

# METODOS MULTIVARIANTES BASADOS EN DISTANCIAS

C.M. Cuadras

Enero 2007

**Abstract**

Curso de doctorado sobre Análisis Multivariante y Regresión basados en el concepto y propiedades de las distancias.

# Índice

<b>1</b>	<b>Distancias, similaridades y aplicaciones</b>	<b>5</b>
1.1	Introducción . . . . .	5
1.2	Distancias . . . . .	5
1.3	Similaridades en general . . . . .	7
1.4	Distancias para variables cuantitativas . . . . .	7
1.5	Similaridades y distancias con variables binarias . . . . .	8
1.6	Similaridad con variables mixtas . . . . .	9
1.7	Otras similaridades y distancias . . . . .	10
1.8	Teorema de caracterización . . . . .	11
1.9	La fórmula de añadir un punto . . . . .	13
1.10	Análisis canónico de poblaciones . . . . .	14
1.11	Análisis de coordenadas principales y biplot . . . . .	14
<b>2</b>	<b>Regresión y predicción DB</b>	<b>17</b>
2.1	El modelo de regresión lineal . . . . .	17
2.2	Regresión DB en dimensión reducida . . . . .	20
2.3	Predicción DB sobre un nuevo individuo . . . . .	21
2.4	Predicción con variables continuas, categóricas y mixtas . . . . .	22
2.5	Regresión no lineal DB . . . . .	23
<b>3</b>	<b>Aspectos computacionales en predicción DB</b>	<b>29</b>
3.1	Selección de variables en regresión DB . . . . .	29
3.2	Selección para un número grande de individuos . . . . .	30
3.3	El caso de tener valores propios negativos o muy pequeños . . . . .	31
<b>4</b>	<b>Análisis discriminante DB</b>	<b>33</b>
4.1	Introducción . . . . .	33
4.2	La función de proximidad de un individuo a una población . . . . .	35
4.3	La regla discriminante DB . . . . .	36
4.4	Propiedades de la función de proximidad . . . . .	37
4.5	La regla DB comparada con algunas reglas clásicas . . . . .	38
4.6	La regla DB en el caso de muestras . . . . .	40
4.7	Ventajas del método DB . . . . .	42
4.8	Discriminación en el caso de varias poblaciones . . . . .	43

<b>5</b>	<b>Asociación DB y predicción multivariante</b>	<b>45</b>
5.1	Relacionando dos conjuntos de variables . . . . .	45
5.2	Relación DB entre variables mixtas . . . . .	47
5.3	Predicción con correlaciones canónicas . . . . .	48
5.4	Planteamiento mediante related MDS . . . . .	50
<b>6</b>	<b>Comparación DB de poblaciones</b>	<b>53</b>
6.1	Comparando dos conjuntos de datos mixtos . . . . .	53
6.2	Planteamiento mediante coordenadas principales . . . . .	54
6.3	Planteamiento mediante funciones de proximidad . . . . .	56
6.3.1	Prueba ji-cuadrado . . . . .	56
6.3.2	Prueba de Mann-Whitney . . . . .	57
6.4	Comparando muestras múltiples . . . . .	57
6.4.1	Partición de la variabilidad geométrica . . . . .	58
6.4.2	Tests con coordenadas principales . . . . .	59
6.4.3	Tests con funciones de proximidad . . . . .	59
<b>7</b>	<b>Distintividad</b>	<b>61</b>
7.1	Planteamiento bajo normalidad . . . . .	61
7.2	Planteamiento DB . . . . .	62
7.3	Planteamiento mediante la razón de proximidades . . . . .	63

# Capítulo 1

## Distancias, similaridades y aplicaciones

### 1.1 Introducción

Muchos métodos de estadística y análisis de datos utilizan el concepto geométrico de distancia entre individuos, entre poblaciones, y de un individuo a una población. Esto es especialmente cierto en técnicas de representación de datos (análisis de correspondencias, análisis de coordenadas principales, análisis de proximidades), donde la distancia, entendida como medida de diferenciación entre objetos, constituye la base fundamental de la presentación de los resultados.

Las distancias, aparecen también en muchos otros aspectos de la estadística: contraste de hipótesis, estimación, regresión, análisis discriminante, etc. En este curso aprenderemos cómo utilizar metodología basada en distancias para abordar estas partes de la estadística.

### 1.2 Distancias

Una distancia  $\delta$  sobre un conjunto (finito o no)  $\Omega$  es una aplicación que a cada par de individuos  $(\omega_i, \omega_j) \in \Omega \times \Omega$ , le hace corresponder un número real  $\delta(\omega_i, \omega_j) = \delta_{ij}$ , que cumple con las siguientes propiedades básicas:

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

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

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

Cuando, además, se cumple la desigualdad triangular

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

diremos que la distancia es métrica.

Si  $\Omega$  es un conjunto finito, que indicaremos por  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ , las distancias  $\delta_{ij}$  se expresan mediante la matriz simétrica  $\Delta$ , llamada matriz de

distancias sobre  $\Omega$ :

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2n} \\ \dots & \dots & \dots & \dots \\ \delta_{n1} & \delta_{n2} & \dots & \delta_{nn} \end{pmatrix} \quad \delta_{ii} = 0, \quad \delta_{ij} = \delta_{ji}$$

Se llama *preordenación* de  $\Omega$  asociada a  $\Delta$ , a la ordenación de menor a mayor de los  $m = n \times (n - 1)/2$  pares de distancias no nulas:

$$\delta_{i_1 j_1} \leq \delta_{i_2 j_2} \leq \dots \leq \delta_{i_m j_m},$$

es decir, la ordenación de los pares  $(\omega_i, \omega_j)$  de  $\Omega$ , de acuerdo con su proximidad.

**Ejercicio 1.1** La matriz de distancias genéticas entre los géneros humano ( $H$ ), chimpancé ( $Ch$ ), gorila ( $Go$ ), orangután ( $O$ ) y gibón ( $Gi$ ) es:

	$H$	$Ch$	$Go$	$O$	$Gi$
$H$	0	0.094	0.111	0.180	0.207
$Ch$		0	0.115	0.194	0.218
$Go$			0	0.188	0.288
$O$				0	0.216
$Gi$					0

Cada distancia mide el número de sustituciones nucleótidas en el DNA mitocondrial. Verifica si se cumple la propiedad métrica y escribe la preordenación asociada.

Una matriz de distancias  $\Delta$  puede ser transformada de diversos modos. Por ejemplo:

$$\delta_{ij}^* = \begin{cases} 0 & i = j \\ \delta_{ij} + c & i \neq j \end{cases} \quad (1.1)$$

La transformación (1.1), que consiste en sumar una constante fuera de la diagonal de  $\Delta$ , se llama **aditiva**. Otra transformación es:

$$\tilde{\delta}_{ij}^2 = \begin{cases} 0 & i = j \\ \delta_{ij}^2 + c & i \neq j \end{cases} \quad (1.2)$$

que afecta al cuadrado de la distancia y la que llamaremos **q-aditiva**. Las transformaciones (1.1) y (1.2) son útiles para conseguir que la nueva distancia cumpla propiedades que la distancia original no posee, pero conservando la preordenación, es decir, las relaciones de proximidad entre los individuos.

**Ejercicio 1.2** Sea  $\Delta = (\delta_{ij})$  una matriz de distancias  $n \times n$  sobre un conjunto finito  $\Omega$ .

1. Demuestra que las transformaciones (1.1) y (1.2) conservan la preordenación de  $\Omega$ .
2. Suponiendo que  $\Delta$  no es métrica, prueba que la transformación aditiva tomando

$$c = \max \{ \delta_{ij} - \delta_{ik} - \delta_{jk} \} \text{ para cada } i, j, k \in \Omega$$

convierte  $\Delta$  en  $\Delta^* = (\delta_{ij}^*)$  que sí posee la propiedad métrica.

### 1.3 Similaridades en general

En muchas aplicaciones es conveniente trabajar con similaridades, concepto dual al de distancias. Una similaridad  $s$  en un conjunto  $\Omega$ , es una aplicación que asigna a cada par  $(\omega_i, \omega_j) \in \Omega \times \Omega$  un número real  $s_{ij} = s(i, j)$ , que cumple:

$$\text{S.1.} \quad 0 \leq s_{ij} \leq s_{ii} = 1.$$

$$\text{S.2.} \quad s_{ij} = s_{ji}.$$

Cuando  $\Omega$  es un conjunto finito, entonces la matriz

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ s_{n1} & s_{n2} & \dots & s_{nn} \end{pmatrix}$$

se denomina matriz de similaridades sobre  $\Omega$ .

Es inmediato pasar de similaridad a distancia y recíprocamente. Las dos transformaciones básicas son:

$$\delta_{ij} = 1 - s_{ij} \quad (1.3)$$

$$\delta_{ij} = \sqrt{1 - s_{ij}} \quad (1.4)$$

En general, una matriz de similaridades puede tener en su diagonal elementos  $s_{ii} \neq 1$ . La transformación que nos permite pasar de similaridad a distancia es entonces:

$$\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}} \quad (1.5)$$

Por diversas razones, que justificaremos más adelante, (1.4) es preferible a (1.3). En general, (1.5) es la transformación más apropiada (ver Ejercicio 1.9).

### 1.4 Distancias para variables cuantitativas

Supongamos ahora que cada individuo de  $\Omega$  puede ser representado por un punto  $\mathbf{x} = (x_1, x_2, \dots, x_p) \in R^p$ . Algunas distancias especialmente interesantes entre dos puntos  $\mathbf{x}, \mathbf{y} \in R^p$ , son:

a) la distancia Euclídea,

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}, \quad (1.6)$$

b) la distancia “ciudad”

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p |x_i - y_i|, \quad (1.7)$$

c) la distancia “valor absoluto”

$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^p |x_i - y_i|}. \quad (1.8)$$

Cuando  $\Omega$  se puede asimilar a una población normal multivariante  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , con  $\boldsymbol{\Sigma}$  no singular, la distancia estadística (al cuadrado) más apropiada es

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}) \quad (1.9)$$

llamada distancia de Mahalanobis. Naturalmente, esta distancia puede ser definida en poblaciones  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , es decir, con vector de medias  $\boldsymbol{\mu}$  y matriz de covariancias  $\boldsymbol{\Sigma}$ , sin necesidad de asumir normalidad. Véase (1.20).

**Ejercicio 1.3** Comprueba que la distancia  $d_E$  puede ser escrita como

$$d_E^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})' (\mathbf{x} - \mathbf{y})$$

de manera que  $d_E$  es un caso particular de  $d_M$ . Especifica la matriz de covarianzas  $\boldsymbol{\Sigma}$ .

**Ejercicio 1.4** Sean  $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ ,  $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$  dos poblaciones normales multivariantes. El discriminador lineal de Fisher, para asignar  $\mathbf{x} \in R^p$  a una de las dos poblaciones es

$$L(\mathbf{x}) = \left[ \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right]' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Expresa  $L(\mathbf{x})$  como la diferencia entre las distancias de Mahalanobis de  $\mathbf{x}$  a  $\boldsymbol{\mu}_1$  y a  $\boldsymbol{\mu}_2$ .

## 1.5 Similaridades y distancias con variables binarias

Supongamos que tenemos  $p$  variables binarias  $X_1, X_2, \dots, X_p$ , donde cada  $X_i$  toma los valores 0 ó 1 según la presencia de una cierta característica. Son bien conocidos los siguientes coeficientes de similaridad entre cada par de individuos  $\omega_i, \omega_j$ :



$$s_{ij} = \frac{a + d}{p} \quad (\text{Sokal} - \text{Michener}) \quad (1.10)$$

$$s_{ij} = \frac{a}{a + b + c} \quad (\text{Jaccard}) \quad (1.11)$$

siendo  $a, b, c, d$  las frecuencias de (1,1), (1,0), (0,1) y (0,0), respectivamente. Nótese que  $p = a + b + c + d$ . Estas similaridades pueden ser transformadas en distancias utilizando (1.3) o preferentemente (1.4).

**Ejercicio 1.5** *Cinco herramientas cortantes arqueológicas A, B, C, D, E han sido encontradas en un yacimiento. Estaban fabricadas con Piedra, Bronce y Hierro, según la matriz de incidencias*

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

*Calcula las matrices de similaridad de Sokal-Michener y de Jaccard.*

**Ejercicio 1.6** *Sea  $\mathbf{X}$  la matriz  $n \times p$  con los datos binarios de  $n$  objetos respecto a  $p$  características. Sea  $\mathbf{J}$  la matriz  $n \times p$  formada por unos.*

1. *Demuestra que la matriz de similaridades de Sokal-Michener puede expresarse como*

$$\mathbf{S} = [\mathbf{X}\mathbf{X}' + (\mathbf{J} - \mathbf{X})(\mathbf{J} - \mathbf{X})'] / p$$

2. *Intenta encontrar una expresión parecida para la similaridad de Jaccard*

## 1.6 Similaridad con variables mixtas

Si las variables son mixtas, continuas, binarias o cualitativas, es entonces adecuado utilizar la distancia de Gower,  $d_{ij}^2 = 1 - s_{ij}$ , siendo

$$s_{ij} = \left( \sum_{h=1}^{p_1} (1 - |x_{ih} - x_{jh}| G_h) + a + \alpha \right) / (p_1 + (p_2 - d) + p_3) \quad (1.12)$$

una similaridad, donde  $p_1$  es el número de variables cuantitativas,  $a$  y  $d$  corresponden al número de coincidencias y no coincidencias para las  $p_2$  variables binarias, respectivamente, y  $\alpha$  es el número de coincidencias para las  $p_3$  variables cuantitativas.  $G_h$  es el rango de la  $h$ -ésima variable cuantitativa. Este coeficiente admite la posibilidad de tratar datos faltantes y se reduce al coeficiente de Jaccard (1.11) cuando  $p_1 = p_3 = 0$ .

**Ejercicio 1.7** Para la siguiente tabla de datos, obtenidos sobre 10 individuos, calcula la matriz de similaridades de Gower

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<b>1</b>	180	76	1	1	0	0	<b>6</b>	181	72	3	2	1	0
<b>2</b>	174	67	1	3	1	1	<b>7</b>	171	60	3	2	0	0
<b>3</b>	174	68	2	3	1	1	<b>8</b>	162	58	1	3	1	1
<b>4</b>	170	64	2	2	0	1	<b>9</b>	170	66	2	2	0	1
<b>5</b>	177	70	1	3	0	1	<b>10</b>	179	81	1	1	1	0

*A*= talla en cm., *B*= peso en kg., *C*= color ojos (1 azul, 2 verde-gris, 3 castaño), *D*= color cabello (1 rubio, 2 castaño, 3 oscuro), *E*= gafas (1 si, 0 no), *F*= vestimenta (1 clásica, 0 moderna).

## 1.7 Otras similaridades y distancias

Si disponemos de una densidad de probabilidad  $f(x_1, \dots, x_p)$ , podemos definir distancias entre poblaciones siguiendo un camino no heurístico.

Supongamos que  $f(\mathbf{x}) = f(x_1, \dots, x_p)$  pertenece a un modelo estadístico regular  $\{f(\mathbf{x}, \theta), \theta \in \Theta\}$ . Consideremos el vector columna aleatorio:

$$\mathbf{Z} = \frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta).$$

Entonces, la llamada matriz de información de Fisher es el valor esperado

$$\mathbf{F} = E(\mathbf{Z}\mathbf{Z}'),$$

y una definición de distancia (al cuadrado) entre  $\mathbf{x} = (x_1, \dots, x_p)$ ,  $\mathbf{y} = (y_1, \dots, y_p)$ , que generaliza (1.8), es la distancia de Rao

$$d_R^2(\mathbf{x}, \mathbf{y}) = (\mathbf{z}_x - \mathbf{z}_y)' \mathbf{F}^{-1} (\mathbf{z}_x - \mathbf{z}_y)$$

Esta distancia (propuesta por Cuadras, 1989 y Oller, 1989), depende de  $\theta$ . Véase también Miñarro y Oller (1992).

**Ejercicio 1.8** Sea  $X$  una variable aleatoria. Encuentra  $d_R$  cuando:

1.  $X$  es exponencial de parámetro  $\alpha$ .
2.  $X$  es Poisson de parámetro  $\lambda$ .
3.  $\mathbf{X}$  es un vector aleatorio  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , donde  $\boldsymbol{\Sigma}$  es una matriz constante.

## 1.8 Teorema de caracterización

Sea  $\Delta = (\delta_{ij})$  una matriz  $n \times n$  de distancias Euclídeas. Es decir, existe una configuración de puntos  $\mathbf{x}_1, \dots, \mathbf{x}_n \in R^m$ , tales que

$$\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j).$$

¿Cómo se puede saber si una matriz de distancias es Euclídea? El siguiente teorema nos proporciona un criterio, que es a la vez condición necesaria y suficiente.

