números reales en el intervalo $[0, 1]$, verificándose que $\sum_{i=1}^m p_i = 1$

La función de probabilidad de x es

$$f(x/n, p) = \frac{n!}{x!}\, p^x$$

donde se han empleado las notaciones

$$x! \equiv \prod_{i=1}^m x_i! \qquad \text{y} \qquad p^x \equiv \prod_{i=1}^m p_i^{x_i}$$

Emplearemos la distancia entre individuos para esta distribución dada por Miñarro [79, pág. 66]

$$\boldsymbol{\delta}^2(x,\, y) = \frac{1}{n} \sum_{i=1}^m \frac{(x_i - y_i)^2}{p_i}$$

Consideramos el problema de asignar a una de las poblaciones π_1, π_2 un individuo ω para el que se ha observado el valor $u = (u_1, \ldots, u_m)$ siendo los parámetros n y $p = (p_1, \ldots, p_m)$ en π_1, y n y $q = (q_1, \ldots, q_m)$ en π_2.

Proposición 5.4.2 *Las funciones discriminantes coinciden con el clásico estadístico χ^2 de K. Pearson en cada una de las poblaciones. Es decir:*

$$\phi_1(\omega) = \sum_{i=1}^m \frac{(u_i - n\, p_i)^2}{n\, p_i} \qquad \phi_2(\omega) = \sum_{i=1}^m \frac{(u_i - n\, q_i)^2}{n\, q_i}$$

Demostración:

Calculamos $\phi_1(\omega)$. El mismo cálculo sirve para $\phi_2(\omega)$, cambiando p por q.

$$\begin{aligned}
\mathbf{H}_{1\,0} &= \mathsf{E}\left(\boldsymbol{\delta}^2(u,x)\right) \\
&= \frac{1}{n}\sum_{i=1}^{m}\frac{1}{p_i}\,\mathsf{E}\left(u_i^2 - 2\,u_i\,x_i + x_i^2\right) \\
&\qquad \text{como cada marginal } x_i \text{ es } B(n,\,p_i), \text{ resulta} \\
&= \frac{1}{n}\sum_{i=1}^{m}\frac{1}{p_i}\left(u_i^2 - 2\,u_i\,n\,p_i + n\,p_i\,(1-p_i) + n^2\,p_i^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{m}\frac{u_i^2}{p_i} - n + (m-1)
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_1 &= \mathsf{E}\left(\boldsymbol{\delta}^2(x,y)\right) = \frac{1}{n}\sum_{i=1}^{m}\frac{1}{p_i}\,\mathsf{E}\left(x_i^2 - 2\,x_i\,y_i + y_i^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{m}\frac{2}{p_i}\,\left(n\,p_i\,(1-p_i)\right) \\
&= 2\,(m-1)
\end{aligned}$$

Finalmente, sustituyendo en $\phi_1(\omega) = \mathbf{H}_{1\,0} - (1/2)\,\mathbf{H}_1$, y agrupando términos, se llega al enunciado. □

5.4.3 Distribución multinomial negativa

Sea $x = (x_1, \ldots, x_m)$ una variable aleatoria con distribución multinomial negativa, de parámetros r y $p = (p_1, \ldots, p_m)$.

Los x_i son enteros positivos o nulos, r es un entero positivo, los p_i son números reales en el intervalo $[0, 1]$, verificándose que $\sum_{i=1}^{m} p_i < 1$

La función de probabilidad de x es

$$f(x/p,r) = \frac{(|x|+r-1)!}{x!\,(r-1)!}\,p^x\,(1-|p|)^r$$

donde se han empleado las notaciones

$$x! \equiv \prod_{i=1}^{m} x_i! \qquad p^x \equiv \prod_{i=1}^{m} p_i^{x_i}$$
$$|x| \equiv \sum_{i=1}^{m} x_i \qquad \text{y} \qquad |p| \equiv \sum_{i=1}^{m} p_i$$

Emplearemos la distancia entre individuos para esta distribución dada por Miñarro [79, pág. 67]

$$\boldsymbol{\delta}^2\,(x,\,y) = \frac{1-|p|}{r}\left[\sum_{i=1}^{m}\frac{1}{p_i}\,(x_i-y_i)^2 - (|x|-|y|)^2\right]$$

Consideramos el problema de asignar a una de las poblaciones π_1, π_2 un individuo ω para el que se ha observado el valor $u = (u_1, \ldots, u_m)$ siendo los parámetros $p = (p_1, \ldots, p_m)$ en π_1, y $q = (q_1, \ldots, q_m)$ en π_2, con igual r en ambas poblaciones.

Para el cálculo de las funciones discriminantes necesitamos usar la siguiente propiedad

Lema 5.4.1 *Si $x = (x_1, \ldots, x_m)$ es multinomial negativa con parámetros $p = (p_1, \ldots, p_m)$ y r, entonces*

- *Cada marginal x_i es binomial negativa, con parametros r y*

$$p_i' = \frac{p_i}{1-|p|+p_1}$$

- *La suma $|x|$ es binomial negativa, con parámetros r y $|p|$*

Demostración: La función característica de la distribución multinomial negativa es

$$\psi(t/p,r) = \left(\frac{1-|p|}{1-\sum_{k=1}^{m} p_k \exp(i\,t_k)}\right)^r \tag{5.15}$$

donde $t = (t_1, \ldots, t_m)$, y $i = \sqrt{-1}$.

Haciendo $t_1 = \cdots = t_{m-1} = 0$ obtenemos la función característica de la marginal t_m.

$$\psi(t_m/p, r) = \\ = \left(\frac{1-|p|}{1-|p|+p_m - p_m \exp(i\, t_m)}\right)^r = \left(\frac{1-p'_m}{1-p'_m \exp(i\, t_m)}\right)^r$$

que es la función característica de una distribución binomial negativa. Por simetría tenemos las demás marginales.

La distribución de la suma $|x|$ se obtiene por recurrencia:

Consideramos las variables y_i definidas por

$$\begin{aligned} y_i &= x_i \qquad (1 \le i \le m-2) \\ y_{m-1} &= x_{m-1} + x_m \end{aligned}$$

Se obtiene en primer lugar la distribución conjunta de las y, que resulta ser multinomial negativa con los parámetros

$$\begin{aligned} p'_i &= p_i \qquad (1 \le i \le m-2) \\ p'_{m-1} &= p_{m-1} + p_m \end{aligned}$$

Esto puede verse empleando la variable auxiliar

$$y_m = x_m$$

y efectuando el cambio de variables $x \longrightarrow y$. El cambio inverso es

$$\begin{aligned} x_i &= y_i \qquad (1 \le i \le m-2) \\ x_{m-1} &= y_{m-1} - y_m \\ x_m &= y_m \end{aligned}$$

Hay que notar que el recorrido de la variable y_m es el intervalo $[0,\ y_{m-1}]$.

La función de probabilidad conjunta de las y es

$$f(y(m-1), y_m) = \\ = \frac{(|y(m-1)|+r-1)!}{y(m-2)!\,(y_{m-1}-y_m)!\,y_m!\,(r-1)!} \times \\ \times (1-|p|)^r\, p(m-2)^{y(m-2)}\, p_{m-1}^{y_{m-1}-y_m}\, p_m^{y_m}$$

donde la notación $y(m-1)$ se emplea para indicar el vector $(y_1, \ldots, y_{m-1})$, y notaciones análogas se interpretan de la misma manera.

Finalmente se calcula la marginal de $y(m-1)$, sumando para todos los valores de y_m, de 0 a y_{m-1}. □

Como consecuencia del lema, tenemos

$$\begin{aligned}
\mathbf{E}\,(x_i) &= \tfrac{r\,p_i}{1-|p|} \\
\mathbf{E}\,(x_i^2) &= \tfrac{r\,p_i}{1-|p|} + \tfrac{r\,(r+1)\,p_i^2}{(1-|p|)^2} \\
\mathbf{E}\,(|x|) &= \tfrac{r\,|p|}{1-|p|} \\
\mathbf{E}\,(|x|^2) &= \tfrac{r\,|p|\,(1+r\,|p|)}{(1-|p|)^2}
\end{aligned}$$

Proposición 5.4.3 *La función discriminante para π_1 es*

$$\phi_1(\omega) = \frac{1-|p|}{r}\left[\sum_{i=1}^{m}\frac{u_i^2}{p_i} + \frac{r^2}{1-|p|} - (|u|+r)^2\right]$$

La función discriminante para π_2 se obtiene cambiando p por q en esta expresión.

Demostración: $\mathbf{H}_{1\,0} =$

$$
\begin{aligned}
&= \frac{1-|p|}{r}\left[\sum_{i=1}^{m}\frac{1}{p_i}\,\mathbf{E}\left(u_i^2-2\,u_i\,x_i+x_i^2\right)\right.\\
&\qquad\left.-\,\mathbf{E}\left(|u|^2-2\,|u|\,|x|+|x|^2\right)\right]\\
&= \frac{1-|p|}{r}\left[\sum_{i=1}^{m}\frac{1}{p_i}\left\{u_i^2-2\,u_i\frac{r\,p_i}{1-|p|}+\frac{r\,p_i}{1-|p|}+\frac{r(r+1)\,p_i^2}{(1-|p|)^2}\right\}\right.\\
&\qquad\left.-\left\{|u|^2-\frac{2\,|u|\,r\,|p|}{1-|p|}+\frac{r\,|p|\,(1+r\,|p|)}{(1-|p|)^2}\right\}\right]\\
&= \frac{1-|p|}{r}\left[\sum_{i=1}^{m}\frac{u_i^2}{p_i}-\frac{2\,r\,|u|}{1-|p|}+\frac{r\,m}{1-|p|}+\frac{r\,(r+1)\,|p|}{(1-|p|)^2}-|u|^2+\right.\\
&\qquad\left.+2\,\frac{r\,|p|\,|u|}{1-|p|}-\frac{r\,|p|\,(1+r\,|p|)}{(1-|p|)^2}\right]\\
&= m+\frac{1-|p|}{r}\left[\sum_{i=1}^{m}\frac{u_i^2}{p_i}+\frac{r^2}{1-|p|}-(|u|+r)^2\right]
\end{aligned}
$$

Análogamente, $\mathbf{H}_1 =$

$$
\begin{aligned}
&= \frac{1-|p|}{r}\left[\sum_{i=1}^{m}\frac{1}{p_i}\,\mathbf{E}\left(x_i^2-2\,x_i\,y_i+y_i^2\right)-\mathbf{E}\left(|x|^2-2\,|x|\,|y|+|y|^2\right)\right]\\
&= \frac{1-|p|}{r}\left[\sum_{i=1}^{m}\frac{2}{p_i}\left(\frac{r\,p_i}{1-|p|}+\frac{r\,(r+1)\,p_i^2}{(1-|p|)^2}-\frac{r^2\,p_i^2}{(1-|p|)^2}\right)-2\,\frac{r\,|p|}{(1-|p|)^2}\right]\\
&= 2\,\frac{1-|p|}{r}\left[\frac{r\,m}{1-|p|}+\frac{r\,|p|}{(1-|p|)^2}-\frac{r\,|p|}{(1-|p|)^2}\right]\\
&= 2\,m
\end{aligned}
$$

□

5.4.4 Distribución normal univariante

Sea $x=(x_1,\ldots,x_m)$ una muestra aleatoria simple de una variable normal univariante $N(\mu,\sigma)$.

Emplearemos la distancia entre individuos para esta distribución dada por Miñarro [79, pág. 68]

$$\boldsymbol{\delta}^2\,(x,\,y) = \frac{m}{2\,\sigma^4}\left[2\,\sigma^2(\overline{x}-\overline{y})^2 + (S'^2_x - S'^2_y)^2\right]$$

donde

$$\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$S'^2_x = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2 = \frac{\sigma^2}{m}\sum_{i=1}^{m}\left(\frac{(x_i-\mu)}{\sigma}\right)^2 \equiv \frac{\sigma^2}{m}\,Q_x$$

La variable aleatoria $\overline{x}$ tiene distribución $N(\mu,\,\sigma/\sqrt{m})$, y Q_x tiene distribución $\chi^2\,(m)$.

Consideramos el problema de asignar a una de las poblaciones π_1, π_2 un individuo ω para el que se ha observado el valor $u = (u_1, \ldots, u_m)$. Los parámetros μ y σ son distintos en las dos poblaciones, pero como no tenemos que escribirlos simultáneamente, omitiremos subíndices, entendiéndose que las fórmulas obtenidas han de aplicarse por separado a cada población, sustituyendo los valores pertinentes de los parámetros.

Proposición 5.4.4 *La función discriminante correspondiente a la población π_k, en la que los parámetros son (μ, σ) es*

$$\phi_k(\omega) = \frac{m}{2\,\sigma^4}\left[2\,\sigma^2\,(\overline{u}-\mu)^2 + (S'^2_u - \sigma^2)^2\right]$$

Demostración:

$$\begin{aligned}
\mathbf{H}_{k\,0} &= \\
&= \frac{m}{\sigma^2}\,\mathbf{E}\left(\overline{u}^2 - 2\,\overline{u}\,\overline{x} + \overline{x}^2\right) + \frac{1}{2\,m}\,\mathbf{E}\left(Q_u^2 - 2\,Q_u\,Q_x + Q_x^2\right) \\
&= \frac{m}{\sigma^2}\left(\overline{u}^2 - 2\,\overline{u}\,\mu + \frac{\sigma^2}{m} + \mu^2\right) + \\
&\qquad + \frac{1}{2\,m}\left(Q_u^2 - 2\,Q_u\,m + m^2 + 2\,m\right) \\
&= 2 + \frac{m}{\sigma^2}\,(\overline{u}-\mu)^2 + \frac{1}{2\,m}(Q_u - m)^2
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_k &= \frac{m}{\sigma^2}\,\mathbf{E}\left(\overline{x}^2 - 2\,\overline{x}\,\overline{y} + \overline{y}^2\right) + \frac{1}{2\,m}\,\mathbf{E}\left(Q_x^2 - 2\,Q_x\,Q_y + Q_y^2\right) \\
&= \frac{m}{\sigma^2}\,2\left(\frac{\sigma^2}{m}\right) + \frac{1}{2\,m}\,2\,(2\,m) \\
&= 4
\end{aligned}$$

□

5.4.5 Distribución normal multivariante con Σ conocida

Sea x una variable aleatoria m–dimensional con distribución $N(\mu_1, \Sigma)$ en π_1 y distribución $N(\mu_2, \Sigma)$ en π_2.

La distancia entre dos individuos cuyas observaciones son x, y se calcula por la expresión (formalmente idéntica a la distancia de Mahalanobis entre poblaciones, véase Miñarro [79, pp. 70–71])

$$\boldsymbol{\delta}(x,\, y) = (x-y)' \cdot \Sigma^{-1} \cdot (x-y) \tag{5.16}$$

Proposición 5.4.5 *Las funciones discriminantes para un individuo ω cuya observación és x_0 son*

$$\phi_k(\omega) = (x_0 - \mu_k)' \cdot \Sigma^{-1} \cdot (x_0 - \mu_k) \qquad k = 1,\, 2$$

Demostración:

$$\begin{aligned}
\mathbf{H}_{k\,0} &= \mathbf{E}_{\pi_k}\left((x-x_0)' \cdot \Sigma^{-1} \cdot (x-x_0)\right) \\
&\quad \text{haciendo } x - x_0 = (x-\mu_k) + (\mu_k - x_0) \\
&= \mathbf{E}_{\pi_k}\Big((x-\mu_k)' \cdot \Sigma^{-1} \cdot (x-\mu_k) + \\
&\qquad 2\,(\mu_k - x_0)' \cdot \Sigma^{-1} \cdot (x-\mu_k) \; + \\
&\qquad + (\mu_k - x_0)' \cdot \Sigma^{-1} \cdot (\mu_k - x_0)\Big) \\
&= m + (\mu_k - x_0)' \cdot \Sigma^{-1} \cdot (\mu_k - x_0)
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_{k} &= \mathbf{E}_{\pi_k \times \pi_k}\left((x-y)' \cdot \Sigma^{-1} \cdot (x-y)\right) \\
&\quad \text{haciendo } x - y = (x-\mu_k) + (\mu_k - y) \\
&= \mathbf{E}_{\pi_k \times \pi_k}\Big((x-\mu_k)' \cdot \Sigma^{-1} \cdot (x-\mu_k) + \\
&\qquad 2\,(x-\mu_k)' \cdot \Sigma^{-1} \cdot (\mu_k - y) \; + \\
&\qquad + (y-\mu_k)' \cdot \Sigma^{-1} \cdot (y-\mu_k)\Big) \\
&= 2\,m
\end{aligned}$$

□

5.5 Distancias entre poblaciones

5.5.1 Distancia basada en la diferencia de Jensen

Si se dispone de una distancia entre individuos global, de forma que sea posible calcular la distancia entre dos individuos, uno de cada subpoblación, puede proponerse una distancia entre poblaciones a partir de una distancia entre individuos, por la diferencia de Jensen

$$\Delta_{1\,2} = \mathbf{H}_{1\,2} - \frac{1}{2}\left(\mathbf{H}_1 + \mathbf{H}_2\right) \tag{5.17}$$

donde los $\mathbf{H}_k$ son, como en (5.11)

$$\mathbf{H}_k = \mathbf{E}_{\pi_k \times \pi_k}\left[\boldsymbol{\delta}^2\left(\xi, \eta\right)\right]$$

y $\mathbf{H}_{1\,2}$ se define por

$$\mathbf{H}_{1\,2} = \mathbf{E}_{\pi_1 \times \pi_2}\left[\boldsymbol{\delta}^2\left(\xi, \eta\right)\right]$$

Para una muestra de $n = n_1 + n_2$ individuos, la expresión correspondiente es

$$\widehat{\Delta}_{1\,2} = \widehat{\mathbf{H}}_{1\,2} - \frac{1}{2}\left(\widehat{\mathbf{H}}_1 + \widehat{\mathbf{H}}_2\right) \tag{5.18}$$

donde ahora los $\widehat{\mathbf{H}}_1$, $\widehat{\mathbf{H}}_2$, $\widehat{\mathbf{H}}_{1\,2}$ se calculan como medias de las cajas de la matriz $(n,\, n)$ de distancias global $\boldsymbol{\Delta}^{(2)}$.

$$\widehat{\mathbf{H}}_1 = \frac{1}{n_1^2} D_{1\,1} \qquad \widehat{\mathbf{H}}_2 = \frac{1}{n_2^2} D_{2\,2} \qquad \widehat{\mathbf{H}}_{1\,2} = \frac{1}{n_1\, n_2} D_{1\,2} \tag{5.19}$$

siendo

$$D_{1\,1} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \boldsymbol{\delta}_{i\,j} \qquad D_{1\,2} = \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \boldsymbol{\delta}_{i\,j} \qquad D_{2\,2} = \sum_{i=n_1+1}^{n} \sum_{j=n_1+1}^{n} \boldsymbol{\delta}_{i\,j} \tag{5.20}$$

Si se ha empleado una distancia no paramétrica, como por ejemplo, la distancia de Gower o la distancia Valor Absoluto, la matriz de distancias global $\boldsymbol{\Delta}^{(2)}$ aparece de modo natural.