Sea  $\mathbf{1}_n$  el vector columna que contiene unos. Entonces  $\mathbf{J} = \mathbf{1}_n \mathbf{1}_n'$  es una matriz  $n \times n$  también de unos. Sea  $\mathbf{H} = \mathbf{I}_n - \mathbf{J}/n$  la matriz de centrado,  $\mathbf{A} = (a_{ij})$  la matriz con elementos  $a_{ij} = -\delta_{ij}^2/2$  y calculemos  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$ . Recordemos que una matriz simétrica es semidefinida positiva si todos sus valores propios son no negativos.

**Teorema 1.1** *La matriz de distancias  $\Delta$  es Euclídea en dimensión  $m$  si y sólo si  $\mathbf{B} \geq \mathbf{0}$  ( $\mathbf{B}$  es semidefinida positiva) y el rango de  $\mathbf{B}$  es  $\text{rang}(\mathbf{B}) = m \leq n - 1$ .*

Hagamos ahora explícito este teorema. Si  $\Delta$  es matriz de distancias Euclídeas, es entonces posible obtener la descomposición espectral

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}' = \mathbf{X}\mathbf{X}', \quad (1.13)$$

donde  $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}$  contiene los  $m$  vectores propios de  $\mathbf{B}$ ,  $\mathbf{\Lambda}$  es una matriz diagonal que contiene los valores propios ordenados  $\lambda_1 > \dots > \lambda_m > 0$ . La matriz  $\mathbf{B}$  proporciona las coordenadas Euclídeas del conjunto  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ . Cada fila  $\mathbf{x}_i$  de  $\mathbf{X}$  contiene las coordenadas, llamadas *coordenadas principales* del individuo  $i$ .

Las coordenadas principales tienen interesantes propiedades:

1. Las filas  $\mathbf{x}_1, \dots, \mathbf{x}_n$  de  $\mathbf{X}$  verifican  $\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$ , es decir, sus distancias Euclídeas se igualan a los términos  $\delta_{ij}$  en  $\Delta$ .
2. Las columnas  $X_1, \dots, X_m$  de  $\mathbf{X}$ , entendidas como variables, tienen media 0.
3. Cada columna  $X_j$  de  $\mathbf{X}$  tiene varianza igual a  $\lambda_j/n$ .
4. Las columnas de  $\mathbf{X}$  son ortogonales (incorrelacionadas).
5. Las columnas de  $\mathbf{X}$  pueden ser interpretadas como componentes principales.
6. La representación de los elementos  $\omega_1, \omega_2, \dots, \omega_n$  utilizando las filas de  $\mathbf{X}$  es óptima.

Por otra parte, si indicamos por  $\delta_{ij}(k)$  la distancia utilizando las  $k < m$  primeras coordenadas, entonces la cantidad

$$\sum_{i,j=1}^n \delta_{ij}^2(k) = 2n(\lambda_1 + \dots + \lambda_k) \quad (1.14)$$

es máxima en dimensión reducida  $k$ , siendo  $\lambda_1 > \dots > \lambda_k$  los  $k$  primeros valores propios de  $\mathbf{B}$ , ordenados de mayor a menor.

**Ejercicio 1.9** Se pide:

1. Analiza si la matriz de distancias del Ejercicio 1.1 es Euclídea.
2. Demuestra las cuatro primeras propiedades de las coordenadas principales.
3. Demuestra que si  $\mathbf{S} = (s_{ij})$  es una matriz de similaridades tal que  $\mathbf{S} \geq 0$ , entonces la distancia  $\delta_{ij}$  tal que

$$\delta_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij} \quad (1.15)$$

es Euclídea.

4. Observa que salvo un factor constante, (1.4) es un caso particular de (1.5). Comprueba que (1.3) no proporciona, en general, una distancia Euclídea.
5. A partir de la matriz de similaridades del Ejercicio 1.5, representa las 5 herramientas en dimensión 2. Interpreta la primera dimensión.

**Ejercicio 1.10** Se pide:

1. Demuestra que si las distancias  $\delta_1(.,.), \delta_2(.,.)$  definidas sobre un mismo  $\Omega$  son Euclídeas, entonces la distancia  $\delta$  tal que

$$\delta^2(.,.) = \delta_1^2(.,.) + \delta_2^2(.,.) \quad (1.16)$$

es también Euclídea.

2. Utiliza (1.16) para probar que la distancia “valor absoluto” (1.8) es Euclídea.

Cuando la matriz de distancias  $\Delta$  no es Euclídea, sus elementos deben transformarse para proporcionar una distancia Euclídea, pero conservando la preordenación de  $\Omega$ . La transformación puede ser no lineal (obtenida numéricamente) o algebraica. El siguiente teorema proporciona dos transformaciones algebraicas.

**Teorema 1.2** Sea  $\Delta = (\delta_{ij})$  una matriz de distancias no Euclídeas. Entonces  $\mathbf{B}$  tendrá valores propios positivos y negativos:  $\lambda_1 > \dots > \lambda_k > 0 > \lambda'_1 > \dots > \lambda'_k$ , con  $k + k' = n - 1$ . Se verifica:

1. La transformación  $q$ -aditiva con  $c \geq -2\lambda'_k$ , convierte  $\Delta$  en  $\tilde{\Delta}$  Euclídea.
2. La transformación aditiva con  $c \geq \lambda$ , donde  $\lambda$  es el mayor valor propio de la matriz no simétrica

$$\begin{pmatrix} \mathbf{0} & 2\mathbf{B} \\ -\mathbf{I} & -4\mathbf{B}_r \end{pmatrix}$$

siendo  $\mathbf{B}$  la matriz asociada a  $\Delta$  y  $\mathbf{B}_r$  la matriz asociada a  $\Delta_r = (\sqrt{d_{ij}})$ , convierte  $\Delta$  en  $\Delta^*$  Euclídea.

**Ejercicio 1.11** Sea  $\Delta = (\delta_{ij})$  una matriz de distancias no Euclídeas.

1. Demuestra el primer apartado del Teorema 1.2.
2. Demuestra el segundo apartado del Teorema 1.2.

## 1.9 La fórmula de añadir un punto

Supongamos que  $\Delta = (\delta_{ij})$  es una matriz de distancias Euclídeas. Indiquemos por  $\omega_{n+1}$  un nuevo individuo. Supongamos conocidas las distancias entre  $\omega_{n+1}$  y  $\omega_1, \omega_2, \dots, \omega_n$ :

$$\delta_j = \delta(\omega_j, \omega_{n+1}), \quad \omega_j \in \Omega \quad (1.17)$$

Si cada  $\omega_j$  está representado por el punto  $\mathbf{x}_j \in R^m$ ,  $\mathbf{X}$  es la matriz cuyas filas son  $\mathbf{x}'_j$ , y además  $\omega_{n+1}$  viene representado por  $\mathbf{x} \in R^m$ , las coordenadas de  $\mathbf{x}$  (escritas como un vector columna) se dan a continuación.

**Teorema 1.3** Las coordenadas  $\mathbf{x} \in R^m$  de  $\omega_{n+1}$  en función de  $\mathbf{X}$  y de las distancias (1.17) es

$$\mathbf{x} = \frac{1}{2} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (\mathbf{b} - \mathbf{d}) \quad (1.18)$$

donde  $\mathbf{d} = (\delta_1^2, \dots, \delta_n^2)'$ , y  $\mathbf{b} = (b_{11}, \dots, b_{nn})'$  son los vectores columna que contienen las distancias (al cuadrado) y la diagonal de  $\mathbf{B}$ , respectivamente. En particular, si  $\mathbf{X}$  es la matriz que contiene las coordenadas principales, entonces

$$\mathbf{x} = \frac{1}{2} \mathbf{\Lambda}^{-1} \mathbf{X}' (\mathbf{b} - \mathbf{d}) \quad (1.19)$$

Las fórmulas (1.18) y (1.19) son debidas a Gower (1968). Veremos aplicaciones en la Sección 1.11 y en el Capítulo 2.

**Ejercicio 1.12** La matriz de distancias (Tabla 1.1) entre 6 bebidas refrescantes, se obtuvo a partir de la valoración dada por 38 estudiantes, siguiendo una escala (1= no similar, 9= muy similar). Las similaridades acumuladas se transformaron en distancias (disimilaridades).

1. Verifica si esta matriz de distancias es Euclídea o no.
2. Representa las seis bebidas en dimensión 2 a lo largo de los dos primeros ejes principales.
3. Una nueva bebida refrescante LC mantiene las siguientes distancias con las demás:

	PC	CC	CCC	DPC	DtUP	7-UP
LC	220	265	244	210	99	85

Sitúa LC, utilizando (1.19), dentro de la representación anterior de las seis bebidas ya conocidas.

TABLE 1.1

	PC	CC	CCC	DPC	D7UP	7-UP
Pepsi Cola	0					
Coca Cola	127	0				
Classic CC	167	143	0			
Diet Pepsi	207	235	243	0		
Diet 7-UP	320	322	327	288	0	
Seven Up	321	318	318	317	136	0

## 1.10 Análisis canónico de poblaciones

Supongamos que tenemos  $g > 2$  poblaciones  $\Omega_1, \dots, \Omega_g$ , representadas por los vectores de medias  $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_g$  y la matriz de variancias y covariancias (que se supone común)  $\boldsymbol{\Sigma}$ , en relación a  $p$  variables cuantitativas. Suponiendo que se dispone de una matriz de datos para cada población, el **análisis canónico de poblaciones** es un método multivariante que permite representar las  $g$  poblaciones en dimensión reducida, usualmente en una representación bidimensional. Una exposición del tema resultaría algo extensa (ver Cuadras, 1991), por lo que nos limitaremos a decir que, esencialmente, es equivalente a un análisis de coordenadas principales utilizando la distancia de Mahalanobis entre vectores de medias

$$d_M^2(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \quad (1.20)$$

La representación, a lo largo de los llamados ejes canónicos, incluye además regiones confidenciales para los vectores de medias..

**Ejercicio 1.13** Utilizando los datos del fichero *cranis.dat*, que corresponden a 4 variables y 5 poblaciones, y el programa *canp06s.exe*, realiza un análisis canónico de poblaciones. Interpreta la primera dimensión canónica.

## 1.11 Análisis de coordenadas principales y biplot

Dada una matriz de datos cuantitativos  $\mathbf{Y}$ , de orden  $n \times p$ , el método biplot (Gabriel, 1971) es una representación sobre el mismo gráfico de las  $n$  filas (= individuos) y  $p$  columnas (= variables) de  $\mathbf{Y}$ . Usualmente los individuos se representan como puntos y las variables como vectores. La presentación conjunta Biplot puede ser obtenida por una descomposición singular de  $\mathbf{Y}$ , pero Gower y Harding (1988) probaron que también se puede obtener una solución biplot partiendo de (1.19).

Supongamos que deseamos representar la primera variable  $Y_1$ , es decir la primera columna de  $\mathbf{Y}$ . Podemos identificar  $Y_1$  con el conjunto de coordenadas

$$\alpha_1 \mathbf{u}_1 = (\alpha_1, 0, \dots, 0) \quad \alpha_1 \in R_1$$

siendo  $R_1$  el recorrido de  $Y_1$ . La distancia al cuadrado entre  $\alpha_1 \mathbf{u}_1$  y la fila  $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})$  de  $\mathbf{Y}$  es

$$(\alpha_1 - y_{i1})^2 + y_{i2}^2 + \dots + y_{ip}^2$$

Observando que  $b_{ii} = x_{i1}^2 + \dots + x_{ip}^2 = y_{i1}^2 + \dots + y_{ip}^2$ , resulta que

$$\mathbf{b} - \mathbf{d} = 2\alpha_1 Y_1 - \alpha_1^2 \mathbf{1}_n$$

y, debido a que  $\mathbf{X}'\mathbf{1}_n = \mathbf{0}$ , tenemos que las coordenadas representando al eje  $Y_1$  son  $\mathbf{x}_1(\alpha_1) = \alpha_1 \Lambda^{-1} \mathbf{X}'\mathbf{Y}_1$

En general, los  $p$  ejes vendrán representados por el haz de segmentos

$$\mathbf{X}(\alpha) = \Lambda^{-1} \mathbf{X}'\mathbf{Y}\mathbf{D}_\alpha \quad \alpha_i \in R_i$$

donde  $\mathbf{X}$  es la matriz con las coordenadas principales y  $\mathbf{D}_\alpha = \text{diag}(\alpha_1, \dots, \alpha_p)$ .

**Ejercicio 1.14** La puntuación, en una escala de 1 a 10, que 8 ciudadanos dan a 5 políticos  $Zp$ ,  $Az$ ,  $Go$ ,  $Pu$ ,  $Fr$  es la siguiente:

	$Zp$	$Az$	$Go$	$Pu$	$Fr$
1	6	3	5	4	3
2	7	2	5	5	2
3	1	8	3	6	7
4	2	5	5	4	6
5	6	5	4	7	3
6	1	8	2	6	7
7	2	7	3	4	8
8	8	2	6	4	3

Centrando primero la matriz de datos por columnas, represéntela mediante un biplot.





## Capítulo 2

# Regresion y predicción DB

### 2.1 El modelo de regresión lineal

Consideremos el modelo lineal que relaciona una variable respuesta con diversas variables explicativas:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \quad (2.1)$$

Aquí  $\mathbf{y}(n \times 1)$  es un vector (conocido) con  $n$  observaciones de una variable respuesta cuantitativa  $Y$ , la matriz  $\mathbf{X}(n \times m)$  es conocida de  $\text{rang}(\mathbf{X}) = m$ ,  $\boldsymbol{\beta}(m \times 1)$  es un vector (desconocido) de parámetros y  $\mathbf{e} = (e_1, \dots, e_n)'$  es un vector aleatorio tal que

$$\begin{aligned} E(e_i) &= 0 & \text{var}(e_i) &= \sigma^2 & i &= 1, \dots, n \\ E(e_i e_j) &= 0 & i &\neq j \end{aligned}$$

En muchas ocasiones se cumple que  $\mathbf{e}$  es  $N_n(0, \sigma^2 \mathbf{I}_n)$  y entonces se dice que (2.1) es un modelo lineal normal.

Recordemos que, la estimación LS (mínimos cuadrados) de  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)'$  es

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}, \quad (2.2)$$

la suma de cuadrados residual es

$$R_0^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{e}}'\hat{\mathbf{e}}, \quad (2.3)$$

y una estimación insesgada de  $\sigma^2$  es

$$\hat{\sigma}^2 = R_0^2 / (n - m). \quad (2.4)$$

**Ejercicio 2.1** *Consideremos el modelo lineal*

$$7 = \beta_1 + \beta_2 + e_1; \quad 12 = 2\beta_1 + \beta_2 + e_2;$$

$$2 = \beta_1 - \beta_2 + e_3; \quad 14 = \beta_1 + 3\beta_2 + e_4.$$

Escribe este modelo en la forma matricial (2.1) y estima los parámetros  $\beta = (\beta_1, \beta_2)$  y  $\sigma^2$ .

A continuación vamos a interpretar el modelo lineal desde la perspectiva de las distancias. Indiquemos por  $\mathbf{x}_1, \dots, \mathbf{x}_n$  las filas de  $\mathbf{X}$ . La distancia Euclídea cuadrática entre cada par de individuos  $\omega_i, \omega_j$  es

$$d_E^2(\omega_i, \omega_j) = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)' \quad (2.5)$$

Sobre la matriz de distancias Euclídeas

$$\mathbf{D} = (d_E(\omega_i, \omega_j))$$

podemos aplicar el Teorema 1.1 de la Sección 1.8. Construimos  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$  y hallamos la descomposición espectral

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}' = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'. \quad (2.6)$$

Resulta entonces que las distancias entre las filas  $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n$  de  $\tilde{\mathbf{X}}$ , reproducen exactamente las distancias (2.5). Podemos, en consecuencia, reescribir el modelo (2.1) como

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\gamma} + \mathbf{e} \quad (2.7)$$

donde  $\beta \rightarrow \gamma$  es una reparametrización.

Lo que importa aquí es notar que (2.7) ha sido construido utilizando *sólo* distancias (en este caso la distancia Euclídea), y le llamaremos modelo DB (distance-based).

El vector proyector de  $\mathbf{y}$  del modelo (2.1) es

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2.8)$$

Análogamente podríamos encontrar  $\hat{\tilde{\mathbf{y}}}$  para el modelo (2.7). Pero ambos modelos son esencialmente el mismo, puesto que

$$\hat{\mathbf{y}} = \hat{\tilde{\mathbf{y}}} \quad (2.9)$$

Como veremos enseguida, el modelo DB presenta ventajas con otras distancias.