Si se tiene un modelo paramétrico para π_1 y π_2, y se dispone de la distancia de Rao entre las dos poblaciones, ésta será la distancia óptima entre las dos poblaciones.

Pero en general, la distancia de Rao entre las poblaciones no es accesible, o presenta graves problemas computacionales. En estos casos puede recurrirse a (5.17) definiendo de algún modo apropiado la distancia entre un individuo x de π_1 y un individuo y de π_2, por ejemplo como un promedio

entre la distancia entre x, y considerados ambos en π_1 y la distancia entre x, y considerados ambos en π_2.

La expresión (5.18) puede interpretarse en términos de una representación euclídea (o pseudo–euclídea) de modo análogo a la interpretación dada en el teorema (5.2.1) de las funciones discriminantes.

Tenemos ahora una representación común

$$\psi : \pi_1 \cup \pi_2 \longrightarrow E^m$$

donde, en general, $E^m = \mathbf{R}^p \oplus i\,\mathbf{R}^q$, $(p + q = m \leq n - 1$, verificándose que

$$\boldsymbol{\delta}^2(x,\, y) = \|\psi(x) - \psi(y)\|^2$$

para cualquier par de individuos de la muestra conjunta. Como es usual, empleamos la descomposición

$$B = H \cdot (-\frac{1}{2}\boldsymbol{\Delta}^{(2)}) \cdot H = X \cdot X'$$

siendo H la matriz de centrado n–dimensional. Las filas x_i de X contienen los vectores de E^m que representan a los n individuos de la muestra. Suponemos que las n_1 primeras filas corresponden a π_1 y las n_2 filas comprendidas entre la $n_1 + 1$ y la n a π_2.

Tendremos los dos centroides

$$\overline{x}(1) = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \qquad \overline{x}(2) = \frac{1}{n_2}\sum_{i=n_1+1}^{n} x_i$$

Podemos ahora enunciar el teorema que da la interpretación de (5.18)

TEOREMA 5.5.1

$$\widehat{\Delta}_{1\,2} = \|\overline{x}(1) - \overline{x}(2)\|^2$$

Demostración:

$$\|\overline{x}(1) - \overline{x}(2)\|^2 = \overline{x}(1) \cdot \overline{x}(1)' + \overline{x}(2) \cdot \overline{x}(2)' - 2\,\overline{x}(1) \cdot \overline{x}(2)'$$

Teniendo en cuenta que $b_{i\,j} = x_i \cdot x'_j$, y empleando para las sumas de cajas de B expresiones análogas a las definidas en (5.20) para la matriz $\boldsymbol{\Delta}^{(2)}$, resulta

$$\|\overline{x}(1) - \overline{x}(2)\|^2 = \frac{1}{n_1^2}\,B_{1\,1} + \frac{1}{n_2^2}\,B_{2\,2} - \frac{2}{n_1\,n_2}\,B_{2\,2} \qquad (5.21)$$

Ahora, puesto que

$$-2\,b_{i\,j} = \boldsymbol{\delta}^2_{i\,j} - \frac{1}{n}\,s_i - \frac{1}{n}\,s_j + \frac{1}{n^2}\,D$$

(siendo s el vector que contiene las sumas de las filas (o columnas) de $\boldsymbol{\Delta}^{(2)}$ y D la suma de todos los elementos de esta matriz), tenemos

$$\begin{aligned} -2\sum_{i=1}^{n_1} b_{i\,j} &= \sum_{i=1}^{n_1} \boldsymbol{\delta}^2_{i\,j} - \frac{1}{n}\,(D_{1\,1} + D_{1\,2}) - \frac{n_1}{n}\,s_j + \frac{n_1}{n^2}\,D \\ -2\sum_{i=n_1+1}^{n} b_{i\,j} &= \sum_{i=n_1+1}^{n} \boldsymbol{\delta}^2_{i\,j} - \frac{1}{n}\,(D_{2\,2} + D_{1\,2}) - \frac{n_2}{n}\,s_j + \frac{n_2}{n^2}\,D \end{aligned}$$

De estas igualdades resulta

$$\begin{aligned} B_{1\,1} &= \frac{n_1 - n_2}{2\,n}\,D_{1\,1} + \frac{n_1}{n}\,D_{1\,2} - \frac{n_1^2}{2\,n^2}\,D \\ B_{2\,2} &= \frac{n_2 - n_1}{2\,n}\,D_{2\,2} + \frac{n_2}{n}\,D_{1\,2} - \frac{n_2^2}{2\,n^2}\,D \\ -2\,B_{1\,2} &= -\frac{n_2}{n}\,D_{1\,1} - \frac{n_1}{n}\,D_{2\,2} - \frac{n_1\,n_2}{n^2}\,D \end{aligned}$$

y al substituir estas expresiones en (5.21) se llega al enunciado.
□

La fórmula (5.18) es, al igual que todo el método DB, fácilmente generalizable al caso de disponer de $k > 2$ subpoblaciones, dando lugar en dicho caso, a una matriz $\boldsymbol{\Delta}_P{}^{(2)}$ de dimensión $(k,\,k)$ con las distancias entre cada par de subpoblaciones.

Sometiendo esta matriz a Multidimensional Scaling, se llega a una representación euclídea, que puede considerarse un análogo no paramétrico del Análisis Canónico de Poblaciones para poblaciones normales (véase, por ejemplo Cuadras [20]), y de hecho produce resultados coincidentes cuando es aplicable este análisis.

Esto es consecuencia de la siguiente

Proposición 5.5.1 *Supongamos que las poblaciones π_i son normales, con $\sum$ común conocida, y que se emplea la distancia entre individuos (5.16) (basada en* efficient scores*).*

Entonces la distancia $\Delta_{i\,j}$ entre las poblaciones π_i, y π_j coincide con la distancia de Mahalanobis. En particular, la matriz $\boldsymbol{\Delta}^2_P$ de distancias entre poblaciones es euclídea.

Demostración:

Ya se ha visto que $\mathbf{H}_i = \mathbf{H}_j = 2\,m$, siendo m el número de variables. Calculamos ahora

$$\begin{aligned}\mathbf{H}_{i\,j} = \\ &= \mathbf{E}_{\pi_i \times \pi_j}\left[\boldsymbol{\delta}^2\,(x,y)\right] \\ &= \mathbf{E}_{\pi_i \times \pi_j}\left((x-y)' \cdot \Sigma^{-1} \cdot (x-y)\right) \\ &= \mathbf{E}_{\pi_i \times \pi_j}\left(x' \cdot \Sigma^{-1} \cdot x + y' \cdot \Sigma^{-1} \cdot y - x' \cdot \Sigma^{-1} \cdot y - y' \cdot \Sigma^{-1} \cdot x\right)\end{aligned}$$

El primer sumando dentro del paréntesis es

$$\mathrm{tr}\,(\Sigma^{-1} \cdot x \cdot x')$$

por lo que su valor esperado es igual a

$$m + \mu_i \cdot \Sigma^{-1} \cdot \mu_i$$

Análogamente con el segundo sumando. Sustituyendo, resulta

$$\mathbf{H}_{i\,j} = 2\,m + (\mu_i - \mu_j)' \cdot \Sigma^{-1} \cdot (\mu_i - \mu_j)$$

y finalmente

$$\Delta_{i\,j} = (\mu_i - \mu_j)' \cdot \Sigma^{-1} \cdot (\mu_i - \mu_j)$$

□

La equivalencia entre Coordenadas Canónicas y Coordenadas Principales de la distancia de Mahalanobis se debe a Gower [49], y una generalización al caso de representación canónica de funciones paramétricas estimables puede verse en Cuadras [19].

En consecuencia, obtenemos un método de representación de poblaciones basado en distancias con las ventajas de los métodos DB (puede aplicarse sin hipótesis sobre la distribución de probabilidad de los datos y con variables mixtas), que puede considerarse una extensión del Análisis Canónico de Poblaciones clásico.

A diferencia de éste, no se dispone de regiones confidenciales para los centroides de las poblaciones. Sin embargo, según veremos en la sección 6.4 del Capítulo 6, la expresión (5.18) es especialmente adecuada para obtener por el método *bootstrap*, una estimación de la función de distribución de la distancia entre dos poblaciones, y de ella, intervalos de confianza para esta distancia.

Chapter 6

Aspectos computacionales y ejemplos

6.1 Consideraciones generales

En este capítulo se describen algunos aspectos de la implementación de los métodos DB.

De la descripción de estos métodos se puede observar que las implementaciones son, en principio, de programación simple en comparación con otros métodos de funcionalidad parecida, al no contener algoritmos iterativos ni dificultades especiales con matrices quasi–singulares.

Por ello no será necesario incluir listados completos de programas. En lugar de ello se discutirán especialmente algunos puntos relevantes y posiblemente no obvios de los algoritmos.

Los programas que implementan los métodos DB han sido publicados [5], y forman el núcleo del paquete de programas **MULTICUA** .

Quizás el aspecto más importante que cabe considerar es la necesidad de gran capacidad de almacenamiento.

En primer lugar, para mantener una matriz de distancias de orden igual al número de individuos. (Para poner un ejemplo, para 1500 individuos se requieren unos 9 Megabytes de almacenamiento, con números de punto flotante en doble precisión).

Este problema se presenta solamente en el caso de la Regresión DB o en la estimación *bootstrap* de la distribución de las distancias entre poblaciones, pues, como se verá en la sección 6.3, para el Análisis Discriminante DB no se necesitan en realidad los elementos individuales de la matriz de distancias.

Tambien se requiere considerable cantidad de memoria si se desea almacenar la matriz Y de dimensión (n, p) conteniendo las coordenadas de la muestra. Esta es la solución más eficiente en caso de ser posible, pues para calcular la matriz de distancias se debe acceder a todos los pares de individuos.

Para la mayor parte de los cálculos, el almacenamiento de Y no es un problema tan grave, al crecer sólo linealmente con n, pero si el número de variables o de individuos se hace muy grande, se debe recurrir a un fichero en un dispositivo externo. En este caso es totalmente imprescindible asignar dos *buffers* distintos al fichero para poder acceder rápidamente a pares de registros alejados físicamente.

6.2 Implementación del modelo de regresión DB

Este método presenta más dificultades que el discriminante. No por el tipo de cálculo a realizar, una diagonalización, que posiblemente sea el problema mejor estudiado en Análisis Numérico, sino por las dimensiones implicadas.

En primer lugar, se requiere la presencia individual de todos los elementos de la matriz de distancias o de la matriz B a diagonalizar, de igual dimensión. Además, se necesitan (al menos algunos de) los vectores propios de B.

Por ello, si se exige tener en memoria tanto la matriz a diagonalizar como los vectores propios, se restringe la aplicación a problemas muy pequeños (o máquinas muy grandes).

Se deduce la necesidad de implementar algoritmos de diagonalización *out–of–core*, sea del tipo Lanczos o del tipo Householder por bandas (véase Golub y Van Loan [46]). Es decir, se calcula la matriz de distancias, guardandose en un dispositivo externo. A partir de ella se obtiene la matriz B, que tambien se guarda en un dispositivo externo, y finalmente se calculan vectores propios sin que en ningún momento estas matrices existan completas en memoria.

Estos algoritmos de diagonalización son en principio parciales, es decir, obtienen solamente algunos de los valores y vectores propios.

Concretamente, en las implementaciones utilizadas, los Lanczos producen pares (valor propio/vector propio) según algún criterio preestablecido, por ejemplo, los correspondientes a los mayores valores propios, mientras que el algoritmo Householder por bandas produce todos los valores propios, y a partir de ellos se pueden obtener vectores propios según petición.

Aunque usualmente no se emplea el modelo global (con todas las coordenadas principales como variables regresoras) sino que, según se discute en [25], se calcula la regresión con un número de coordenadas principales pequeño en comparación con el número de individuos de la muestra, la selección de las coordenadas a utilizar como variables regresoras es según el criterio de mayor correlación con la variable dependiente.

En la práctica, esta condición impone calcular todos los vectores propios y por tanto, limita enormemente la elección de algoritmo. De los algoritmos *out–of–core* a que se ha tenido acceso, solamente el Householder por bandas cumple el requerimiento.

Actualmente está en estudio un criterio para seleccionar las variables regresoras teniendo en cuenta solamente los valores propios. Si este criterio se muestra eficaz, permitirá emplear los algoritmos Lanczos, mucho más eficientes.

6.3 Implementación del Análisis Discriminante DB

Como se ha mencionado en la nota 5.2.1, en la exposición del modelo DB de Análisis Discriminante se han usado dos subpoblaciones sólo por comodidad de notación. Naturalmente, la implementación de este método se ha realizado para un número arbitrario k de subpoblaciones, limitado por la capacidad del ordenador.

En este apartado se discuten algunos detalles del algoritmo, con especial énfasis en las necesidades de memoria.

Si se desea solamente asignar nuevas observaciones a una de k subpoblaciones, es suficiente calcular inicialmente las sumas de distancias internas de cada subpoblación, y asignar secuencialmente cada nueva observación calculando el vector de distancias a las n observaciones de las k muestras. Por tanto, en este caso es suficiente asignar memoria para este vector.

Si se desea hacer una estimación de la probabilidad de error de asignación por el método *leave–one–out* no sería eficiente realizar todo el proceso anterior para cada individuo, pues ello equivaldría a repetir muchas veces el cálculo de cada distancia entre cada par.

El caso más sencillo aparece cuando se elige una función global como distancia entre individuos, tal como la distancia valor absoluto, o la de Gower, de manera que se puede calcular una matriz global de distancias al cuadrado D para el conjunto de las dos subpoblaciones, que se puede considerar subdividida en cajas

$$D = (D_{\alpha\,\beta}) \qquad 1 \leq \alpha,\, \beta \leq k$$

siendo $D_{\alpha\,\beta}$ de dimensión (n_α, n_β).

Si el número total de individuos es pequeño, de forma que se pueda almacenar la matriz D completa (es decir, el triángulo superior) en memoria, el cálculo correspondiente a eliminar el individuo i y asignarlo según las funciones discriminantes obtenidas de los restantes equivale a extraer la fila (columna) $d(i)$ correspondiente a este individuo en D y calcular las sumas

$$a_i(\alpha) = \sum_j d(i)_j$$

donde j recorre los índices de la subpoblación α. A continuación se obtienen la funciones discriminantes como en las fórmulas (5.8) y (5.9).

Esto no es práctico si el número de individuos es grande de modo que la matriz D no pueda mantenerse en memoria. Obsérvese además que aparte del aumento de tiempo de cálculo debido a los accesos a disco o cinta, este algoritmo obligaría a almacenar la matriz desplegada (es decir, los dos triangulos) en el dispositivo externo, de modo que el acceso a las filas individuales

no sea extremadamente ineficiente, como ocurriría de tener sólo la mitad de ella, al quedar los elementos contiguos de una misma fila en posiciones distantes entre sí, según se muestra en el diagrama

$$
\begin{array}{cccccc}
1 & \underline{2} & 4 & 7 & 11 & \ldots \\
 & \underline{3} & \underline{5} & \underline{8} & \underline{12} & \ldots \\
 & & 6 & 9 & 13 & \ldots \\
 & & & 10 & 14 & \ldots \\
 & & & & 15 & \ldots \\
 & & & & \ldots &
\end{array}
$$

y por esta misma razón, se tendría que calcular dos veces la distancia entre cada par de individuos, una para la posición $(i,\, j)$ y otra para la $(j,\, i)$.

La solución a este problema consiste en no almacenar la matriz D sino solamente una matriz DIP de dimensión $(n,\, k)$ definida por

$$
\begin{array}{rcl}
DIP(i,\, \alpha) & = & \text{Suma de las distancias del individuo } i \\
 & & \text{a todos los individuos de la población } \alpha
\end{array}
$$

que podemos suponer que cabe en memoria en todos los casos.

A partir de ella se obtiene, para cada α, la suma $D(\alpha)$ de todas las distancias internas de la población α, sumando los elementos pertenecientes a dicha población de la columna α.

Los elementos $a_i(\alpha)$ necesarios para la estimación del error de asignación son los elementos de la fila i de DIP. [Estos equivalen a los a_i, b_i en la notación para dos poblaciones de la sección 5.2.2 del Capítulo 5].