**Ejercicio 2.2** Se pide:

1. Encuentra la transformación  $\gamma = \mathbf{T}\beta$ .
2. Demuestra (2.9).

La distancia Euclídea (2.5) ha sido calculada utilizando las filas de la matriz de diseño  $\mathbf{X}$  que es conocida. Vamos a introducir el modelo de regresión DB (distance based) en general.

Supongamos que tenemos  $p$  variables  $W_1, W_2, \dots, W_p$  observables, de tipo continuo, binario o categórico, o incluso los tres tipos a la vez, en cuyo caso diremos que los datos son **mixtos**. Sea  $d(\omega_i, \omega_j)$  una distancia adecuada entre pares  $\omega_i, \omega_j$  de individuos. Si los datos son binarios  $d(\omega_i, \omega_j)$  se puede basar en (1.10) ó (1.11), y si son mixtos en el coeficiente de similitud de Gower (1.12). Supongamos que la distancia tiene la propiedad de ser Euclídea. A partir de  $d(\omega_i, \omega_j)$  se puede obtener la matriz  $n \times n$  de distancias  $\mathbf{\Delta}$ , y aplicando la descomposición espectral (1.13), obtendremos la matriz  $\mathbf{X}$ , de coordenadas principales que reproducen las distancias originales. El modelo DB que proponemos es entonces

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (2.10)$$

donde  $\mathbf{1}$  es el vector de unos, mientras que  $\mathbf{X}$ ,  $\boldsymbol{\beta}$  y  $\mathbf{e}$  tienen el mismo significado que en el modelo (2.1). Observemos que, como  $\mathbf{B}\mathbf{1} = \mathbf{0}$ , tanto  $\mathbf{1}$  como las columnas  $X_1, \dots, X_m$  de  $\mathbf{X}$ , son vectores propios de  $\mathbf{B}$ .

Podemos también escribir

$$\mathbf{y} = \beta_0 \mathbf{1} + \sum_{i=1}^m \beta_i X_i + \mathbf{e} \quad (2.11)$$

donde  $m = \text{rang}(\mathbf{B})$  y  $X_1, \dots, X_m$ , juegan el papel de **variables predictoras**.

Las propiedades básicas del modelo DB son:

1. Las estimaciones de los parámetros de regresión son

$$\hat{\beta}_0 = \bar{y} = \mathbf{y}'\mathbf{1}/n \quad \hat{\beta}_i = \mathbf{y}'X_i/\lambda_i \quad (2.12)$$

2. El vector predictor o proyección ortogonal de  $\mathbf{y}$  es

$$\hat{\mathbf{y}} = \bar{y}\mathbf{1} + \mathbf{X}\mathbf{\Lambda}^{-1}\mathbf{X}'\mathbf{y} \quad (2.13)$$

3. El coeficiente de correlación simple  $r_i = r(\mathbf{y}, X_i)$  es

$$r_i^2 = (\mathbf{y}'X_i) / ns_y^2 \lambda_i \quad (2.14)$$

donde  $s_y^2$  es la variancia muestral de  $\mathbf{y}$ .

4. El coeficiente de correlación múltiple (al cuadrado)  $R^2$  entre  $\mathbf{y}$  y  $X_1, \dots, X_m$  es

$$R^2 = \mathbf{y}'\mathbf{X}\mathbf{\Lambda}^{-1}\mathbf{X}'\mathbf{y} / ns_y^2 = \sum_{i=1}^m r_i^2, \quad (2.15)$$

donde  $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_m)$  contiene los valores propios de  $\mathbf{B}$ .

**Ejercicio 2.3** Demuestra las fórmulas (2.12) a (2.15).

## 2.2 Regresión DB en dimensión reducida

El modelo DB (2.10-2.11), es el **modelo completo**. A veces, sin embargo,  $m = \text{rango}(\mathbf{B})$  crece con  $n$  (incluso puede darse el caso en que  $m = n - 1$ ). Entonces, el número de variables  $X_1, \dots, X_m$  (las columnas de  $\mathbf{X}$ ) puede resultar excesivo y de esa manera encontrarnos con un coeficiente de determinación  $R^2$  arbitrariamente próximo a 1. Para evitar este problema, es conveniente partir  $\mathbf{X}$  en dos partes

$$\mathbf{X} = (\mathbf{X}_{(k)}, \mathbf{Z}) \quad (2.16)$$

donde  $\mathbf{X}_{(k)} = (X_1, \dots, X_k)$  contiene  $k$  columnas adecuadas de  $\mathbf{X}$ , y la matriz  $\mathbf{Z}$  contiene las restantes columnas.

Definimos de este modo el **modelo DB en dimensión  $k$** , que puede ser expresado de dos maneras equivalentes:

$$\begin{aligned} \mathbf{y} &= \beta_0 \mathbf{1} + \mathbf{X}_{(k)} \beta_{(k)} + \mathbf{e}_k, \\ \mathbf{y} &= \beta_0 \mathbf{1} + \sum_{i=1}^k X_i \beta_i + \mathbf{e}_k. \end{aligned} \quad (2.17)$$

Como valor de  $k$ , se puede tomar  $k =$  número inicial de variables observables explicativas.

Una buena selección de las columnas  $X_1, \dots, X_k$  de  $\mathbf{X}$  consiste en escogerlas por orden de correlación con  $\mathbf{y}$ , es decir,

$$r(\mathbf{y}, X_1) > r(\mathbf{y}, X_2) > \dots > r(\mathbf{y}, X_k). \quad (2.18)$$

Otra selección obvia consiste en ordenarlas de acuerdo con la variabilidad explicada por las variables predictoras (columnas de  $\mathbf{X}$ ):  $\lambda_1 > \dots > \lambda_k$ , es decir, seleccionar los  $k$  primeros ejes principales. Pero si resultara que la variable  $\mathbf{X}_{k+1}$  tiene una correlación  $r_{k+1} = r(\mathbf{y}, X_{k+1})$  relativamente alta, podríamos haber perdido una variable predictiva importante. Véase Cuadras (1993) para una discusión de este problema en términos de una desigualdad.

**Ejercicio 2.4** Una revista de automóviles publicó el consumo de gasolina  $Y$  en función de 10 características, sobre una muestra de 32 coches. Si 2 variables son binarias ( $b$ ), 3 cualitativas ( $q$ ) y el resto continuas ( $c$ ), compara el modelo de regresión clásico con el modelo DB utilizando la distancia de Gower. La variable dependiente se encuentra en la columna  $Y$ :

Y	c	c	c	c	c	b	b	q	q	q	Y	c	c	c	c	c	b	b	q	q	q
21.0	160.0	110	3.90	2.620	16.46	0	1	6	4	4	14.7	440.0	230	3.23	5.345	17.42	0	0	8	3	4
21.0	160.0	110	3.90	2.875	17.02	0	1	6	4	4	32.4	78.7	66	4.08	2.200	19.47	1	1	4	4	1
22.8	108.0	93	3.85	2.320	18.61	1	1	4	4	1	30.4	75.7	52	4.93	1.615	18.52	1	1	4	4	2
21.4	258.0	110	3.08	3.215	19.44	1	0	6	3	1	33.9	71.1	65	4.22	1.835	19.90	1	1	4	4	1
18.7	360.0	175	3.15	3.440	17.02	0	0	8	3	2	21.5	120.1	97	3.70	2.465	20.01	1	0	4	3	1
18.1	225.0	105	2.76	3.460	20.22	1	0	6	3	1	15.5	318.0	150	2.76	3.520	16.87	0	0	8	3	2
14.3	360.0	245	3.21	3.570	15.84	0	0	8	3	4	15.2	304.0	150	3.15	3.435	17.30	0	0	8	3	2
24.4	146.7	62	3.69	3.190	20.00	1	0	4	4	2	13.3	350.0	245	3.73	3.840	15.41	0	0	8	3	4
22.8	146.7	62	3.69	3.190	20.00	1	0	4	4	2	19.2	400.0	175	3.08	3.845	17.05	0	0	8	3	2
19.2	167.6	123	3.92	3.440	18.30	1	0	6	4	4	27.3	79.0	66	4.08	1.935	18.90	1	1	4	4	1
17.8	167.6	123	3.92	3.440	18.90	1	0	6	4	4	26.0	120.3	91	4.43	2.140	16.70	0	1	4	5	2
16.4	275.8	180	3.07	4.070	17.40	0	0	8	3	3	30.4	120.3	91	4.43	2.140	16.70	0	1	4	5	2
17.3	275.8	180	3.07	3.730	17.60	0	0	8	3	3	15.8	351.0	264	4.22	3.170	14.50	0	1	8	5	4
15.2	275.8	180	3.07	3.780	18.00	0	0	8	3	3	19.7	145.0	175	3.62	2.770	15.50	0	1	6	5	6
10.4	472.0	205	2.93	5.250	17.98	0	0	8	3	4	15.0	301.1	335	3.54	3.570	14.60	0	1	8	5	8
10.4	460.0	215	3.00	5.424	17.82	0	0	8	3	4	21.4	121.0	109	4.11	2.780	18.60	1	1	4	4	2

**Ejercicio 2.5** Escribe la versión de las fórmulas (2.12-2.15), para el modelo DB en dimensión  $k$ .

## 2.3 Predicción DB sobre un nuevo individuo

Supongamos que, sobre las variables (mixtas) explicativas, hemos obtenido la observación  $\mathbf{w}_{n+1} = (w_1, \dots, w_p)$  sobre un nuevo individuo  $\omega_{n+1}$ . Entonces debería ser posible calcular las distancias

$$d_i^2 = d^2(\omega_i, \omega_{n+1}) \quad i = 1, \dots, n \quad (2.19)$$

entre  $\omega_{n+1}$  y los otros individuos cuyas observaciones conocemos para la variable respuesta  $Y$ . Queremos evaluar

$$y_{n+1} = Y(\omega_{n+1}),$$

es decir, la predicción de  $Y$  sobre  $\omega_{n+1}$ .

Se puede obtener esta predicción mediante la fórmula de añadir un punto (1.18), hallando las coordenadas del nuevo individuo a partir de las distancias. Estas coordenadas, utilizando el modelo completo, son

$$\mathbf{x} = (x_1, \dots, x_m)' = \frac{1}{2} \mathbf{\Lambda}^{-1} \mathbf{X}'(\mathbf{b} - \mathbf{d}).$$

La predicción según el modelo (2.10) es

$$\hat{y}_{n+1} = \bar{y} + \sum_{i=1}^m \hat{\beta}_i x_i = \bar{y} + \mathbf{x}' \hat{\boldsymbol{\beta}}.$$

Substituyendo, se obtiene

$$\hat{y}_{n+1} = \bar{y} + \mathbf{x}' \mathbf{\Lambda}^{-1} \mathbf{X}' \mathbf{y}. \quad (2.20)$$

Si consideramos ahora el modelo DB en dimensión  $k$ , y hacemos las particiones

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_{(k)} \\ \mathbf{z} \end{pmatrix} \quad \mathbf{X} = (\mathbf{X}_{(k)}, \mathbf{Z}) \quad \mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{m-k} \end{pmatrix}$$

donde  $\mathbf{x}_{(k)} = (x_1, \dots, x_k)'$  son las  $k$  coordenadas relativas a las  $k$  dimensiones predictivas, y la diagonal de  $\mathbf{\Lambda}_k$  contiene los valores propios, obtenemos

$$\hat{y}_{n+1}(k) = \bar{y} + \mathbf{x}'_{(k)} \mathbf{\Lambda}_k^{-1} \mathbf{X}'_k \mathbf{y} + \mathbf{z}' \mathbf{\Lambda}_{m-k}^{-1} \mathbf{Z}' \mathbf{y}.$$

Si ahora tenemos en cuenta que  $\mathbf{Z}' \mathbf{y} \approx \mathbf{0}$  (ya que  $\mathbf{Z}$  contiene las variables menos correlacionadas con  $Y$ ), obtenemos finalmente:

$$\hat{y}_{n+1}(k) = \bar{y} + \mathbf{x}'_{(k)} \mathbf{\Lambda}_k^{-1} \mathbf{X}'_k \mathbf{y}. \quad (2.21)$$

**Ejercicio 2.6** Demuestra que, considerando el modelo completo, la predicción (2.20) se puede escribir como

$$\hat{y}_{n+1} = \bar{y} + \frac{1}{2} (\mathbf{b} - \mathbf{d})' \mathbf{B}^- \mathbf{y},$$

donde  $\mathbf{B}^-$  es la  $g$ -inversa de  $\mathbf{B} = \mathbf{X}\mathbf{X}'$ , es decir, tal que  $\mathbf{B}\mathbf{B}^-\mathbf{B} = \mathbf{B}$ . (Observa que  $\mathbf{B}$  es singular y por lo tanto no tiene inversa en el sentido tradicional).

## 2.4 Predicción con variables continuas, categóricas y mixtas

El modelo DB se reduce al modelo clásico de regresión cuando la distancia utilizada es Euclídea (1.6) y las variables son continuas. Cuadras y Arenas (1990) demuestran que:

- Para variables cuantitativas y distancia Euclídea, la fórmula de predicción (2.20) brinda los mismos resultados. (Para otras distancias, los resultados son diferentes. Véase más abajo y en la próxima sección).
- Para  $p$  variables cualitativas  $W_1, W_2, \dots, W_p$ , donde  $W_r$  tiene  $q_r$  estados ( $1 \leq r \leq p$ ), podemos tomar como distancia

$$d_{ij}^2 = 2(p - m_{ij}), \quad (2.22)$$

donde  $m_{ij}$  es el número de coincidencias entre los individuos  $i$  y  $j$ . Por ejemplo, si  $i = (010, 1000, 001, 10)$ ,  $j = (010, 0100, 001, 10)$ , entonces  $p = 4$ ,  $m_{ij} = 3$ ,  $d_{ij}^2 = 2$ . En ese caso, el modelo DB con la distancia (2.22) vuelve a dar los mismos resultados que el modelo de regresión clásica, es decir las predicciones son las mismas. Naturalmente los resultados son diferentes si consideramos otras distancias.

- La situación cambia si las variables son mixtas, esto es, una mezcla de continuas, binarias y categóricas. Entonces la distancia no es Euclídea en el sentido de antes. Una buena elección consiste en emplear la distancia de Gower (1.12). Existen muchos ejemplos que prueban que utilizando el método de regresión DB con esta distancia podemos obtener mejores predicciones que con el método clásico. Los ejercicios 2.4 y 2.7 brindan sendos ejemplos donde se puede comprobar este hecho

El método DB fue introducido por Cuadras (1989), Cuadras y Arenas (1990), y continuado por Cuadras (1990), Cuadras y Fortiana (1993), Cuadras, Arenas y Fortiana (1996). Estos artículos contienen ejemplos con datos reales.

**Ejercicio 2.7** *Se pide:*

1. Demuestra que la distancia (2.22) es Euclídea y que el método DB y el clásico (codificando los estados como 0= ausente, 1= presente) son equivalentes.
2. Comprueba que, para los siguientes datos ( $n=28$ ), donde la primera columna contiene la variable respuesta  $Y$  y las cinco columnas siguientes las variables explicativas, que el método DB con la distancia de Gower, proporciona mejores resultados que el método clásico de regresión múltiple. (Nota: utiliza el programa MULTICUA).

6.03	2	3	3	3	1	5.69	1	2	2	1	2
5.60	1	1	1	0	1	6.08	3	2	4	3	1
5.90	1	5	2	2	1	6.05	2	1	3	0	1
5.50	2	1	1	0	1	5.90	1	3	3	2	2
5.58	2	1	2	1	1	5.86	0	1	1	1	2
5.79	2	2	5	0	1	6.68	3	3	4	3	1
6.38	2	3	3	0	1	5.71	3	3	4	2	1
5.54	2	1	2	0	2	5.86	3	4	4	0	1
5.69	1	3	2	3	1	5.68	2	1	3	0	2
5.49	3	3	4	1	1	5.60	1	1	2	0	1
5.72	3	3	4	0	2	5.34	1	1	5	0	2
6.74	1	5	2	2	2	5.32	1	4	1	1	1
7.48	0	1	4	0	2	6.36	3	2	3	0	2
6.93	1	3	2	1	1	5.55	3	3	4	0	1

## 2.5 Regresión no lineal DB

Consideremos la regresión de  $Y$  sobre  $p$  variables cuantitativas  $X_1, \dots, X_p$ , según un modelo de regresión no lineal

$$Y = f(X_1, \dots, X_p; \theta) + e \quad (2.23)$$

donde  $\theta$  es el vector de parámetros,  $f$  es una función no lineal (posiblemente desconocida),  $e$  es el término aleatorio de error.