Un último detalle a tener en cuenta para calcular la matriz DIP es que es suficiente calcular una sola vez la distancia entre cada par de individuos, empleando el siguiente algoritmo:

1. Inicializar DIP a 0
2. Recorrer secuencialmente la lista de los $\frac{1}{2}\, n\, (n-1)$ pares $(i,\, j)$ de individuos distintos, calculando la distancia al cuadrado $d(i, j)$
3. Sumar $d(i, j)$ a las posiciones $\mathrm{DIP}(i, p(j))$ y $\mathrm{DIP}(j, p(i))$, siendo

$$
p(i) = \text{ Población a que pertenece } i
$$

Si no existe una distancia global entre todos los individuos, se puede modificar el método anterior para este caso. Obsérvese, sin embargo, que se requerirán k matrices del tipo DIP.

Finalmente, para el caso paramétrico, en que las coordenadas de los individuos se reemplazan por los *efficient scores*, no se emplea tampoco la matriz de distancias, ni una matriz DIP, pues al ser bilineal la fórmula de la distancia entre dos individuos, las fórmulas para las funciones discriminantes se simplifican, dando lugar a expresiones parecidas a las del discriminador cuadrático, como en el caso de poblaciones con distribución normal multivariante con Σ conocida (Proposición 5.4.5 del Capítulo 5).

La mayor dificultad aparece para hacer una estimación de la probabilidad de error por el método *leave–one–out*, pues al eliminar un individuo de una sub–muestra se deben recalcular para ésta las estimaciones máximo–verosímiles de los parámetros, los *efficient scores* y en general también la matriz de la métrica y su inversa. Posiblemente en este caso sería más asequible algún otro estimador de la probabilidad de error, como el basado en *bootstrap*.

6.4 Estimación *bootstrap* de la distancia entre poblaciones

Se parte de una muestra de n individuos pertenecientes a k poblaciones $\pi_1, \ldots, \pi_k$, con vector $(n_1, \ldots, n_k)$ de número de individuos en cada población conocido y constante.

Suponemos dada una función distancia global, que permite llegar a una matriz $\mathbf{\Delta}^{(2)}$ de dimensión (n, n) de distancias al cuadrado entre los individuos. A partir de ella se obtiene una matriz $\widehat{\Delta}$ de dimensión (k, k) de distancias entre las poblaciones de la diferencia de Jensen

$$\widehat{\Delta}_{\alpha\,\beta} = \widehat{\mathbf{H}}_{\alpha\,\beta} - \frac{1}{2}\,(\widehat{\mathbf{H}}_{\alpha\,\alpha} + \widehat{\mathbf{H}}_{\beta\,\beta})$$

donde

$$\widehat{\mathbf{H}}_{\alpha\,\beta} = \frac{1}{n_\alpha\, n_\beta} \sum_{i \in I_\alpha} \sum_{j \in I_\beta} \boldsymbol{\delta}^2_{i\,j} \qquad (1 \leq \alpha, \beta \leq k)$$

y I_α es el subconjunto de $[1, n]$ correspondiente a los indices de los individuos de π_α.

Se propone el problema de realizar, por el método *bootstrap*, una estimación de las funciones de distribución de los elementos de la matriz $\widehat{\Delta}$.

En este apartado se exponen los puntos relevantes del algoritmo empleado para este cálculo, y en especial, la técnica empleada para sortear la dificultad de las grandes dimensiones implicadas.

El método consiste en realizar una sucesión de B remuestreos del total de n individuos, cada uno de ellos obtenido por concatenación de remuestreos realizados en cada población por separado. Se mantienen constantes los números n_α. Dentro de cada población, un remuestreo viene determinado por una permutación (con repetición) del conjunto de individuos que la forman.

Para cada remuestreo se evaluan las matrices $\mathbf{\Delta}^{(2)}$ y $\widehat{\Delta}$. Finalmente se obtiene la función de distribución empírica para cada elemento de $\widehat{\Delta}$.

Dado que cada remuestreo consiste en individuos de la muestra original, todos los elementos de la nueva matriz $\mathbf{\Delta}^{(2)}$ ya existen en alguna posición de la inicial, por lo que parece superfluo recalcularlos, y más razonable leerlos simplemente de esta posición.

La gran dimensión de $\mathbf{\Delta}^{(2)}$ fuerza a guardarla en un dispositivo externo, preferentemente de tipo secuencial por razones de eficiencia. Como el elemento $(i,\, j)$ de la nueva matriz $\mathbf{\Delta}^{(2)}$, correspondiente al remuestreo dado por la permutación p, se encuentra en el lugar $(p_i,\, p_j)$ de la matriz inicial, la aplicación de este algoritmo exige un dispositivo de acceso aleatorio (es decir un fichero en disco), y además, daría lugar a una actividad de la cabeza lectora que destruiría físicamente la unidad en breve plazo.

La solución que se ha encontrado se basa en el hecho que la ordenación de individuos dentro de cada población no influye en el resultado (la matriz $\widehat{\Delta}$). Por tanto podemos representar cada remuestreo b en cada población π_α como un vector de multiplicidades $\mu_{\alpha\, b}$ con n_α elementos.

El vector de multiplicidades correspondiente a la muestra original es el

$$(1, \ldots, 1)$$

Los sucesivos remuestreos se generan como vectores $\mu_{\alpha\, b}$ cuyos elementos tienen una distribución multinomial de dimensión n_α, suma de frecuencias n_α, y vector de probabilidades

$$\left(\frac{1}{n_\alpha}, \ldots, \frac{1}{n_\alpha}\right)$$

Veamos que a partir de los vectores de multiplicidades es posible calcular los $\widehat{\mathbf{H}}_{\alpha\,\beta}$ correspondientes al remuestreo b con solamente una sola lectura secuencial de los $n\,(n-1)/2$ elementos de $\mathbf{\Delta}^{(2)}$. El proceso es como sigue:

1) Se inicializan a 0 los $\widehat{\mathbf{H}}_{\alpha\,\beta}$.

2) Se recorren por orden los elementos de $\mathbf{\Delta}^{(2)}$. El elemento $(i,\, j)$ debe ser sumado a $\widehat{\mathbf{H}}_{\alpha\,\beta}$ con una multiplicidad $m\,(i,\, j)$ que se calcula por

$$m(i,\, j) = \begin{cases} \mu_{\alpha\, i_\alpha}\, \mu_{\beta\, j_\beta} & \text{si} \quad \alpha \neq \beta \\ \mu^2_{\alpha\, i_\alpha} & \text{si} \quad \alpha = \beta \quad \text{y} \quad i = j \\ 2\, \mu_{\alpha\, i_\alpha}\, \mu_{\alpha\, j_\alpha} & \text{si} \quad \alpha = \beta \quad \text{y} \quad i \neq j \end{cases}$$

siendo i_α un índice que varía de 1 a n_α y da el número de orden dentro de π_α del individuo i–ésimo de la muestra conjunta. La relación entre los dos índices es

$$i = \sum_{\beta < \alpha} n_\beta + i_\alpha$$

De esta manera cada remuestreo requiere una sola lectura de la matriz de distancias, siendo el tiempo de cálculo empleado muy razonable. Por ejemplo, en un ordenador personal con procesador 80486 a 33MHz, un cálculo con 1000 remuestreos sobre una matriz de distancias de $n = 150$ individuos (el ejemplo de la sección siguiente), ha empleado 2 minutos y 10 segundos.

Puede verse además que este algoritmo es susceptible de ulterior optimización, al ser fácilmente vectorizable, realizando simultáneamente varios remuestreos (es decir calculando varios vectores de multiplicidades) para cada lectura de la matriz $\mathbf{\Delta}^{(2)}$.

6.5 Ejemplos de aplicación de modelos DB

6.5.1 Regresión DB

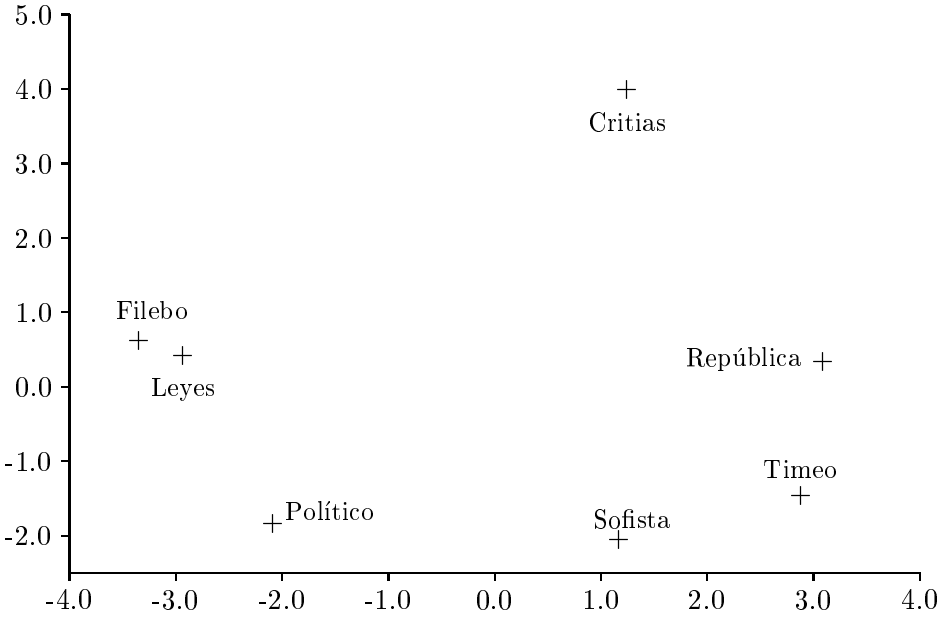
Como ilustración del método de regresión DB trataremos un problema de ordenación temporal de diálogos de Platón citado por Mardia et al. [78, pág. 313] y procedente del trabajo de Cox y Brandwood [16]. Estos autores emplean técnicas de Análisis Discriminante (método ML) para proponer una solución.