La regresión DB con la distancia Euclídea es equivalente a interpretar (2.23) en términos de un modelo lineal. Pero si consideramos la distancia valor absoluto (1.8) entonces la regresión DB es no lineal. Cuadras y Fortiana (1993) demuestran que, si  $p = 1$ , la regresión DB con la distancia

$$d(x, y) = \sqrt{|x - y|} \quad (2.24)$$

equivale a una regresión sobre polinomios ortogonales. Concretamente sobre polinomios de Tchebychev, cuando los valores  $x_1, x_2, \dots, x_n$  observados de la variable  $X$  son equidistantes.

De momento no existen resultados teóricos conocidos para  $p > 2$ , pero la utilidad del método DB con distancia (1.8) en regresión no lineal se pone de manifiesto con ejemplos reales.

**Ejercicio 2.8** Sobre el conjunto  $\Omega = \{0, 1, 2, \dots, n\}$  consideramos la distancia

$$d(i, j) = \sqrt{|i - j|} \quad (2.25)$$

1. Demuestra que es Euclídea (compárala con Ejercicio 1.10).
2. Prueba que para  $n = 8$ , los 4 ejes principales obtenidos por análisis de coordenadas principales, es decir utilizando (1.13), constituyen las dimensiones lineal, cuadrática, cúbica y cuártica.

**Ejercicio 2.9** En una reacción química,  $y$  es la fracción de material original remanente luego de  $x_1$  minutos de reacción a  $x_2$  grados Kelvin. El modelo de regresión no lineal que predice  $y$  es

$$y = \exp \left( -\theta_1 x_1 \exp \left[ -\theta_2 \left( \frac{1}{x_2} - \frac{1}{620} \right) \right] \right) + e \quad (2.26)$$

donde  $\theta_1, \theta_2$ , son parámetros desconocidos.

1. Estima los parámetros utilizando (2.26) y aplica también una regresión DB (distancia "valor absoluto") con  $k=2$  y  $k=3$ , calculando en cada caso las predicciones  $\hat{y}$  de  $y$  sobre los siguiente datos ( $n=8$ ):

COMPLETA ESTAS TRES COLUMNAS

$y$	$x_1$	$x_2$	$\hat{y}$ (con (2.26))	$\hat{y}$ (DB, $k=2$ )	$\hat{y}$ (DB, $k=3$ )
0.912	109	600			
0.382	65	640			
0.397	1180	600			
0.376	66	640			
0.342	1270	600			
0.358	69	640			
0.348	1230	600			
0.376	68	640			



2. Compara las tres predicciones calculando la cantidad

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

3. En las predicciones DB, ¿has tenido necesidad de conocer el modelo de regresión no lineal (2.26)? ¿Qué ventajas le encuentras al método DB?

**Ejercicio 2.10** La Tabla 2.1 contiene la renta per capita y una matriz binaria de relaciones comerciales entre 15 países (0 si no hay relación significativa; 1 si hay relación significativa).

1. Representa los 15 países por análisis de coordenadas principales utilizando el coeficiente de Jaccard. Interpreta las dos primeras dimensiones.
2. Estudia si la matriz binaria predice bien la renta y realiza la predicción de la renta per capita en España mediante:
  - (a) Regresión múltiple clásica.
  - (b) Regresión DB con 5 dimensiones.
3. Comenta los problemas encontrados con la regresión clásica y discute las ventajas del método DB para estos datos.

TABLA 2.1

		Ar	Au	Br	Ca	Cz	Eg	Fr	It	Ja	NZ	Sw	US	Ru	UK	Wg
Arge	4124	0	0	1	0	0	0	0	1	1	0	0	1	0	0	1
Aust	10981	0	0	0	0	0	0	0	0	1	1	0	1	0	1	1
Braz	2232	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1
Cana	13034	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0
Czec	2853	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Egyp	4368	0	0	0	0	0	0	1	1	0	0	0	1	1	1	1
Fran	8115	0	0	0	0	0	1	0	1	0	0	0	1	0	1	1
Ital	5549	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1
Japa	9928	1	1	1	1	0	0	0	0	0	1	0	1	0	0	0
N.Ze	6736	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0
Swed	10570	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
USA	13968	1	1	1	1	0	1	1	0	1	1	1	0	0	1	1
USSR	2588	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
U.K.	6514	0	1	0	1	0	1	1	0	0	1	1	1	0	0	1
W.Ge	9064	1	1	1	0	0	1	1	1	0	0	1	1	0	1	0
Esp	?	1	0	0	0	0	0	1	1	1	0	0	1	0	1	1

TABLA 2.2

	N	C	R	E	Z	Ei	J	G	M	prizes
N	0									120
C	.318	0								110
R	.270	.101	0							130
E	.311	.223	.061	0						145
Z	.378	.243	.236	0.061	0					160
Ei	.392	.236	.176	.088	.007	0				170
J	.399	.311	.345	.176	.074	.128	0			200
G	.392	.345	.297	.101	.209	.182	.027	0		225
M	.426	.358	.318	.230	.264	.128	.142	.128	0	?

**Ejercicio 2.11** Consideremos que  $1, 2, \dots, n$  son  $n$  estímulos. Supongamos que existe una escala de preferencia  $\theta_1, \theta_2, \dots, \theta_n$ , donde  $\theta_i$  es un valor asociado al estímulo  $i$ . El estímulo  $j$  es mejor que el estímulo  $i$  si  $\theta_j > \theta_i$ , lo que indicaremos como  $j > i$ . A fin de obtener la ordenación de los estímulos, los ubicamos en un eje unidimensional

$$i_1 < i_2 < \dots < i_n$$

es decir

$$\theta_{i_1} < \theta_{i_2} < \dots < \theta_{i_n}$$

evaluamos las proporciones

$$p_{ij} = P(j > i)$$

que indican, en una muestra grande de individuos, la proporción de individuos que prefieren  $j$  sobre  $i$ . Observemos que si  $p_{ij} > 0.5$  es que realmente  $j$  es preferible sobre  $i$ . Consideremos entonces el modelo basado en la distribución normal

$$p_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\theta_j - \theta_i} e^{-x^2/2} dx \quad (2.27)$$

y que permite calcular  $\theta_1, \theta_2, \dots, \theta_n$  en función de  $p_{ij}$ ,  $i, j = 1, \dots, n$ .

1. Interpreta el modelo (2.27) razonando los casos:

$$a) p_{ij} = 0.5 \quad b) p_{ij} \cong 1 \quad c) p_{ij} \cong 0$$

2. Definimos la siguiente distancia entre cada par de estímulos

$$d(i, j) = \begin{cases} |p_{ij} - 0.5| & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases} \quad (2.28)$$

Prueba que cumple con la propiedad simétrica  $d(i, j) = d(j, i) \geq 0$ .

**Ejercicio 2.12** En un estudio de nutrición se desean ordenar  $n = 9$  vegetales comestibles según el orden de preferencias que le dan una amplia gama de consumidores. Los vegetales son: nabo ( $N$ ), col ( $C$ ), remolacha ( $R$ ), espárrago ( $E$ ),

zanahoria ( $Z$ ), espinaca ( $Ei$ ), judía verde ( $J$ ), guisante ( $G$ ) y maíz ( $M$ ). Algunas de las proporciones son:

$$P(C > N) = 0.818 \quad P(E > C) = 0.723 \quad P(R > G) = 0.203$$

Utilizando la distancia (2.28) se ha obtenido la matriz de distancias de la Tabla 2.2. Aplica un análisis de coordenadas principales y encuentra la escala de preferencias, identificando cada vegetal con las coordenadas de la primera dimensión.

1. Comprueba que la matriz de distancias de la Tabla 2.2 (parte izquierda) no es Euclídea y encuentra una transformación que la convierta en Euclídea (ver Teorema 1.2).
2. Los precios (por kilo de producto) de los primeros 8 vegetales son:

120   110   130   145   160   170   200   225

Predice el precio del maíz.



## Capítulo 3

# Aspectos computacionales en predicción DB

### 3.1 Selección de variables en regresión DB

En la Sección 2.1 hemos sugerido dos métodos para seleccionar las dimensiones predictivas para el modelo de regresión DB en dimensión  $k$ .

El criterio de selección de los vectores propios de  $\mathbf{B}$  ordenados de acuerdo con los valores propios, es un método para obtener las dimensiones principales que permitan realizar la mejor representación, en dimensión  $k$ , de  $\Omega$ , de modo que se maximice la variabilidad geométrica en esta dimensión.

$$\frac{1}{2n^2} \sum_{i,j=1}^n \delta_{ij}^2(k) = \frac{1}{n} \sum_{i=1}^n \lambda_i. \quad (3.1)$$

Un buen criterio para seleccionar variables predictivas, consiste en ordenar las columnas de  $\mathbf{X}$  en orden descendente de las correlaciones con  $\mathbf{y}$ , es decir, seleccionar  $X_{i_1}, \dots, X_{i_k}$  tales que

$$r^2(\mathbf{y}, X_{i_1}) > \dots > r^2(\mathbf{y}, X_{i_k}). \quad (3.2)$$

El coeficiente de determinación  $R^2(k)$  (ver (2.15)), es:

$$R^2(k) = \sum_{\alpha=1}^k r^2(\mathbf{y}, X_{i_\alpha})$$

La proyección de  $\mathbf{y}$ , de acuerdo con el modelo (2.17), es:

$$\hat{\mathbf{y}} = \hat{\beta}_0 \mathbf{1} + \sum_{\alpha=1}^k \hat{\beta}_{i_\alpha} X_{i_\alpha}$$

La variabilidad geométrica DB condicional, se define como:

$$\frac{1}{2n^2} \sum_{i,j=1}^n (\hat{y}_i - \hat{y}_j)^2. \quad (3.3)$$

Es fácil probar que

$$\frac{1}{2n^2} \sum_{i,j=1}^n (\hat{y}_i - \hat{y}_j)^2 = \frac{1}{n} \sum_{\alpha=1}^k \hat{\beta}_{i_\alpha} \lambda_{i_\alpha} = s_y^2 R^2(k), \quad (3.4)$$

donde  $s_y^2$  es la varianza muestral de  $Y$ . Esta ecuación puede ser interpretada como la versión DB de (3.1). Nótese que el criterio (3.2) maximiza  $R^2(k)$ , es decir proporciona el subespacio en dimensión  $k$ , de las coordenadas principales que tienen máxima correlación múltiple con  $\mathbf{Y}$ .

**Ejercicio 3.1** Sea  $\bar{\mathbf{X}}$  la matriz de datos centrados y sea  $\mathbf{S} = \bar{\mathbf{X}}' \bar{\mathbf{X}}/n$  la matriz de covarianzas. La transformación de componentes principales es  $\mathbf{X} = \tilde{\mathbf{X}}\mathbf{T}$ , donde  $\mathbf{S} = \mathbf{T}\mathbf{D}\mathbf{T}'$  es la descomposición espectral de  $\mathbf{S}$ . Relacionando  $\mathbf{S}$  con  $\mathbf{B} = \mathbf{X}\mathbf{X}'$  (ver 1.12), demuestra que las coordenadas principales (columnas de  $\mathbf{X}$ ) pueden ser interpretadas como componentes principales.

**Ejercicio 3.2** Determina la distancia implícita en la definición (3.3) y demuestra (3.4).

## 3.2 Selección para un número grande de individuos

Los dos métodos de la Sección 4.1 se basan en maximizar (3.3) ó (3.4) y necesitan calcular los vectores propios de la matriz  $\mathbf{B}$ . Cuando  $n$  es muy grande, ese cálculo puede volverse muy arduo o imposible. Un procedimiento que solamente requiere calcular los primeros  $k$  vectores propios adecuados, es el siguiente.

Se particiona  $\mathbf{X}$  en:

$$\mathbf{X} = (\mathbf{X}_{(i)}, \mathbf{Z}_i)$$

donde  $\mathbf{X}_{(i)}$  contiene las primeras columnas de  $\mathbf{X}$ , es decir los primeros vectores propios de  $\mathbf{B}$ , ordenados de acuerdo con sus valores propios. Se define la secuencia:

$$c(0) = 0, \quad c(i) = \mathbf{y}'\mathbf{B}_{(i)}\mathbf{y}/\mathbf{y}'\mathbf{B}\mathbf{y} \quad i = 1, 2, \dots, m, \quad (3.5)$$

donde  $m = \text{rango}(\mathbf{B})$  y  $\mathbf{B}_{(i)} = \mathbf{X}_{(i)}\mathbf{X}_{(i)}'$ . El coeficiente  $c(i)$  verifica

$$c(i) = \frac{\sum_{\alpha=1}^i r_\alpha^2 \lambda_\alpha}{\sum_{\alpha=1}^m r_\alpha^2 \lambda_\alpha} \quad (3.6)$$

donde  $r_\alpha = \text{corr}(\mathbf{y}, \mathbf{X}_\alpha)$ . Cada  $c(i)$  mide la predictibilidad de las primeras  $i$  dimensiones, ponderadas por los correspondientes valores propios. El valor

### 3.3. EL CASO DE TENER VALORES PROPIOS NEGATIVOS O MUY PEQUEÑOS 31

inicial  $c(0) = 0$ , puede ser interpretado como la falta de predicibilidad de  $\mathbf{1}$ , el vector constante de unos, que es también un valor propio de  $\mathbf{B}$ . Esta secuencia satisface:

- $c(0) = 0 \leq c(i) \leq 1, \quad i = 1, \dots, m.$
- $c(m) = 1.$
- $c(i) \leq c(i+1) \quad i = 0, \dots, m-1.$
- $c(i) = c(i+1)$ , si la dimensión  $(i+1)$  es no predictiva.

La selección debe ser realizada representando gráficamente los puntos

$$(i, 1 - c(i)) \quad i = 0, 1, \dots, m^* < m$$

donde  $m^*$  es tal que  $1 - c(i)$  esté muy próximo a 0. Esto es, el corte en  $m^*$  es tal que, a la derecha de  $m^*$  el gráfico está muy próximo al eje horizontal, indicando que las dimensiones superiores no deben ser tenidas en cuenta. La dimensión principal  $1 \leq i \leq m^*$  debe ser seleccionada si se aprecia una caída entre el punto  $(i-1, 1 - c(i-1))$  y el  $(i, 1 - c(i))$ . Entonces la dimensión  $i$  es aceptada o rechazada según si  $r_i^2$  o  $\lambda_i$  sean grandes o pequeños.

Finalmente, es necesario hacer notar que este procedimiento solamente requiere el cálculo de los primeros  $m^*$  vectores propios.

**Ejercicio 3.3** *Se pide:*

1. Construye el gráfico de  $(i, 1 - c(i))$  para los datos de coches del ejercicio 2.4.
2. Prueba (3.6) e interpreta una dimensión problemática, es decir una dimensión donde  $\lambda_i$  es grande y  $r_i^2$  es pequeño o recíprocamente  $\lambda_i$  es pequeño y  $r_i^2$  es grande.

### 3.3 El caso de tener valores propios negativos o muy pequeños

Supongamos ahora que la matriz  $n \times n$  de distancias observadas,  $\Delta = (\delta_{ij})$ , no es Euclídea. Puede ocurrir, por ejemplo, que se emplee el coeficiente de similitud de Gower con algunos datos faltantes. Entonces la matriz  $\mathbf{B}$  tendrá algunos valores propios negativos (Teorema 1.2), por lo tanto nos encontremos con algunas variables (columnas de  $\mathbf{X}$ ) con “varianzas negativas”. Consecuentemente el modelo de regresión DB no funcionaría correctamente.

Para solucionar este inconveniente, podemos considerar la transformación q-aditiva de la distancia:

$$\tilde{\delta}_{ij}^2 = \begin{cases} 0 & i = j \\ \delta_{ij}^2 + 2a & i \neq j \end{cases} \quad (3.7)$$

que transforma la matriz  $\mathbf{B}$  en

$$\widehat{\mathbf{B}} = \mathbf{B} + a\mathbf{H} \quad (3.8)$$

Los vectores propios de  $\widehat{\mathbf{B}}$  son los mismos y los valores propios son  $\lambda_i + a$

$$\widehat{\mathbf{B}}X_i = (\lambda_i + a)X_i \quad (3.9)$$

Si  $\lambda_i < 0$  para algún  $i$ , podemos escoger  $a > 0$  tal que  $\lambda_i + a > 0$ ,  $i = 1, \dots, m$ , lo que nos permite conseguir la matriz Euclídea  $\tilde{\Delta} = (\tilde{\delta}_{ij})$ .

Considerando el modelo completo y escribiendo la ecuación (2.20) como

$$\widehat{\mathbf{y}}_{n+1} = \bar{\mathbf{y}} + \frac{1}{2}(\mathbf{b} - \mathbf{d})'\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}'\mathbf{y}, \quad (3.10)$$

entonces la transformación q-aditiva (3.7), cambia el modelo DB a

$$\widehat{\mathbf{y}}_{n+1}(a) = \bar{\mathbf{y}} + \frac{1}{2}(\mathbf{b} - \mathbf{d})'\mathbf{U}(\mathbf{\Lambda} + a\mathbf{I})^{-1}\mathbf{U}'\mathbf{y}, \quad (3.11)$$

donde  $\mathbf{I}$  es la matriz identidad  $n \times n$ . Nótese que esta predicción para un nuevo individuo  $\omega_{n+1}$ , depende de la constante  $a$ .