El problema consiste en ordenar cronológicamente los siete diálogos *La República*, *Las Leyes*, *Critias*, *Filebo*, *El Político*, *El Sofista* y *Timeo*, de los cuales se conoce solamente que *La República* es el primero, y *Las Leyes* es el último.

El análisis debe basarse en medidas del estilo. Concretamente, se dispone de una tabla que registra para cada obra las frecuencias de aparición como cinco últimas sílabas de una frase de cada uno de las $2^5 = 32$ posibles combinaciones de sílabas largas y cortas. Se supone que estas medidas evolucionan con el tiempo y que de su estudio se puede deducir la ordenación cronológica de las obras.

Hemos empleado para este análisis el modelo de regresión DB, con distancia valor absoluto, tomando sucesivamente como valores de la variable dependiente cada permutación de los números 1 a 7, (con los 1 y 7 fijados a *República* y *Leyes*, respectivamente, según las especificaciones dadas).

Se toma el coeficiente de determinación como medida de adecuación entre una ordenación dada y las 32 medidas de que se dispone para cada obra.



La figura es la representación gráfica de las dos primeras Coordenadas Principales obtenidas de los datos con la distancia Valor Absoluto. Este análisis proporciona los siguientes porcentajes acumulados de variabilidad para los cuatro primeros ejes principales

34.11 54.13 69.61 81.91

La figura, junto con la consideración de los porcentajes de variabilidad, muestran que la aproximación unidimensional es muy deficiente, y de hecho muestran que para tener un buen ajuste a los datos, se requieren al menos tres o cuatro ejes.

Por ello se ha registrado, para cada una de las 120 permutaciones, los coeficientes de determinación que corresponden a tomar como variables regresoras uno, dos, tres y cuatro de los ejes principales.

Se han ordenado en cada caso las permutaciones según el orden de los coeficientes de determinación obtenidos. En la tabla siguiente se muestra la zona que contiene las permutaciones que dan los valores más grandes.

	Permutación	Coef. Determ.
Un eje		
	3 6 5 4 2	0.9112
	4 6 5 3 2	0.9071
	3 5 6 4 2	0.8447
	4 5 6 3 2	0.8407
Dos ejes		
	3 5 6 4 2	0.9606
	3 6 5 4 2	0.9587
	4 6 5 3 2	0.9372
	4 5 6 3 2	0.9283
Tres ejes		
	5 2 4 6 3	0.9923
	3 6 2 4 5	0.9899
	3 5 6 4 2	0.9815
	4 3 2 5 6	0.9792
Cuatro ejes		
	4 2 5 6 3	0.9994
	3 6 5 2 4	0.9989
	2 6 3 4 5	0.9988
	5 2 4 6 3	0.9972

La solución propuesta por Cox y Brandwood es la ordenación *Timeo*, *Sofista*, *Critias*, *Político*, y *Filebo*, lo que corresponde en nuestra notación a la permutación 4 6 5 3 2.

Observamos en nuestra tabla que encontramos esta ordenación en caso de tomar regresión sobre el primer eje principal (de hecho aparece en segunda posición, con poca diferencia de la 3 6 5 4 2, que aparece en primer lugar.

Como hemos visto que el primer eje principal explica solamente un porcentaje de variabilidad del 34.11%, podemos afirmar que el método de Cox y Branwood es una aproximación *lineal* a la ordenación, y proponemos la ordenación 4 2 5 6 3 como más verosímil, teniendo en cuenta que para su obtención se han tomado en cuenta cuatro ejes principales, que explican un 81.91% de la variabilidad de los datos.

Nota: Cox y Brandwood hacen constar en su trabajo que la ordenación obtenida por ellos no coincide con la mantenida por la mayoría de estudiosos, pero no citan en su artículo cual es dicha ordenación.

6.5.2 Análisis Discriminante DB

El siguiente ejemplo, para el Análisis Discriminante DB, es el (casi obligado para cualquier método discriminante) estudio de los datos *Iris* de Fisher [42]. Consisten en medidas de las cuatro variables longitud de sépalo, anchura de sépalo, longitud de pétalo y anchura de pétalo para $n = 150$ individuos de *Iris*, repartidos en tres grupos procedentes de las tres especies *Iris setosa*, *Iris versicolor*, *Iris virginica*, con $n_1 = n_2 = n_3 = 50$ individuos en cada grupo.

Se ha empleado la distancia Valor Absoluto, y se ha calculado por el método *leave–one–out* la matriz de asignación, es decir la matriz de dimensión (3, 3) que contiene en el lugar (i, j) el número de elementos del grupo i que han sido asignados al grupo j por el algoritmo discriminante. El cociente entre la suma de elementos no diagonales de esta matriz y el total de individuos es la estimación de la probabilidad de error de asignación. Se obtiene el siguiente resultado

$$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 48 & 2 \\ 0 & 3 & 47 \end{pmatrix}$$

con estimación $\widehat{p} = 0.0333$ de la probabilidad de asignación errónea. Para comparación, con el discriminador lineal (LDF) se obtiene

$$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 48 & 2 \\ 0 & 1 & 49 \end{pmatrix}$$

con estimación $\widehat{p} = 0.0200$ de la probabilidad de asignación errónea, y con el discriminador cuadrático (QDF) la matriz de asignación es

$$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 1 & 49 \end{pmatrix}$$

y la estimación de la probabilidad de error de asignación es $\widehat{p} = 0.0267$.

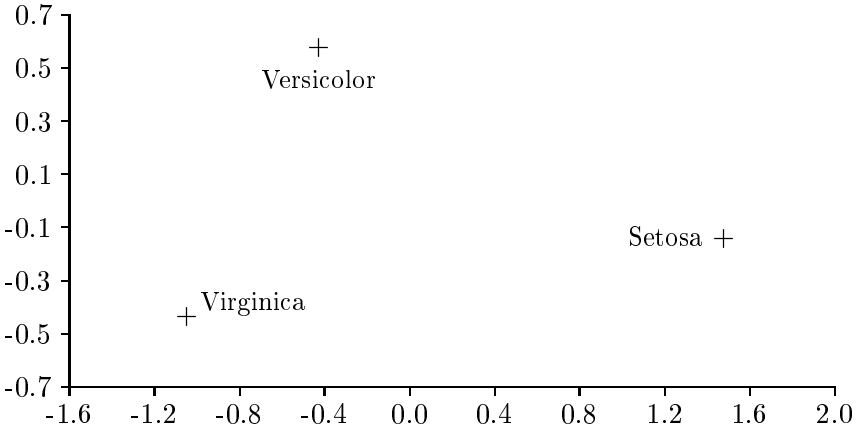
La matriz de de distancias entre los grupos, calculada por la diferencia de Jensen (5.18) a partir de la matriz de distancias entre los individuos es

$$\begin{pmatrix} 0 & 4.1495 & 6.4756 \\ & 0 & 1.4022 \\ & & 0 \end{pmatrix}$$

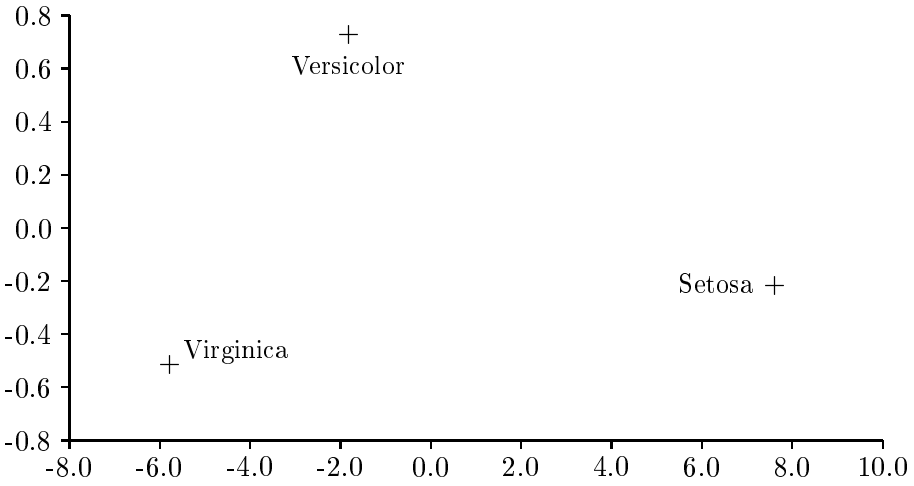
Un Análisis de Coordenadas Principales realizado sobre esta matriz, lleva a la siguiente configuración euclídea bidimensional con tres puntos que representan a los tres grupos

$$\begin{pmatrix} 1.4783 & -.14115 \\ -.42876 & .57477 \\ -1.0496 & -.43362 \end{pmatrix}$$

siendo 86.573 el porcentaje de variabilidad correspondiente al primer eje principal. La representación gráfica de estos puntos es



Compárese este diagrama con el siguiente, obtenido a partir de los mismos datos pero esta vez mediante un Análisis Canónico de Poblaciones.