**Ejercicio 3.4** *Se pide:*

1. Demuestra las ecuaciones (3.9)-(3.11).
2. Propón un criterio para seleccionar el mejor  $a$  en (3.11).
3. Escribe una versión apropiada de (3.11) para el modelo DB en dimensión  $k$ .

Supongamos ahora que  $\Delta = (\delta_{ij})$  es una matriz de distancias Euclídea, pero algunos valores propios de  $\mathbf{B}$  son muy pequeños. Entonces el ajuste de (3.10) puede ser muy pobre. Quizás el empleo de (3.11), que requiere la inversa de la matriz  $(\mathbf{\Lambda} + a\mathbf{I})$ , puede mejorar el ajuste. Este inconveniente también se presenta para la estimación de los parámetros de regresión, (véase 2.12):

$$\widehat{\beta}_i(a) = \mathbf{X}_i'\mathbf{y}/(\lambda_i + a) \quad (3.12)$$

Existe una analogía entre (3.11) - (3.12) y la *regresión cresta* (Hoel y Kennard, 1970). La transformación (3.7), puede ser interpretada como una extensión de este procedimiento de regresión para una distancia general.

**Ejercicio 3.5** *Se pide:*

1. Construye un ejemplo donde el ajuste con el modelo DB dé  $R^2 = 0$  y otro ejemplo tal que  $R^2 = 1$ .
2. Describe la regresión cresta y posteriormente comenta la versión DB.



## Capítulo 4

# Análisis discriminante DB

### 4.1 Introducción

Supongamos que  $\Omega_1, \Omega_2$  son dos poblaciones,  $X_1, \dots, X_p$  son  $p$  variables observables y  $\mathbf{x} = (x_1, \dots, x_p)$  contiene las observaciones de las variables sobre un individuo  $\omega$ . Queremos asignar  $\omega$  a una de las dos poblaciones.

Una **regla discriminante** es un criterio que permite asignar  $\omega \in \Omega_1 \cup \Omega_2$ , y que a menudo se plantea en términos de una **función discriminante**  $D(x_1, \dots, x_p)$ . Entonces la regla de clasificación es

$$\begin{array}{ll} \text{Si } D(x_1, \dots, x_p) \geq 0 & \text{asignamos } \omega \text{ a } \Omega_1, \\ \text{en otro caso} & \text{asignamos } \omega \text{ a } \Omega_2. \end{array}$$

Las cuatro reglas mas importantes del análisis discriminante son:

1. **Regla MV o de la máxima verosimilitud**

Sea  $f_i(x)$  la densidad de  $\mathbf{x}$  en  $\Omega_i$ ,  $i = 1, 2$ . La función discriminante es:

$$V(\mathbf{x}) = \log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) \quad (4.1)$$

2. **Regla B o regla de Bayes**

Supongamos que son conocidas las probabilidades a priori:

$$q_1 = P(\Omega_1), \quad q_2 = P(\Omega_2), \quad q_1 + q_2 = 1$$

Entonces la función discriminante es:

$$B(\mathbf{x}) = \log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) + \log(q_1/q_2) \quad (4.2)$$

3. **Regla M o de Matusita**

Supongamos que estamos en condiciones de calcular una distancia  $\delta_i = \delta(\omega, \Omega_i)$ ,  $i = 1, 2$ , de  $\omega$  a cada población. Usualmente esta distancia es del tipo  $\delta_i = \delta(\mathbf{x}, \boldsymbol{\mu}_i)$ ,  $i = 1, 2$ , donde  $\mathbf{x}$  es el vector de observaciones y  $\boldsymbol{\mu}_i$  es

un vector representante de  $\Omega_i$ , por ejemplo el vector de medias. La regla M está basada en la función discriminante

$$M(\mathbf{x}) = \delta_2^2(\mathbf{x}) - \delta_1^2(\mathbf{x}). \quad (4.3)$$

#### 4. Regla de Fisher del discriminador lineal

Si  $\Omega_i \sim (\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ , es decir, se puede identificar  $\Omega_i$  haciendo uso de un vector de medias  $\boldsymbol{\mu}_i$  y una matriz de covariancias  $\boldsymbol{\Sigma}$ , entonces el discriminador lineal es

$$L(\mathbf{x}) = \left[ \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right]' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \quad (4.4)$$

Todas estas reglas de clasificación tienen una interpretación sencilla. Se cumple que:

- La regla MV asigna  $\omega$  a una población  $\Omega_i$ , tal que la verosimilitud  $f_i(\mathbf{x})$  es mayor.
- La regla de Bayes lo que tiene en cuenta es el valor mas grande de la probabilidad “a posteriori”  $P(\Omega_i/\mathbf{x})$ . MV es un caso particular de la regla B si  $q_1 = q_2 = 1/2$ .
- La regla M simplemente asigna  $\omega$  a la población que tiene mas próxima.
- El primer discriminador lineal que fue estudiado, es un caso particular de la regla M cuando  $\Omega_i \sim (\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ ,  $i = 1, 2$ , y

$$\delta_j^2 = (\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \quad (4.5)$$

es la distancia de Mahalanobis.

Todas esas reglas esencialmente coinciden en el caso particular de que las poblaciones sean normales multivariantes  $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ , es decir, la función de densidad en  $\Omega_i$  es:

$$f_i(\mathbf{x}) = \frac{|\boldsymbol{\Sigma}_i^{-1}|^{1/2}}{(2\pi)^{p/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right\}$$

con  $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ .

**Ejercicio 4.1** Supóngase que  $\Omega_i$  es una población  $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ ,  $i = 1, 2$ .

1. Demuestra que, si  $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ , las reglas MV, B con  $q_1 = q_2 = 1/2$ , M y de Fisher, coinciden.
2. Demuestra que, si  $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ , la regla MV está basada en el discriminador cuadrático

$$Q(\mathbf{x}) = \frac{1}{2} \mathbf{x}' (\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}) \mathbf{x} + \mathbf{x}' (\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2) + \frac{1}{2} \boldsymbol{\mu}_2' \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2 - \frac{1}{2} \boldsymbol{\mu}_1' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \log |\boldsymbol{\Sigma}_2| - \frac{1}{2} \log |\boldsymbol{\Sigma}_1|. \quad (4.6)$$

## 4.2 La función de proximidad de un individuo a una población

La distancia (4.5) proporciona una medida de la proximidad o de la distancia de  $\mathbf{x}$  a una población. Queremos generalizar esta medida haciendo uso solamente de **distancias entre observaciones**, sin necesidad de trabajar con medias, que son conceptos relativos a la población.

Sea  $\delta$  una distancia sobre  $\Omega$ ,  $\mathbf{x} = (x_1, \dots, x_p)$  un vector aleatorio con densidad  $f(x_1, \dots, x_p)$  con soporte  $S$ . Llamaremos **variabilidad geométrica** relativa a la distancia  $\delta$ , a la cantidad

$$V_\delta(\mathbf{X}) = \frac{1}{2} \int_{S \times S} \delta^2(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y} \quad (4.7)$$

$V_\delta(\mathbf{X})$  es el valor esperado de todas las interdistancias (al cuadrado). Podemos entender  $V_\delta(\mathbf{X})$  como una varianza generalizada.

**Ejercicio 4.2** Demuestra que:

1. Si  $X$  es una variable aleatoria y  $\delta(x, y) = |x - y|$ , entonces

$$V_\delta(X) = \text{var}(X).$$

2. Si  $\mathbf{X}$  es un vector aleatorio con matriz de covarianzas  $\Sigma$ , y  $\delta = d_E$  es la distancia Euclídea (1.5), entonces

$$V_{d_E}(\mathbf{X}) = \text{tra}(\Sigma).$$

Sea  $\omega$  un individuo de  $\Omega$ , y  $\mathbf{x} = (x_1, \dots, x_p)$  las observaciones de  $\mathbf{X}$  sobre  $\omega$ . Dada una distancia  $\delta$  sobre  $\Omega$ , definiremos la **función de proximidad** (o simplemente **proximidad**) de  $\omega$  a  $\Omega$  en relación con  $\mathbf{X}$  a la función

$$\phi_\delta^2(\mathbf{x}) = E[\delta^2(\mathbf{x}, \mathbf{Y})] - V_\delta(\mathbf{X}) = \int_S \delta^2(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y} - V_\delta(\mathbf{X}) \quad (4.8)$$

$\phi_\delta^2(\mathbf{x})$  es el valor esperado de la distancia al cuadrado de  $\mathbf{x}$  (que es fijo) a  $\mathbf{y}$  (que es aleatorio). Como veremos en el Teorema 3.1,  $\phi_\delta^2(\mathbf{x})$  puede expresarse como una distancia (al cuadrado) entre  $\mathbf{x}$  y un vector medio asociado a la población.

Supongamos finalmente que tenemos dos poblaciones  $\Omega_1, \Omega_2$  asociadas a dos vectores aleatorios  $\mathbf{X}, \mathbf{Y}$  con funciones de densidad  $f_1(\mathbf{x}), f_2(\mathbf{y})$  y variabilidades geométricas  $V_\delta(\mathbf{X}), V_\delta(\mathbf{Y})$ . La distancia (al cuadrado) entre las dos poblaciones se define como

$$\Delta^2(\Omega_1, \Omega_2) = \int_{S \times S} \delta^2(\mathbf{x}, \mathbf{y}) f_1(\mathbf{x}) f_2(\mathbf{y}) d\mathbf{x} d\mathbf{y} - V_\delta(\mathbf{X}) - V_\delta(\mathbf{Y}).$$

**Ejercicio 4.3** Se pide:

1. Sea  $X$  una variable aleatoria con distribución exponencial, de parámetro  $\alpha$ . Calcula la función de proximidad para las distancias:

$$a) \quad \delta(x, y) = |x - y| \quad b) \quad \delta(x, y) = \sqrt{|x - y|}$$

2. Sea  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Demuestra que si  $\delta$  es la distancia de Mahalanobis, entonces

$$\phi_\delta^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (4.9)$$

Supongamos ahora que existe una aplicación  $\psi : R^p \rightarrow R^m$ , tal que

$$\begin{aligned} \mathbf{x} &\rightarrow \psi(\mathbf{x}) \\ \delta(\mathbf{x}, \mathbf{y}) &= d_E(\psi(\mathbf{x}), \psi(\mathbf{y})) \end{aligned} \quad (4.10)$$

donde  $d_E$  es la distancia Euclídea (1.6). Diremos que  $\psi$  proporciona una representación Euclídea de  $\delta$ .

El siguiente Teorema generaliza (4.9), de manera que la función de proximidad se puede interpretar como una distancia del individuo a la población, y  $\Delta^2(\Omega_1, \Omega_2)$  como la distancia entre dos vectores medios.

**Teorema 4.1** *Supongamos que existe una representación Euclídea de  $(\Omega, \delta)$  en un espacio  $L$  (Euclídeo o de Hilbert separable) con un producto escalar  $\langle \cdot, \cdot \rangle$  y una norma  $\|\mathbf{z}\|^2 = \langle \mathbf{z}, \mathbf{z} \rangle$ , tal que*

$$\delta^2(\mathbf{x}, \mathbf{y}) = \|\psi(\mathbf{x}) - \psi(\mathbf{y})\|^2, \quad (4.11)$$

donde  $\psi(\mathbf{x}), \psi(\mathbf{y}) \in L$  son las imágenes de  $\mathbf{x}, \mathbf{y}$ . Entonces se verifica que

$$\phi_\delta^2(\mathbf{x}_0) = \|\psi(\mathbf{x}_0) - E(\psi(\mathbf{X}))\|^2, \quad (4.12)$$

donde  $\mathbf{x}_0$  contiene las coordenadas de un individuo fijado  $\omega$ . Además se tiene que

$$\Delta^2(\Omega_1, \Omega_2) = \|E(\psi(\mathbf{X})) - E(\psi(\mathbf{Y}))\|^2.$$

### 4.3 La regla discriminante DB

Sean ahora  $\Omega_1, \Omega_2$  dos poblaciones y  $\delta$  una función de distancia.  $\delta$  es la misma en cada población, pero puede tener versiones ligeramente diferentes, digamos  $\delta_1, \delta_2$ , cuando nos encontramos en  $\Omega_1, \Omega_2$ , respectivamente. Por ejemplo, si las poblaciones son normales  $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ , y consideramos las distancias de Mahalanobis

$$\begin{aligned} \delta_1^2(\mathbf{x}, \mathbf{y}) &= (\mathbf{x} - \mathbf{y})' \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \mathbf{y}), \\ \delta_2^2(\mathbf{x}, \mathbf{y}) &= (\mathbf{x} - \mathbf{y})' \boldsymbol{\Sigma}_2^{-1} (\mathbf{x} - \mathbf{y}), \end{aligned}$$

entonces lo único que cambia es la matriz  $\boldsymbol{\Sigma}$ . Lo que debe quedar bien claro es que  $\delta$  depende del vector aleatorio  $\mathbf{X}$ , que en general tendrá diferente distribución en  $\Omega_1$  y  $\Omega_2$ .

Seguidamente, mediante (4.8), calcularemos las funciones de proximidad  $\phi_1^2, \phi_2^2$ , correspondientes a  $\Omega_1, \Omega_2$ . Sea  $\omega$  un individuo a clasificar en  $\Omega_1$  o en  $\Omega_2$ , con valores  $\mathbf{x}_0 = \mathbf{X}(\omega)$ .

La regla de clasificación DB es:

$$\begin{aligned} &\text{Asignar } \omega \text{ a } \Omega_1 \text{ si } \phi_1^2(\mathbf{x}_0) \leq \phi_2^2(\mathbf{x}_0) \\ &\text{en otro caso, asignar } \omega \text{ a } \Omega_2. \end{aligned} \quad (4.13)$$

Teniendo en cuenta que, si la aplicación (4.10) existe, se cumple que

$$\begin{aligned} \phi_1^2(\mathbf{x}_0) &= \|\psi(\mathbf{x}_0) - E_{\Omega_1}(\psi(\mathbf{X}))\|^2 \\ \phi_2^2(\mathbf{x}_0) &= \|\psi(\mathbf{x}_0) - E_{\Omega_2}(\psi(\mathbf{X}))\|^2 \end{aligned} \quad (4.14)$$

concluimos que la regla DB asigna  $\omega$  a la población que tiene más próxima. Por lo tanto, la regla DB es una regla M, pero que depende solamente de las distancias individuales.

## 4.4 Propiedades de la función de proximidad

Transformando la distancia  $\delta$  también se transforma la proximidad  $\phi^2$ . Se cumple:

- Si  $\delta^2 = c\delta^2$ , donde  $c > 0$  es una constante, entonces

$$\tilde{\phi}^2 = c\phi^2 \quad (4.15)$$

- Si  $\tilde{\delta}^2 = \delta^2 + b$  es una transformación q-aditiva de  $\delta$  (ver (1.2)) entonces

$$\tilde{\phi}^2 = \phi^2 + b/2 \quad (4.16)$$

- Si  $\delta^2 = \delta_1^2 + \delta_2^2$  entonces

$$\phi^2 = \phi_1^2 + \phi_2^2 \quad (4.17)$$

**Ejercicio 4.4** Se pide:

1. Demuestra las propiedades (4.15) a (4.17).
2. Razona que la propiedad aditiva (4.17) resulta apropiada cuando consideramos dos variables independientes  $X, Y$ , y  $\delta_1$  está asociada a  $X$  y  $\delta_2$  está asociada a  $Y$ .

## 4.5 La regla DB comparada con algunas reglas clásicas

El discriminador lineal es un caso particular de la regla DB ya que se puede expresar

$$L(\mathbf{x}_0) = \frac{1}{2} [\phi_2^2(\mathbf{x}_0) - \phi_1^2(\mathbf{x}_0)] \quad (4.18)$$

donde  $\phi_i^2(\mathbf{x}_0)$  se calcula tomando una distancia adecuada.