La coincidencia es más notable si se toma en consideración que el primer diagrama procede de un cálculo no paramétrico, libre de cualquier hipótesis sobre la distribución seguida por los datos, y que por tanto, es igualmente aplicable cuando el Análisis Canónico de Poblaciones no lo es.

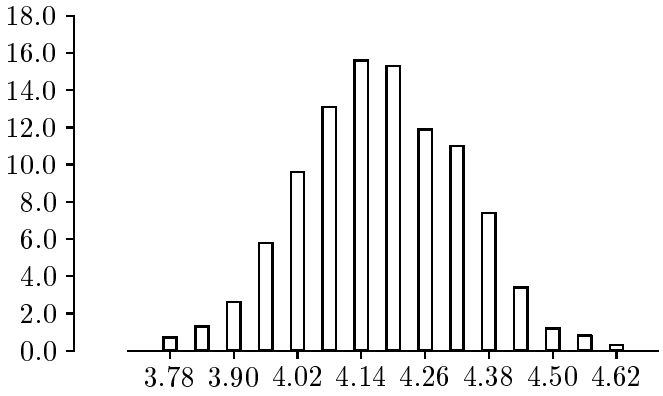
Se ha realizado una estimación *bootstrap* de las distribuciones de probabilidad de los elementos de la matriz de de distancias entre los grupos, empleando $B = 1000$ remuestras de los datos originales.

Los valores medios y desviaciones típicas de los valores obtenidos para los elementos $\widehat{\Delta}_{1\,2}$ $\widehat{\Delta}_{1\,3}$ y $\widehat{\Delta}_{2\,3}$ son

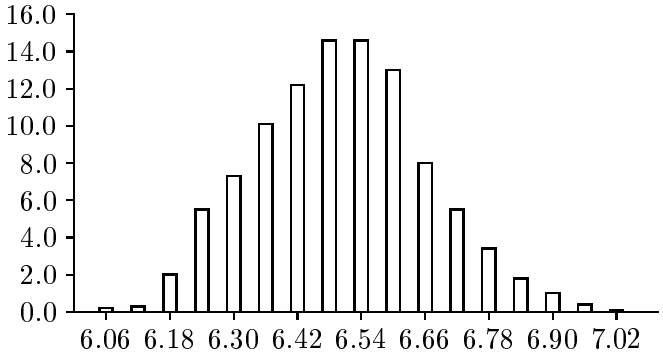
	Media	Desv. Típica
$\widehat{\Delta}_{1\,2}$	4.1775	0.1502
$\widehat{\Delta}_{1\,3}$	6.5022	0.1609
$\widehat{\Delta}_{2\,3}$	1.4348	0.1775

(6.1)

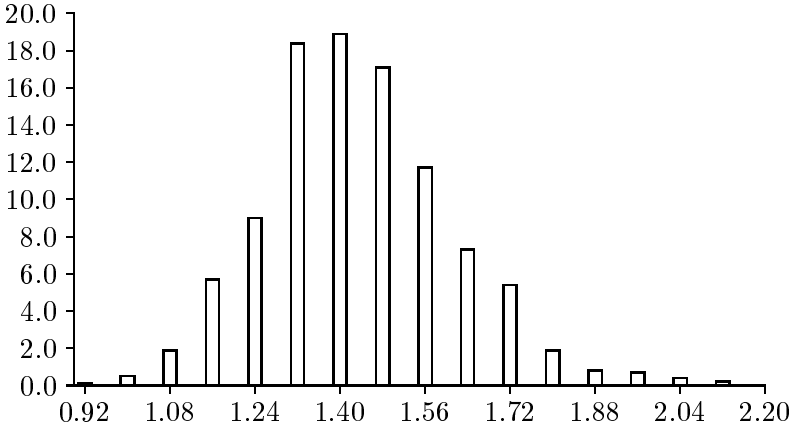
En las figuras siguientes se representan los diagramas de frecuencias para estos elementos de matriz.



$\widehat{\Delta}_{1\,2}$



$\widehat{\Delta}_{1\,3}$



$\widehat{\Delta}_{2\,3}$

Se han calculado los coeficientes ρ^+ descritos en la ecuación (4.36) del Capítulo 4, para tener una medida de la proximidad de las distribuciones empíricas obtenidas a una distribución normal. Los resultados son

	ρ^+ (Normal)
$\widehat{\Delta}_{1\,2}$	0.99914
$\widehat{\Delta}_{1\,3}$	0.99846
$\widehat{\Delta}_{2\,3}$	0.99365

lo que indica proximidad de las distribuciones empíricas a la distribución normal.

Para verificar esta proximidad calculamos los cuatro primeros de los coeficientes β_j descritos en la ecuación (4.37) del Capítulo 4. Los resultados se tabulan a continuación.

En primer lugar, se calculan los β_j correspondientes a los datos originales (primera columna). La segunda columna contiene los valores teóricos (obtenidos por integración numérica) para la distribución normal.

En segundo lugar, se aplica a cada tabla de valores $\widehat{\Delta}_{i\,j}$ la transformación $y = F(x)$, siendo F la distribución normal con parámetros iguales a los empíricos (6.1). Como resultado deberíamos obtener una distribución uniforme, de ser cierta la hipótesis de normalidad. La tercera columna contiene los β_j para los datos transformados, y finalmente en la cuarta columna se reproducen los valores teóricos para la distribución uniforme, calculados de la ecuación (4.38).

		Datos Originales	Teóricos Normal	Datos Transformados	Teóricos Uniforme
$\widehat{\Delta}_{1\,2}$					
	β_1	0.9540	0.9484	0.9937	0.9927
	β_2	0.0093	0	0.0111	0
	β_3	0.2293	0.2407	0.1014	0.1103
	β_4	0.0084	0	0.0090	0
$\widehat{\Delta}_{1\,3}$					
	β_1	0.9523	0.9484	0.9924	0.9927
	β_2	0.0113	0	0.0023	0
	β_3	0.2416	0.2407	0.1149	0.1103
	β_4	0.0307	0	0.0172	0
$\widehat{\Delta}_{2\,3}$					
	β_1	0.9326	0.9484	0.9878	0.9927
	β_2	0.0113	0	0.0591	0
	β_3	0.2627	0.2407	0.1334	0.1103
	β_4	0.0283	0	0.0079	0

Conclusiones

En esta memoria se han presentado algunas aportaciones al estudio de los Métodos de Análisis Multivariante Basados en Distancias introducidos por Cuadras.

1) Se ha realizado un estudio teórico de las propiedades de la función distancia Valor Absoluto para su empleo en dichos métodos, analizando los sistemas de Coordenadas Principales asociados a esta distancia para un sistema unidimensional de puntos equidistantes, obteniendo el teorema (2.4.1), que da una justificación del buen comportamiento de esta función distancia al interpretar los ejes principales (es decir, las variables regresoras en el modelo DB) como funciones polinómicas de grado creciente, por lo que la regresión basada en esta distancia equivale a una regresión no lineal.

2) Motivado por este estudio, se ha estudiado la estructura de valores y vectores propios de una familia de matrices de relevantes propiedades algebraicas y combinatorias.

3) Se ha generalizado a una configuración unidimensional arbitraria de puntos el estudio realizado para el caso equidistante, llegando a la proposición (4.1.2)), que permite una interpretación cualitativa de los ejes principales análoga a la obtenida en el caso equidistante.

4) Se ha encontrado una técnica, basada en la descomposición de Procesos Estocásticos en Componentes Principales, para generalizar a variables aleatorias continuas el concepto de Coordenadas Principales

5) Empleando dicha técnica, se ha encontrado una sucesión de variables aleatorias que, según el teorema (4.2.1), pueden justificadamente llamarse Coordenadas Principales de la distribución Uniforme respecto la Distancia Valor Absoluto, y que son el análogo en el caso contínuo de los ejes principales para un conjunto unidimensional de puntos.

6) Generalizando la Regresión Basada en distancias a este caso continuo, se propone una medida de bondad de ajuste entre funciones de distribución, con aplicación al problema de decidir la distribución seguida por una variable a partir de una muestra.

7) Se han estudiado algunas propiedades del Análisis Discriminante basado en distancias, llegando en particular a los teoremas (5.2.1) y (5.2.2) que explican la regla de asignación DB como un método de mínima distancia en el espacio (euclídeo real o complejo) de las coordenadas principales.

8) Se han calculado las funciones discriminantes para variables aleatorias que siguen algunas distribuciones conocidas, obteniendo en particular, la proposición (5.4.2), que muestra que para una distribución multinomial, las funciones discriminantes DB coinciden con el tradicional estadístico Ji-cuadrado.