Análogamente si consideramos la distancia

$$\delta_j^2(\mathbf{x}, \mathbf{y}) = \begin{cases} (\mathbf{x} - \mathbf{y})' \Sigma_j^{-1} (\mathbf{x} - \mathbf{y}) + \log |\Sigma_j| / 2 & \mathbf{x} \neq \mathbf{y}, \\ 0 & \mathbf{x} = \mathbf{y}, \end{cases} \quad (4.19)$$

entonces el discriminador cuadrático (4.6) es

$$Q(\mathbf{x}_0) = \frac{1}{2} [\phi_2^2(\mathbf{x}_0) - \phi_1^2(\mathbf{x}_0)] \quad (4.20)$$

Finalmente, el llamado **discriminador Euclídeo**

$$E(\mathbf{x}_0) = \left[ \mathbf{x}_0 - \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right]' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad (4.21)$$

es un caso particular del DB si tomamos la distancia Euclídea como distancia entre individuos.  $E(\mathbf{x}_0)$  tiene la ventaja sobre  $L(\mathbf{x}_0)$  de que no exige el cálculo de la inversa de  $\Sigma$ .

**Ejercicio 4.5** *Se pide:*

1. ¿Cuál es la distancia adecuada a fin de encontrar (4.18)?
2. En el cálculo de (4.21), ¿cuál es la matriz de covarianzas común a  $\Omega_1, \Omega_2$ ?
3. Demuestra (4.20).
4. ¿Qué le pasaría a la distancia (4.19) si  $|\Sigma_j| < 1$ ? ¿Cómo podemos superar este aparente inconveniente?

Las expresiones que dan  $L(\mathbf{x}_0), Q(\mathbf{x}_0), E(\mathbf{x}_0)$ , son apropiadas para variables continuas. Por ejemplo, se demuestra que  $L(\mathbf{x}_0)$  es la mejor función discriminante cuando  $\Omega_j$  es  $N_p(\boldsymbol{\mu}_j, \Sigma)$ ,  $i = 1, 2$ .

Supongamos ahora que  $\Omega_1, \Omega_2$  son dos poblaciones multinomiales y que, respecto a unas características cualitativas  $A_1, \dots, A_m$  (sucesos disjuntos), sus probabilidades son

$$\begin{aligned} \mathbf{p}_1 &= (p_{11}, \dots, p_{1m}), & p_{1k} &\geq 0, & \sum p_{1k} &= 1, \\ \mathbf{p}_2 &= (p_{21}, \dots, p_{2m}), & p_{2k} &\geq 0, & \sum p_{2k} &= 1. \end{aligned}$$

#### 4.5. LA REGLA DB COMPARADA CON ALGUNAS REGLAS CLÁSICAS 39

Si un individuo  $\omega$  a clasificar, presenta la característica  $A_k$  con probabilidad  $p_{1k}$  si  $\omega \in \Omega_1$ ,  $p_{2k}$  si  $\omega \in \Omega_2$ , la regla MV es evidentemente

$$\text{Clasificar } \omega \text{ a } \Omega_i \text{ si } p_{ik} = \max \{p_{1k}, p_{2k}\}. \quad (4.22)$$

Consideremos ahora la distancia entre individuos

$$\delta_i^2(\omega, \omega') = \begin{cases} 0 & \text{fsi } \omega \in A_k, \omega' \in A_k \\ (p_{ik}^{-1} + p_{i1}^{-1}) & \text{si } \omega \in A_k, \omega' \in A_i, \quad k \neq i \end{cases} \quad (4.23)$$

que es la llamada distancia de Rao (ver Sección 1.7) en el caso de poblaciones multinomiales.

Los valores de las funciones de proximidad para un individuo  $\omega$  tal que presenta la característica  $A_k$ , son:

$$\phi_i^2(k) = \frac{1 - p_{ik}}{p_{ik}} \quad i = 1, 2 \quad (k = 1, \dots, m)$$

Se verifica que:

$$\min \{\phi_1^2(k), \phi_2^2(k)\} = \phi_i^2(k) \iff p_{ik} = \max \{p_{1k}, p_{2k}\} \quad (4.24)$$

y por lo tanto, la regla DB es equivalente a una regla MV para poblaciones multinomiales.

**Ejercicio 4.6** *Se pide:*

1. Demuestra que en una población multinomial, la variabilidad geométrica de la distancia (4.23) es igual a  $m - 1$ .
2. Demuestra la implicación (4.24)

**Ejercicio 4.7** *Supongamos dos poblaciones normales*

$$\Omega_1 = N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \quad \Omega_2 = N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

donde  $\boldsymbol{\mu}_1 = (0, 0)'$ ,  $\boldsymbol{\mu}_2 = (1, 2)'$ .

1. Encuentra el discriminador lineal en el caso

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

2. Encuentra el discriminador cuadrático en el caso

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

3. Si el discriminador es lineal, la probabilidad de clasificación errónea  $pce$  se define como:

$$pce = \frac{1}{2}P(\mathbf{x} \rightarrow \Omega_1/\Omega_2) + \frac{1}{2}P(\mathbf{x} \rightarrow \Omega_2/\Omega_1)$$

Demuestra que es igual a

$$pce = \Phi(-M/2) \quad (4.25)$$

donde  $\Phi$  es la función de distribución  $N(0, 1)$  y

$$M^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad (4.26)$$

es la distancia de Mahalanobis entre  $\Omega_1$  y  $\Omega_2$ .

4. Calcula (4.25) bajo las condiciones del apartado 1.

## 4.6 La regla DB en el caso de muestras

En la práctica no disponemos de las funciones de densidad  $f_1(\mathbf{x})$ ,  $f_2(\mathbf{x})$  para cada población, sino de dos muestras de tamaños  $n_1, n_2$  de las variables  $\mathbf{X} = (X_1, \dots, X_p)$ . Indiquemos (las representaciones Euclídeas de las muestras) por

$$\begin{array}{ll} \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_1} & \text{muestra de } \Omega_1 \\ \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{n_2} & \text{muestra de } \Omega_2 \end{array} \quad (4.27)$$

Consideremos primeramente sólo una población. Dada una distancia  $\delta$ , y una muestra  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , una estimación de la variabilidad geométrica (4.7) es

$$\hat{V}_\delta(\mathbf{X}) = \frac{1}{2n^2} \sum_{i,j} \delta^2(\mathbf{x}_i, \mathbf{x}_j), \quad (4.28)$$

Si  $\omega$  es un individuo, una estimación de la función de proximidad (4.8) es

$$\hat{\phi}_\delta^2(\mathbf{x}_0) = \frac{1}{n} \sum_{i=1}^n \delta^2(\mathbf{x}_0, \mathbf{x}_i) - \frac{1}{2n^2} \sum_{i,j=1}^n \delta^2(\mathbf{x}_i, \mathbf{x}_j)$$

siendo  $\delta^2(\mathbf{x}_0, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , las distancias de  $\omega$  a las observaciones de la muestra, donde  $\omega$  está representado por la imagen  $\mathbf{x}_0$  (ver 4.10), aunque esta suposición no es necesaria, ya que las distancias las podemos calcular igualmente partiendo de los datos originales.

Volviendo ahora al caso de dos poblaciones, si  $\omega$  es un individuo a clasificar, las estimaciones, en base a las muestras (4.27), de las dos funciones de proximidad son (ver 4.8):

$$\hat{\phi}_1^2(\mathbf{x}_0) = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta^2(\mathbf{x}_0, \mathbf{x}_i) - \frac{1}{2n_1^2} \sum_{i,j=1}^{n_1} \delta^2(\mathbf{x}_i, \mathbf{x}_j)$$



$$\hat{\phi}_2^2(\mathbf{y}_0) = \frac{1}{n_2} \sum_{i=1}^{n_2} \delta^2(\mathbf{y}_0, \mathbf{y}_i) - \frac{1}{2n_2^2} \sum_{i,j=1}^{n_2} \delta^2(\mathbf{y}_i, \mathbf{y}_j) \quad (4.29)$$

donde  $\mathbf{x}_0, \mathbf{y}_0$  son las dos imágenes de  $\omega$  (ver 4.10), según consideremos que  $\omega$  pertenezca a  $\Omega_1$  o a  $\Omega_2$ . hay que insistir, sin embargo en que el conocimiento de  $\mathbf{x}_0, \mathbf{y}_0$ , no es necesario, sino que *solamente hacen falta las distancias entre observaciones, a fin de obtener  $\hat{\phi}_1^2, \hat{\phi}_2^2$* .

Podemos análogamente encontrar la versión muestral de  $\Delta^2(\Omega_1, \Omega_2)$  :

$$\hat{\Delta}^2(\Omega_1, \Omega_2) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \delta^2(\mathbf{x}_i, \mathbf{y}_j) - \frac{1}{2n_1^2} \sum_{i,j=1}^{n_1} \delta^2(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{2n_2^2} \sum_{i,j=1}^{n_2} \delta^2(\mathbf{y}_i, \mathbf{y}_j). \quad (4.30)$$

La regla de clasificación DB es entonces:

$$\text{Asignar } \omega \text{ a } \Omega_i \quad \text{si} \quad \hat{\phi}_i^2 = \min \{ \hat{\phi}_1^2, \hat{\phi}_2^2 \}. \quad (4.31)$$

El siguiente teorema nos demuestra que podemos entender (4.31) como una regla M, cuando la citada representación existe.

**Teorema 4.2** *Supongamos que en dos espacios Euclídeos (posiblemente diferentes) podemos representar  $\omega$  y las dos muestras como*

$$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_1} \in \mathbf{R}^p \quad \mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{n_2} \in \mathbf{R}^q$$

*respectivamente. Entonces se cumple que*

$$\hat{\phi}_1^2 = d_E^2(\mathbf{x}_0, \bar{\mathbf{x}}) \quad \hat{\phi}_2^2 = d_E^2(\mathbf{y}_0, \bar{\mathbf{y}}) \quad (4.32)$$

donde  $\bar{\mathbf{x}} = (\sum_{i=1}^{n_1} \mathbf{x}_i) / n_1$ ,  $\bar{\mathbf{y}} = (\sum_{i=1}^{n_2} \mathbf{y}_i) / n_2$  son los centroides de las representaciones Euclídeas de las muestras.

Como consecuencia del Teorema 3.2, podemos formular la regla DB como sigue:

$$\begin{aligned} &\text{Clasificar } \omega \text{ a } \Omega_1 \text{ si } d_E^2(\mathbf{x}_0, \bar{\mathbf{x}}) < d_E^2(\mathbf{y}_0, \bar{\mathbf{y}}), \\ &\text{en otro caso a } \Omega_2. \end{aligned} \quad (4.33)$$

Por lo tanto, asignamos  $\omega$  a la población que está más próxima.

**Ejercicio 4.8** Sea  $\Delta_1 = (\delta_{ij})$  la matriz de distancias  $n_1 \times n_1$  entre las muestras de la primera población. Sean  $\delta_1, \delta_2, \dots, \delta_{n_1}$  las distancias de un individuo  $\omega$  a cada uno de los elementos de la muestra. Se pide:

1. ¿Cómo podemos encontrar una representación en los términos del Teorema 3.2?

2. Demuestra que si podemos expresar

$$\begin{aligned}\delta_i^2 &= (\mathbf{x}_i - \mathbf{x}_0)' (\mathbf{x}_i - \mathbf{x}_0) & i &= 1, \dots, n_1 \\ \delta_{ij}^2 &= (\mathbf{x}_i - \mathbf{x}_j)' (\mathbf{x}_i - \mathbf{x}_j) & i, j &= 1, \dots, n_1\end{aligned}$$

entonces

$$\sum_{i,j=1}^{n_1} \delta_{ij}^2 = 2n_1 \sum_{i=1}^{n_1} (\mathbf{x}_i - \bar{\mathbf{x}})' (\mathbf{x}_i - \bar{\mathbf{x}})$$

3. Utiliza este resultado para probar que

$$\hat{\phi}_1^2 = d_E^2(\mathbf{x}_0, \bar{\mathbf{x}}).$$

4. Escribe las expresiones correspondientes a la segunda población y razona por qué  $\mathbf{x}_0$  es posiblemente diferente de  $\mathbf{y}_0$ .

**Ejercicio 4.9** Tenemos dos poblaciones  $\Omega_1, \Omega_2$  y una muestra de cada población. respecto a tres variables mixtas, las matrices de distancias entre 5 individuos de la muestra 1 y de la muestra 2, de cada una de las poblaciones respectivamente, son:

$$\Delta_1 = \begin{pmatrix} 0 & 1 & 1 & 3 & 5 \\ & 0 & 2 & 1 & 5 \\ & & 0 & 1 & 2 \\ & & & 0 & 2 \\ & & & & 0 \end{pmatrix} \quad \Delta_2 = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ & 0 & 2 & 3 & 4 \\ & & 0 & 3 & 4 \\ & & & 0 & 5 \\ & & & & 0 \end{pmatrix}$$

Las distancias de un individuo  $\omega$  a clasificar a cada una de las muestras son

$$(3, 2, 3, 1, 2) \quad (3, 2, 1, 2, 3)$$

Clasifica  $\omega$ .

## 4.7 Ventajas del método DB

El método de discriminación DB no necesita conocer las funciones de densidad, que a menudo es difícil o imposible de determinar si los datos son mixtos. DB depende solamente del conocimiento de una distancia entre individuos. Para variables mixtas, una elección adecuada es la (1.12), basada en el índice de Gower.

Observa que una vez que hemos calculado las distancias, la estimación de las funciones de proximidad (4.29) es muy sencilla.

En resumen, el método DB:

- Es equivalente al discriminador lineal (4.4) si tomamos la distancia de Mahalanobis. Es fácil ver que una variante de esta distancia proporciona el discriminador cuadrático (4.6).

- Puede abordar variables binarias utilizando (por ejemplo) el coeficiente de Jaccard.
- Puede abordar variables nominales (secuencias de ADN, palabras de un idioma...) mediante la distancia de tipo matching coefficient (coeficiente de coincidencia).
- Puede abordar variables mixtas, considerando el coeficiente de Gower.
- Admite un tratamiento sencillo de datos faltantes.
- Como solamente depende de las distancias, podemos tratar el caso de que tengamos más variables que individuos en alguna de las muestras.

El método DB fue propuesto por Cuadras (1989, 1991) e ilustrado con ejemplos en Cuadras (1992). Los Teoremas 3.1 y 3.2 están demostrados en Cuadras, Fortiana y Oliva (1997), trabajo que contiene otros ejemplos aplicados.

**Ejercicio 4.10** *En un estudio de cáncer que afectó a 137 mujeres, se consideraron 11 variables (7 continuas, 2 binarias y 3 categóricas) capaces de clasificar un tumor en benigno o maligno. En un primer grupo de  $n_1 = 78$  mujeres resultó benigno, mientras que en otro grupo de  $n_2 = 59$  mujeres el tumor era maligno. Con los datos del fichero **qlmk1.dat**, calcula el número de errores de clasificación, utilizando los discriminadores lineal (4.4), cuadrático (4.6) y Euclídeo (4.21), así como la regla DB con la distancia de Gower. ¿Qué discriminador clasifica mejor? (los datos se encuentran también en el fichero **ejercici.dat**).*

**Ejercicio 4.11** *Supongamos que la matriz de covarianzas  $\Sigma$  es singular, es decir, existen combinaciones lineales entre las variables.*

1. *Razona entonces que (4.21) puede ser un discriminador adecuado.*
2. *Explica que, en general, el método DB sería aplicable en tal situación de singularidad.*

## 4.8 Discriminación en el caso de varias poblaciones

Supongamos que disponemos de  $g > 2$  poblaciones  $\Omega_1, \dots, \Omega_g$  y muestras de tamaños  $n_1, \dots, n_g$  para cada población, y que en relación a  $p$  variables (posiblemente mixtas) y a una distancia  $\delta$ , hemos calculado las funciones de proximidad  $\hat{\phi}_1^2(\omega), \dots, \hat{\phi}_g^2(\omega)$  correspondientes a un individuo  $\omega$  a clasificar (véase (4.29)). La generalización de (4.31) es inmediata, así que la regla DB para  $g > 2$  poblaciones es:

$$\text{Asignamos } \omega \text{ a } \Omega_i \text{ si } \hat{\phi}_i^2(\omega) = \min \left\{ \hat{\phi}_1^2(\omega), \dots, \hat{\phi}_g^2(\omega) \right\} \quad (4.34)$$

Se puede probar un resultado similar al Teorema 3.2. Por lo tanto la regla DB clasifica  $\omega$  a la población que tiene más próxima.

Finalmente, si  $\mathbf{x}(k)_1, \dots, \mathbf{x}(k)_{n_k}$  representa una muestra de  $\Omega_k$ , utilizando la siguiente distancia (al cuadrado) entre cada par de poblaciones  $\Omega_k, \Omega_{k'}$

$$\Delta^2(k, k') = \frac{1}{n_k n_{k'}} \sum_{i,j} \delta^2(\mathbf{x}(k)_i, \mathbf{x}(k')_j) - V_\delta(k) - V_\delta(k') \quad (4.35)$$

donde  $V_\delta(k)$  es la variabilidad geométrica de  $\Omega_k$ , tenemos la posibilidad de representar las  $g$  poblaciones por análisis de las coordenadas principales sobre la matriz de orden  $g \times g$ , de distancias  $(\Delta(k, k'))$ . Este método de *análisis canónico generalizado*, que utiliza solamente distancias entre individuos para obtener distancias entre poblaciones, y que por consiguiente puede manejar datos mixtos, se puede combinar con la discriminación DB (Cuadras, 1991; Krzanowski, 1994; Cuadras, Fortiana y Oliva, 1997). Cabe finalmente mencionar que, mediante distancias y en general funciones simétricas, se pueden construir funciones de densidad de probabilidad y que la cota inferior de la variabilidad geométrica es la entropía de Shanon (véase Cuadras, Atkinson y Fortiana, 1997).

**Ejercicio 4.12** *Se pide:*

1. *Demuestra que el análisis canónico generalizado se reduce al análisis canónico de poblaciones (Sección 1.10) cuando la distancia entre individuos es la de Mahalanobis, es decir, prueba que (4.35) se reduce a (1.20).*
2. *Con los datos del fichero **mardt01.dat**, que corresponde a 7 grupos de estudiantes en relación a 2 variables mixtas (Mardia, Kent y Bibby, 1979, p. 294), realiza una representación canónica generalizada con la distancia de Gower. (Los datos pueden encontrarse también en el fichero **ejercici.dat**)*

## Capítulo 5

# Asociación DB y predicción multivariante

### 5.1 Relacionando dos conjuntos de variables

Supongamos que sobre un mismo conjunto finito  $\Omega = \{\omega_1, \dots, \omega_n\}$ , se tienen dos conjuntos de variables  $\mathbf{X}$  e  $\mathbf{Y}$ , posiblemente mixtas. Nuestro propósito es relacionar ambos conjuntos a través de la definición de medidas apropiadas de asociación.

Si se dispone de dos matrices de datos cuantitativos  $\mathbf{X}$  e  $\mathbf{Y}$ , de órdenes  $p \times n$  y  $q \times n$ , resultado de observar  $p$  y  $q$  variables sobre  $n$  individuos, con matrices de correlaciones

$$\begin{array}{cc} & \mathbf{X} & \mathbf{Y} \\ \mathbf{X} & \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{Y} & \mathbf{R}_{21} & \mathbf{R}_{22} \end{array}$$

un método bien conocido es el **análisis de correlaciones canónicas**. En síntesis, este método multivariante resuelve las ecuaciones en autovalores

$$\begin{aligned} \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{a}_i &= r_i^2\mathbf{R}_{11}\mathbf{a}_i \\ \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{b}_i &= r_i^2\mathbf{R}_{22}\mathbf{b}_i \quad i = 1, \dots, m = \min(p, q). \end{aligned} \quad (5.1)$$

Estas ecuaciones también pueden resolverse empleando las matrices de covarianzas, con los mismos resultados.

Los vectores propios  $\mathbf{a}_i, \mathbf{b}_i$ , se denominan *vectores canónicos* y

$$\mathbf{U}_i = \mathbf{a}_i' \mathbf{X}, \quad \mathbf{V}_i = \mathbf{b}_i' \mathbf{Y}, \quad i = 1, \dots, m,$$

son las variables canónicas cuyas correlaciones, llamadas *correlaciones canónicas*,

$$r_i = \text{corr}(\mathbf{U}_i, \mathbf{V}_i), \quad i = 1, \dots, m$$

son las raíces cuadradas de los valores propios

$$r_1^2 \geq r_2^2 \geq \dots \geq r_m^2. \quad (5.2)$$

La primer correlación canónica  $r_1$  es la máxima correlación entre una función lineal de  $\mathbf{X}$  y una función lineal de  $\mathbf{Y}$ . Las variables canónicas  $\mathbf{U}_1, \dots, \mathbf{U}_m$  son incorrelacionadas, así como  $\mathbf{V}_1, \dots, \mathbf{V}_m$ .

Medidas globales de asociación pueden basarse en (5.2), por ejemplo:

$$\theta_1^2 = \prod_{i=1}^m r_i^2 \quad (5.3)$$

$$\theta_2^2 = 1 - \prod_{i=1}^m (1 - r_i^2) \quad (5.4)$$

Si se consideran las matrices de covarianzas

$$\begin{array}{cc} & \begin{matrix} \mathbf{X} & \mathbf{Y} \end{matrix} \\ \begin{matrix} \mathbf{X} \\ \mathbf{Y} \end{matrix} & \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \end{array}$$

otra medida de asociación (Escoufier, 1973) es:

$$\text{RV}(\mathbf{X}, \mathbf{Y}) = \text{tr}(\mathbf{S}_{21}\mathbf{S}_{12}) / (\text{tr}(\mathbf{S}_{11}^2)\text{tr}(\mathbf{S}_{22}^2))^{1/2} \quad (5.5)$$

Si se trabaja con las matrices de correlaciones, esta medida se reduce a

$$\text{RV}(\mathbf{X}, \mathbf{Y}) = \left( \sum_{i=1}^m r_i^2 \right) / m. \quad (5.6)$$

$\text{RV}(\mathbf{X}, \mathbf{Y})$  está relacionada con el llamado estadístico Procrustes  $R^2$ , que mide el grado de ajuste entre dos matrices:

$$R^2 = 1 - \{\text{tr}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})^{1/2}\}^2 / \{\text{tr}(\mathbf{X}'\mathbf{X})\text{tr}(\mathbf{Y}'\mathbf{Y})\}.$$

Es también interesante el coeficiente del vector alineación de Hotelling

$$K = \frac{\left| \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \right|}{|\mathbf{S}_{11}| |\mathbf{S}_{22}|} = \frac{|\mathbf{S}_{11}| |\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}|}{|\mathbf{S}_{11}| |\mathbf{S}_{22}|} = \prod_{i=1}^p (1 - r_i^2),$$

que proporciona el llamado *coeficiente del vector de correlación*:

$$R_{xy}^2 = 1 - K = 1 - \prod_{i=1}^p (1 - r_i^2).$$

**Ejercicio 5.1** Se pide:

1. Demuestra que las dos ecuaciones propias (5.1) tienen los mismos valores propios.
2. Comprueba que  $0 \leq \theta_1^2 \leq 1$  y discute los casos  $\theta_1^2 = 0$  y  $\theta_2^2 = 1$ .

La asociación entre variables categóricas puede ser determinada empleando *dual scaling* o *análisis de correspondencias*, un método que presenta una clara relación con el de análisis de correlaciones canónicas (Mardia, Kent y Bibby, 1979, p.237-239 y 290-293; Greenacre, 1984).

## 5.2 Relación DB entre variables mixtas

Supongamos que sobre un mismo conjunto  $\Omega = \{\omega_1, \dots, \omega_n\}$  se han observado dos conjuntos de datos mixtos  $M_1, M_2$ . Se desea relacionar  $M_1$  con  $M_2$ . Una forma de lograrlo es haciendo una extensión de la regresión DB. Suponiendo que se dispone de una función de distancia  $\delta$ , que con los datos  $M_1$  y  $M_2$  proporcione dos matrices  $n \times n$  de distancias,  $\mathbf{\Delta}_1 = (\delta_{ij}(1))$ ,  $\mathbf{\Delta}_2 = (\delta_{ij}(2))$ . Sean  $\mathbf{B}_1, \mathbf{B}_2$  las correspondientes matrices de productos internos (Teorema 1.2) y consideremos las descomposiciones espectrales:

$$\mathbf{B}_1 = \mathbf{U}_1 \mathbf{\Lambda}_1 \mathbf{U}_1' \quad \mathbf{B}_2 = \mathbf{U}_2 \mathbf{\Lambda}_2 \mathbf{U}_2',$$

que proporcionan las matrices de coordenadas principales

$$\mathbf{X} = \mathbf{U}_1 \mathbf{\Lambda}_1^{1/2} \quad \mathbf{Y} = \mathbf{U}_2 \mathbf{\Lambda}_2^{1/2}.$$

Entonces  $M_1, M_2$  pueden representarse mediante las matrices cuantitativas  $\mathbf{X}, \mathbf{Y}$ . Empleando las filas de  $\mathbf{X}$ , coordenadas principales relativas a  $\mathbf{\Delta}_1$ , se puede representar  $\Omega$  en un espacio Euclídeo de menor dimensión. Las filas de  $\mathbf{U}_1$  se denominan *coordenadas estándar* (Greenacre, 1984). Similarmente, las coordenadas principales  $\mathbf{Y}$ , también nos permiten representar  $\Omega$  y las coordenadas estándar son las filas de  $\mathbf{U}_2$ .

Para asociar  $M_1$  con  $M_2$ , a través de la relación de  $\mathbf{X}$  con  $\mathbf{Y}$ , primeramente se deben seleccionar las principales dimensiones predictibles de las columnas de

$$\mathbf{X} = [X_1, \dots, X_{m_1}] \quad \mathbf{Y} = [Y_1, \dots, Y_{m_2}]$$

donde  $m_1 = \text{rang}(\mathbf{B}_1)$ ,  $m_2 = \text{rang}(\mathbf{B}_2)$ .

Podemos introducir dos coeficientes  $c_x(k), c_y(k)$  tales que  $c_x(0) = c_y(0)$  y, para  $k > 0$ ,

$$\begin{aligned} c_x(k) &= \frac{\sum_{i=1}^k Y_i' \mathbf{B}_1 Y_i / \text{tr}(\mathbf{B}_2 \mathbf{B}_1)}{\sum_{i=1}^k X_i' \mathbf{B}_1 X_i / \text{tr}(\mathbf{B}_1 \mathbf{B}_2)} \\ c_y(k) &= \end{aligned} \quad (5.7)$$

que satisfacen

$$0 \leq c_x(k) \leq 1, \quad 0 \leq c_y(k) \leq 1, \quad k = 0, 1, 2, \dots \quad (5.8)$$

En la predicción de  $\mathbf{Y}$ , una dimensión  $i+1$ , sea  $X_{i+1}$ , es no predictiva si  $c_x(i) = c_x(i+1)$ . La dimensión de corte es  $m_1^*$  tal que  $c_x(i)$  esté cercano a 1, o bien que  $c_x(i)$  brinde un valor estable para  $i > m_1^*$ . Un argumento similar se puede sostener para el caso de predecir  $\mathbf{X}$  a partir de  $\mathbf{Y}$ .

Un gráfico conjunto de los puntos

$$(k, c_x(k)), \quad (k, c_y(k)), \quad k = 0, 1, 2, 3, \dots$$

puede ayudarnos a decidir cómo y cuantas dimensiones deben ser seleccionadas.

Supongamos ahora que hemos seleccionado  $p$  y  $q$  dimensiones principales y conservemos la notación

$$\mathbf{X} = [X_1, \dots, X_p] \quad \mathbf{Y} = [Y_1, \dots, Y_q], \quad (5.9)$$

es decir, las columnas de  $\mathbf{X}, \mathbf{Y}$  contienen los vectores propios seleccionados de  $\mathbf{B}_1, \mathbf{B}_2$ . Las correspondientes coordenadas estándar pueden ser también indicadas como  $\mathbf{U}_1, \mathbf{U}_2$ . Podemos suponer  $p \leq q$ .

Una medida global de asociación es, entonces

$$\eta^2 = \det(\mathbf{U}'_1 \mathbf{U}_2 \mathbf{U}'_2 \mathbf{U}_1), \quad (5.10)$$

que indicaremos por  $\eta^2(\mathbf{X}, \mathbf{Y})$ .

Este coeficiente satisface las siguientes propiedades:

1.  $0 \leq \eta^2(\mathbf{X}, \mathbf{Y}) = \eta^2(\mathbf{Y}, \mathbf{X}) \leq 1$ .
2.  $\eta^2(\mathbf{X}, \mathbf{Y}) = \det(\mathbf{X}' \mathbf{Y} \mathbf{Y}' \mathbf{X}) / [\det(\mathbf{X}' \mathbf{X}) \det(\mathbf{Y}' \mathbf{Y})]$ .
3.  $\eta^2(\mathbf{X}, \mathbf{Y})$  no depende de las matrices  $\mathbf{X}, \mathbf{Y}$  que dan lugar a las distancias  $\Delta_1, \Delta_2$ .
4.  $r^2(\mathbf{y}, \mathbf{x}) = \eta^2(\mathbf{y}, \mathbf{x})$  donde  $r$  es el coeficiente de correlación simple entre los vectores de datos  $\mathbf{x}$  e  $\mathbf{y}$ .
5.  $R^2(\mathbf{y}, \mathbf{X}) = \eta^2(\mathbf{y}, \mathbf{X})$  donde  $R$  es el coeficiente de correlación múltiple entre  $\mathbf{y}$  y la matriz  $\mathbf{X}$ .
6. Si  $r_j, j = 1, \dots, p$  son los coeficientes de correlación canónica entre  $\mathbf{X}$  e  $\mathbf{Y}$ , entonces

$$\eta^2(\mathbf{X}, \mathbf{Y}) = \prod_{j=1}^p r_j^2.$$

**Ejercicio 5.2** *Se pide:*

1. *Prueba que los coeficientes (5.7) satisfacen (5.8).*
2. *Prueba las cinco primeras propiedades del coeficiente  $\eta^2(\mathbf{X}, \mathbf{Y})$ .*
3. *Prueba que  $RV(\mathbf{X}, \mathbf{Y})$  en términos de distancias se puede escribir como  $RV(\mathbf{X}, \mathbf{Y}) = \text{tr}(\mathbf{B}_2 \mathbf{B}_1) / (\text{tr}(\mathbf{B}_1^2) \text{tr}(\mathbf{B}_2^2))^{1/2}$ .*

### 5.3 Predicción con correlaciones canónicas

Consideremos las matrices  $\mathbf{X}, \mathbf{Y}$  que contienen las dimensiones principales predictivas (5.9). Las matrices de covarianzas correspondientes a  $\mathbf{X}, \mathbf{Y}$  son (omitiendo la constante  $n$ ):

$$\mathbf{S}_{11} = \Lambda_1 \quad \mathbf{S}_{22} = \Lambda_2 \quad \mathbf{S}_{12} = \mathbf{X}' \mathbf{Y}. \quad (5.11)$$

La matriz de correlaciones entre  $\mathbf{X}$  e  $\mathbf{Y}$  es

$$\mathbf{R}_{12} = \mathbf{U}'_1 \mathbf{U}_2, \quad (5.12)$$



donde  $\mathbf{U}_1, \mathbf{U}_2$  son las coordenadas estándar. A partir de (5.1), empleando las matrices de covariancias, los vectores y correlaciones canónicos son las soluciones de las ecuaciones propias

$$\begin{aligned} \mathbf{X}'\mathbf{Y}\mathbf{\Lambda}_2^{-1}\mathbf{Y}'\mathbf{X}\mathbf{a}_i &= r_i^2 \mathbf{\Lambda}_1 \mathbf{a}_i \\ \mathbf{Y}'\mathbf{X}\mathbf{\Lambda}_1^{-1}\mathbf{X}'\mathbf{Y}\mathbf{b}_i &= r_i^2 \mathbf{\Lambda}_2 \mathbf{b}_i \end{aligned} \quad (5.13)$$

Las relaciones entre las matrices y los vectores canónicos

$$\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_p] \quad \mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_p]$$

son

$$\mathbf{A} = \mathbf{\Lambda}_1^{-1}\mathbf{X}'\mathbf{Y}\mathbf{B}\mathbf{D}_r^{-1} \quad \mathbf{B} = \mathbf{\Lambda}_2^{-1}\mathbf{Y}'\mathbf{X}\mathbf{A}\mathbf{D}_r^{-1} \quad (5.14)$$

donde  $\mathbf{D}_r = \text{diag}(r_1, \dots, r_p)$  contiene las correlaciones canónicas. Dichas correlaciones canónicas pueden también obtenerse a partir de las matrices de correlación. Como  $\mathbf{R}_{11} = \mathbf{R}_{22} = \mathbf{I}$ , de (5.1)

$$\mathbf{R}_{12}\mathbf{R}_{21}\mathbf{a}_i = r_i^2 \mathbf{a}_i \quad (5.15)$$

Por lo tanto la medida global (5.10) se puede expresar como

$$\eta^2 = \prod_{i=1}^p r_i^2. \quad (5.16)$$

**Ejercicio 5.3** *Se pide:*

1. Demuestra que las dos ecuaciones (5.13) proporcionan los mismos valores propios.
2. Demuestra las relaciones (5.14) y prueba (5.16).
3. Expresa el coeficiente  $K$  de Hotelling en términos de  $\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \mathbf{U}_1, \mathbf{U}_2$ .