9) Por aplicación de Multidimensional Scaling sobre la matriz de distancias entre poblaciones obtenida a partir de distancias entre individuos por

medio de (5.18), se llega a una técnica no paramétrica análoga al Análisis Canónico de Poblaciones que produce resultados equivalentes cuando éste es aplicable.

10) Se han implementado los algoritmos DB para las distancias más comunes en los programas REGD y DISC, que forman el núcleo del paquete publicado de programas **MULTICUA** .

Bibliography

[1] **A.A. Afifi and S. P. Azen**,
Statistical Analysis. A computer oriented approach (second edition).
Academic Press (1977)

[2] **T. W. Anderson and D. A. Darling**,
Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes.
Ann. Math. Statist. **23** (1952), pp. 193–212.

[3] **T. W. Anderson and D. A. Darling**,
A test of goodness of fit.
J. Amer. Statist. Ass. **49** (1954), pp. 765–769.

[4] **T. W. Anderson**,
An Introduction to Multivariate Statistical Analysis.
John Wiley & Sons, New York (1958)

[5] **C. Arenas, C. M. Cuadras and J. Fortiana**,
MULTICUA : *Paquete no standard de Análisis Multivariante.*
Publicacions del Departament d'Estadistica **4**.
Universitat de Barcelona, (1991)

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

[7] **Y. Ayant and M. Borg**,
Funciones Especiales.
Alhambra, Madrid (1974)

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

[9] **P. Buneman**,
The recovery of trees from measures of dissimilarity.
in [57, pp. 387–395] (1971)

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

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

[12] **T. Cacoullos, (ed.)**,
Discriminant Analysis and Applications.
Academic Press, New York (1973)

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

[14] **M. S. Cord and R. J. Sylvester**,
The property of Cross–Symmetry.
J. Soc. Indust. Appl. Math. **10** (1962), pp. 632–637.

[15] **R. Courant and D. Hilbert**,
Methods of Mathematical Physics.
Wiley–Interscience. John Wiley & Sons, New York (1953).

[16] **D. R. Cox and L. Brandwood**,
On a discriminatory problem connected with the works of Plato.
J. Roy. Statist. Soc. B **21** (1959), pp. 195–200.

[17] **F. Critchley**,
Hierarchical trees can be perfectly scaled in one dimension.
Journal of Classification **5** (1988), pp. 5–20.

[18] **F. Critchley**,
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 (1989).

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

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

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

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

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

[24] **C. M. Cuadras**,
Distance Analysis in discrimination and classification using both continuous and categorical variables.
in [37, pp. 459–473].

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

[26] **C. M. Cuadras**,
An eigenvector pattern arising in nonlinear regression.
Qüestió **14** (1990), pp. 89–95.

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

[28] **C. M. Cuadras**,
Some aspects of distance based discrimination.
Biometrical Letters **29** (1992), pp. (in press).

[29] **C. M. Cuadras**,
Probability distributions with given multivariate marginals and given dependence structure.
J. of Multivariate Analysis **41** (1992), pp. (in press).

[30] **C. M. Cuadras and J. Fortiana**,
Maximum correlation between random variables and some applications.
7th. Int. Conf. on Mult. Analysis, Penn State Univ. (1992), Report.

[31] **C. M. Cuadras**,
Using suitable distances for solving some statistical problems.
DISTANCIA '92 (Rennes) Report, (1992).

[32] **J. De Leeuw, W. Heiser, J. Meulman and F. Critchley, (eds.)**,
Multidimensional Data Analysis.
DSWO Press, Leiden (1986).

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

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

[35] **E. Diday**,
Orders and overlapping clusters in pyramids.
in [32, pp. 201–234] (1986).

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

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

[38] **J. Durbin and M. Knott**,
Components of Cramér–von Mises Statistics. I.
J. Roy. Stat. Soc. B **34** (1972), pp. 290–307.

[39] **J. Durbin, M. Knott and C. C. Taylor**,
Components of Cramér–von Mises Statistics. II.
J. Roy. Stat. Soc. B **37** (1975), pp. 216–237.

[40] **B. Efron**,
The efficiency of logistic regression compared to normal discriminant analysis.
J. Am. Stat. Assoc. **70** (1975), pp. 892–898.

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

[42] **R. A. Fisher**,
The use of multiple measurements in taxonomic problems.
Ann. Eugen. **7** (1936), pp. 179–188.

[43] **W. L. Frank**,
Computing eigenvalues of complex matrices by determinant evaluation and by methods of Danilevski and Wielandt.
J. Soc. Indust. Appl. Math. **6** (1958), pp. 378–392.

[44] **M. Fréchet**,
Sur les tableaux de correlation dont les marges sont données.
Ann. Univ. Lyon, Section A, Series 3 **14** (1951), pp. 53–77.

[45] **F. R. Gantmacher**,
Théorie des Matrices.
Dunod, Paris (1966)

[46] **G. H. Golub and C. F. Van Loan**,
Matrix Computations (Second Edition).
The Johns Hopkins University Press, Baltimore and London, (1989)

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

[48] **I. J. Good**,
The Inverse of a Centrosymmetric Matrix.
Technometrics **12** (1970), pp. 925–928.

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

[50] **J. C. Gower**,
Adding a point to vector diagrams in Multivariate Analysis.
Biometrika **55** (1968), pp. 582–585.

[51] **J. C. Gower**,
Statistical methods of comparing different multivariate analysis of the same data.
in [57, pp. 138–149] (1971)

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

[53] **F. A. Graybill**,
Matrices with application in Statistics.
Wadsworth, Belmont, California (1983)

[54] **P. J. Green**,
Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives.
J. Royal Stat. Soc. Series B **46** (1984), pp. 149–192.

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

[56] **W. Hoeffding**,
Massstabinvariante Korrelationtheorie.
Schriften Math. Inst. Univ. Berlin **5** (1940), pp. 181–233.

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

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

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

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

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

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

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

[64] **M. G. Kendall**,
A Course in Multivariate Analysis.
Charles Griffin, London (1957)

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

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

[67] **W. J. Krzanowski**,
Discrimination and classification using both binary and continuous variables.
J. Am. Stat. Assoc. **70** (1975), pp. 782–790.

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

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

[70] **W. J. Krzanowski**,
Principles of Multivariate Analysis: a user's perspective.
Clarendon Press, Oxford (1988)

[71] **P. A. Lachenbruch**,
Discriminant Analysis.
Hafner, New York (1975)

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

[73] **V. R. Marco, D. M. Young and D. W. Turner**,
The Euclidean distance classifier: an alternative to linear discriminant function.
Commun. Statist.–Simula. **16** (1987), pp. 485–505.

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

[75] **K. Matusita**,
Decision rule, based on distance for the classification problem.
Ann. Inst. Stat. Math. **8** (1956), pp. 67–77.

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

[77] **K. Matusita**,
Discrimination and the affinity of distributions.
in: [12, pp. 213–223]

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

[79] **A. Miñarro**,
Aspectos geométricos de las poblaciones y los individuos estadísticos.
Tesis doctoral. Dep. Estadística, Univ. de Barcelona (1991)

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

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

[82] **J. M. Oller and C. M. Cuadras**,
Defined distances for some probability distributions.

in Proc. 2nd World Conf. Math. at the Serv. of Man (1982), pp. 563–565.

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

[84] **J. M. Oller**,
Some geometrical aspects of Data Analysis and Statistics.
in [37, pp. 41–58].

[85] **W. C. Parr**,
Minimum distance method.
in Encyclopedia of Statistical Sciences

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

[87] **S. J. Press and S. Wilson**,
Choosing between logistic regression and discriminant analysis.
J. Am. Stat. Assoc. **73** (1978), pp. 699–705.

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

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

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

[91] **D. A. Ratkowsky**,
Nonlinear Regression Modeling.
Marcel Dekker, New York (1983)

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

[93] **W. S. Robinson**,
A method for chronologically ordering archaeological deposits.
Am. Antiq. **16** (1951), pp. 293–301.

[94] **M. M. Royall**,
Model robust confidence intervals using maximum likelihood estimators.
International Statistical Review **54** (1986), pp. 221–226.

[95] **S. Sattah and A. Tversky**,
Additive similarity trees.
Psychometrika **42** (1977), pp. 319–345.

[96] **G. A. F. Seber**,
Multivariate Observations.
John Wiley & Sons (1984).

[97] **G. A. F. Seber and C.J. Wild**,
Nonlinear Regression.
John Wiley & Sons (1989).

[98] **G. R. Shorack and J. A. Wellner**,
Empirical processes with applications to Statistics.
John Wiley & Sons (1986).

[99] **W. S. Torgerson**,
Theory and methods of scaling.
John Wiley & Sons (1958).

[100] **R. Valdés**,
Finding string distances.
Dr. Dobb's Journal **17(4)** (1992), pp. 56–62.

[101] **H. Weyl**,
On the volume of tubes.
Am. J. of Math. **61** (1939), pp. 461–472.

[102] **J. Wolfowitz**,
The Minimum Distance Method.
Ann. Math. Stat. **28** (1957), pp. 75–88.