**Ejercicio 5.4** *La tabla 5.1 relaciona países según que la relación comercial sea significativa (tabla inferior, 1 si lo es, 0 si no lo es) con el número de artículos científicos publicados conjuntamente (tabla superior). Calcula el coeficiente de asociación  $\eta^2$ .*

La predicción canónica de  $\mathbf{Y}$  a partir de  $\mathbf{X}$  es

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{A}\mathbf{D}_r\mathbf{B}^{-1} \quad (5.17)$$

es decir,  $\hat{\mathbf{Y}}$  es la proyección de  $\mathbf{Y}$  (proyección ortogonal de  $\mathbf{Y}$  en  $\mathbf{X}$ ).

Finalmente vamos a predecir  $Y$  sobre un nuevo individuo. Sea éste  $\omega_{n+1}$  con distancias (al cuadrado)  $\delta_1^2, \dots, \delta_n^2$ , a cada uno de los  $n$  individuos de  $\Omega$  obtenidas a partir de las variables explicativas  $X$ . Sea  $\mathbf{b}$  el vector columna que contiene los términos de la diagonal de  $\mathbf{B}_1$ . Indiquemos

$$\mathbf{d} = (\delta_1^2, \dots, \delta_n^2)'$$

	USA	Spa	Fra	U.K.	Ital	Ger	Can	Jap	Chin	Rus
USA	63446	692	2281	2507	1642	2812	2733	1039	1773	893
Spain	1	8597	473	347	352	278	163	69	104	177
France	1	1	17155	532	916	884	496	269	167	606
U.K.	1	1	1	12585	490	810	480	213	339	365
Italy	0	1	1	0	13197	677	290	169	120	512
Germany	1	1	1	1	1	16588	499	350	408	984
Canada	1	0	0	1	0	0	7927	228	601	204
Japan	1	0	0	0	0	0	1	20001	371	193
China	1	0	0	0	0	0	1	1	39140	64
Russia	0	0	0	0	0	0	0	0	1	18213

Tabla 5.1: Realción comercial (1 si es significativa, 0 si no lo es) y relación científica (número de artículos de matemáticas o estadística en colaboración durante 1996-2002) entre diez países.

Entonces la fórmula de añadir un punto

$$\mathbf{x} = \frac{1}{2} \mathbf{\Lambda}_1^{-1} \mathbf{X}'(\mathbf{b} - \mathbf{d}) \quad (5.18)$$

nos brinda las coordenadas  $\mathbf{x} = (x_1, \dots, x_n)'$  de  $\omega_{n+1}$ . La predicción de  $\mathbf{y}$  dada  $\mathbf{x}$  es entonces  $\hat{\mathbf{y}}_{n+1} = (y_1, \dots, y_p)'$  tal que

$$\begin{aligned} \hat{\mathbf{y}}_{n+1} &= \frac{1}{2}(\mathbf{b} - \mathbf{d})\mathbf{X}\mathbf{\Lambda}_1^{-1}\mathbf{A}\mathbf{D}_r\mathbf{B}^{-1} \\ &= \frac{1}{2}(\mathbf{b} - \mathbf{d})\mathbf{X}\mathbf{\Lambda}_1^{-1}\mathbf{X}'\mathbf{Y} \end{aligned} \quad (5.19)$$

Existen diversos procedimientos posibles para predecir las variables mixtas  $M_2$  dados los valores  $M_1 = m_1$ , a partir de (5.19). Un camino muy simple consiste en asignar la información mixta  $m_1$ , de un individuo  $w_i$  con coordenadas  $y_i$  tal que

$$d_E(y_i, \hat{y}_{n+1}) = \min \{d_E(y_i, \hat{y}_{n+1}), \dots, d_E(y_n, \hat{y}_{n+1})\},$$

donde  $d_E$  es la distancia Euclídea.

## 5.4 Planteamiento mediante related MDS

Otro camino para combinar la información dada por dos matrices de distancias  $\mathbf{\Delta}_1 = (\delta_1(i, j))$ ,  $\mathbf{\Delta}_2 = (\delta_2(i, j))$  es definir una distancia conjunta  $\delta_{12}$  que extraiga la redundancia entre ambos conjuntos de variables mixtas.

Sean  $\mathbf{x}_i, \mathbf{y}_i$  los vectores de coordenadas (filas), obtenidos a partir de las distancias  $\mathbf{\Delta}_1, \mathbf{\Delta}_2$  respectivamente. La matriz de distancias conjuntas  $\mathbf{\Delta}_{12} = (\delta_{12}(i, j))$  se define como

$$\delta_{12}^2(i, j) = \delta_1^2(i, j) + \delta_2^2(i, j) - \tau_{12}(i, j) \quad (5.20)$$

donde

$$\tau_{12}(i, j) = (\mathbf{x}_i - \mathbf{x}_j) \mathbf{\Lambda}_1^{-1/2} \mathbf{X}' \mathbf{Y} \mathbf{\Lambda}_2^{-1/2} (\mathbf{y}_i - \mathbf{y}_j)' \quad (5.21)$$

Algunas propiedades de  $\delta_{12}$  son:

1. Si  $\delta_1 \equiv 0$ , entonces  $\delta_{12} \equiv \delta_2$ .
2. Si  $\delta_2 \equiv 0$ , entonces  $\delta_{12} \equiv \delta_1$ .
3. Si  $\delta_1 \equiv \delta_2$ , entonces  $\delta_{12} \equiv \delta_1 \equiv \delta_2$ .
4. Si  $\delta_1, \delta_2$  son ortogonales, es decir  $\mathbf{S}_{12} = \mathbf{X}' \mathbf{Y} = \mathbf{0}$ , entonces  $\delta_{12}^2 = \delta_1^2 + \delta_2^2$ .
5. Si  $\delta_2(i, j) = c$ ,  $i \neq j$ , donde  $c$  es una constante, entonces  $\delta_{12}$  y  $\delta_1$  tienen la misma preordenación.
6. Si  $\delta_1$  y  $\delta_2$  tienen un eje principal común  $k$ , entonces  $\delta_{12}$  conserva este mismo eje, con valor propio asociado

$$\lambda_k = \lambda_k(1) + \lambda_k(2) - (\lambda_k(1)\lambda_k(2))^{1/2}, \quad (5.22)$$

donde  $\lambda_k(1), \lambda_k(2)$  son los valores propios relacionados a  $\delta_1, \delta_2$ .

Las propiedades anteriores demuestran que  $\delta_{12}$  no cambia si las distancias son iguales, es una distancia aditiva en condiciones de ortogonalidad, preserva los ejes principales comunes, etc. El término  $\tau_{12}$  encierra la redundancia entre los dos conjuntos de datos.

La matriz de productos internos relacionada a  $\delta_{12}$  es

$$\mathbf{B}_{12} = \mathbf{B}_1 + \mathbf{B}_2 - \frac{1}{2}(\mathbf{B}_1^{1/2} \mathbf{B}_2^{1/2} + \mathbf{B}_2^{1/2} \mathbf{B}_1^{1/2}) \quad (5.23)$$

donde  $\mathbf{B}^{1/2} = \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{U}'$ . Las coordenadas principales conjuntas  $\mathbf{Z}$  se obtienen a partir de

$$\mathbf{B}_{12} = \mathbf{U}_{12} \mathbf{\Lambda}_{12} \mathbf{U}_{12}' = \mathbf{Z} \mathbf{Z}' \quad (5.24)$$

siendo

$$\mathbf{Z} = \mathbf{U}_{12} \mathbf{\Lambda}_{12}^{1/2} \quad (5.25)$$

Una interpretación de igualdad y ortogonalidad considerando las coordenadas  $\mathbf{Z}$  es:

- Si  $\delta_1 \equiv \delta_2$ , entonces  $\mathbf{Z} = \mathbf{X} = \mathbf{Y}$ .
- Si  $\delta_1$  y  $\delta_2$  son ortogonales, entonces  $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ .

La representación de  $\Omega$  a través de las coordenadas conjuntas  $\mathbf{Z}$ , es la llamada representación mediante *related metric scaling* de  $\Omega$ . Véase Cuadras y Fortiana (1997a), Cuadras (1998).

Si  $\mathbf{V}_{xy}, \mathbf{V}_x, \mathbf{V}_y$  son las variabilidades geométricas (véase (3.1)) correspondientes a  $\delta_{12}, \delta_1, \delta_2$ , entonces se puede probar que

$$\mathbf{V}_{xy} \leq \mathbf{V}_x + \mathbf{V}_y \quad (5.26)$$

Introduciendo la razón

$$\xi = \frac{\mathbf{V}_{xy}}{\mathbf{V}_x + \mathbf{V}_y} = \frac{\text{tr}(\mathbf{B}_{12})}{\text{tr}(\mathbf{B}_1 + \mathbf{B}_2)}, \quad (5.27)$$

otra medida de asociación global es

$$\alpha = 2(1 - \xi) \quad (5.28)$$

con las siguientes propiedades:

1.  $0 \leq \alpha \leq 1$ .
2. Si  $\delta_1 \equiv \delta_2$ , entonces  $\mathbf{V}_{xy} = \mathbf{V}_x = \mathbf{V}_y$ ,  $\alpha = 1$ .
3. Si  $\delta_1$  y  $\delta_2$  son ortogonales, entonces  $\mathbf{V}_{xy} = \mathbf{V}_x + \mathbf{V}_y$ ,  $\alpha = 0$ .
4. Normalizando  $\mathbf{B}_1$  y  $\mathbf{B}_2$  de modo que  $\text{tr}(\mathbf{B}_1) = \text{tr}(\mathbf{B}_2) = \text{tr}(\mathbf{\Lambda}_1^2) = \text{tr}(\mathbf{\Lambda}_2^2) = 1$ , entonces  $\alpha = \text{tr}(\mathbf{U}_1 \mathbf{\Lambda}_1 \mathbf{U}_1' \mathbf{U}_2 \mathbf{\Lambda}_2 \mathbf{U}_2')$ .

Un cuadro resumen de las medidas de asociación basadas en distancias en función de las coordenadas estándar  $\mathbf{U}_1, \mathbf{U}_2$  y las matrices de productos internos  $\mathbf{B}_1, \mathbf{B}_2$  es el siguiente:

$$\begin{aligned} \eta^2 &= \det(\mathbf{U}_1' \mathbf{U}_2 \mathbf{U}_2' \mathbf{U}_1) \\ RV &= \text{tr}(\mathbf{B}_2 \mathbf{B}_1) / (\text{tr}(\mathbf{B}_1^2) \text{tr}(\mathbf{B}_2^2))^{1/2} \\ \alpha &= 2[1 - \text{tr}(\mathbf{B}_{12}) / \text{tr}(\mathbf{B}_1 + \mathbf{B}_2)]. \end{aligned}$$

Para más detalles y generalizaciones, véase Cuadras y Fortiana (1997b), Arenas y Cuadras (2004).

**Ejercicio 5.5** *Se pide:*

1. Demuestra que (5.23) es la matriz de productos internos relativa a la distancia conjunta (5.19).
2. Demuestra (5.26) y las propiedades de  $\alpha$ , véase (5.28).
3. Demuestra la fórmula  $\alpha = \text{tr}(\mathbf{U}_1 \mathbf{\Lambda}_1 \mathbf{U}_1' \mathbf{U}_2 \mathbf{\Lambda}_2 \mathbf{U}_2')$ .

## Capítulo 6

# Comparación DB de poblaciones

### 6.1 Comparando dos conjuntos de datos mixtos

Sean dos poblaciones  $\Omega_1, \Omega_2$  y  $\xi$  un vector aleatorio mixto con valores en  $\mathbf{S}_\xi$ . Supongamos que se dispone de dos muestras de tamaño adecuado  $n_1$  y  $n_2$  provenientes de  $\Omega_1$  y  $\Omega_2$ . Sobre la base de  $\xi$ , ambas muestras proporcionan dos conjuntos de datos  $\mathbf{M}_1, \mathbf{M}_2$ . Sea  $\delta$  una función de distancia en  $\mathbf{S}_\xi$ , que nos brinda las distancias entre observaciones  $\xi$ , posiblemente diferentes según que los individuos provengan de  $\Omega_1$  o  $\Omega_2$  (véase Sección 3.3) y supongmos que se dispone de dos aplicaciones (*embeddings*)

$$\Psi_i : \mathbf{S}_\xi \rightarrow \mathbf{L} \quad i = 1, 2.$$

donde  $\mathbf{L}$  es un espacio Euclídeo (o de Hilbert), que satisface (4.11).  $\Psi_i$  depende de  $\Omega_i, i = 1, 2$ .

El problema consiste en comparar la distribución de  $\xi$  en  $\Omega_1$  con la distribución de  $\xi$  en  $\Omega_2$ , sobre la base de  $\mathbf{M}_1, \mathbf{M}_2$  y la distancia  $\delta$ . Esta comparación se realiza comprobando si podemos aceptar la hipótesis nula

$$H_0 : E(\Psi_1(\xi)) = E(\Psi_2(\xi)) \quad (6.1)$$

Suponiendo que mediante  $\delta$  y calculando las distancias sobre la base de los datos  $\mathbf{M}_1, \mathbf{M}_2$ , se obtienen dos matrices  $n_i \times n_i$  de interdistancias  $\Delta_{11}, \Delta_{22}$ , así como la matriz  $n_1 \times n_2$  de interdistancias  $\Delta_{12}$ . Nuestro propósito es encontrar un procedimiento DB para decidir sobre un test del tipo

$$H_0 : \Delta^2(\Omega_1, \Omega_2) = 0,$$

donde  $\Delta^2(\Omega_1, \Omega_2)$  es una medida de distancia entre las dos poblaciones.

## 6.2 Planteamiento mediante coordenadas principales

Si consideramos la matriz de superdistancias

$$\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}$$

y  $\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}^{1/2}$  es la correspondiente matriz de coordenadas principales particionada como

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix}$$

donde  $\mathbf{Z}_1(n_1 \times p)$  se relaciona con  $\Delta_{11}$ ,  $\mathbf{Z}_2(n_2 \times p)$  se relaciona con  $\Delta_{22}$ . Las distancias Euclídeas entre las filas de  $\mathbf{Z}_1$  y las filas de  $\mathbf{Z}_2$  reproducen  $\Delta_{12}$ . Se puede suponer que la dimensión efectiva  $p$  ha sido obtenida, empleando las ideas de la Sección 4.5.

$\mathbf{Z}_1$  y  $\mathbf{Z}_2$  se pueden interpretar como dos matrices de datos “independientes”, relacionadas con  $\mathbf{M}_1, \mathbf{M}_2$ , con vectores de medias  $\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2$  y matrices de covarianza  $\mathbf{S}_1, \mathbf{S}_2$ .

Sea

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$$

la matriz de producto interno relacionada con  $\Delta$ . Entonces  $\mathbf{B} = \mathbf{Z}\mathbf{Z}'$  y los vectores de medias de  $\mathbf{Z}_1, \mathbf{Z}_2$  verifican las restricciones

$$n_1\bar{\mathbf{z}}_1 + n_2\bar{\mathbf{z}}_2 = \mathbf{0},$$

y

$$\mathbf{Z}_1'\mathbf{Z}_2 = \mathbf{B}_{12}. \quad (6.2)$$

Una test para decidir  $H_0$  puede basarse en el estadístico

$$T_0^2 = (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)' \mathbf{V}^{-1} (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)$$

donde

$$\mathbf{V} = (n_1\mathbf{S}_1 + n_2\mathbf{S}_2) = \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{Z}_2'\mathbf{Z}_2 = \mathbf{\Lambda},$$

Así

$$(\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)' \mathbf{\Lambda}^{-1} (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2) = \|\bar{\mathbf{u}} - \bar{\mathbf{v}}\|^2$$

donde  $\bar{\mathbf{u}}$  y  $\bar{\mathbf{v}}$  son los vectores de medias de las correspondientes coordenadas estándar.

En general, la distribución de  $T_0^2$  será desconocida. Un test de permutaciones consiste en obtener las

$$N = \frac{(n_1 + n_2)!}{n_1!n_2!}$$

permutaciones de las filas de  $\mathbf{Z}$ , obteniendo los diferentes  $T_0^2$  y ordenándolos

$$T_1^2 \leq \dots \leq T_k^2 \leq \dots \leq T_N^2$$

donde  $k$  se relaciona con el nivel de significancia  $\alpha$  mediante

$$\frac{N - k}{N} = \alpha$$

Decidimos entonces aceptar  $H_0$  si  $T_0^2 \leq T_k^2$  y rechazar  $H_0$  si  $T_0^2 > T_k^2$ . Si  $N$  resulta demasiado grande, podemos tomar  $N' < N$  permutaciones y proceder con los  $N'$  valores obtenidos.

Un tests alternativo podría estar basado en

$$E^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)' (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2).$$

Cuando la distancia es tipo Mahalanobis, se puede probar que este test es equivalente a un test basado en la distancia (4.30).

**Teorema 6.1** Sean  $\mathbf{x}'_1, \dots, \mathbf{x}'_{n_1}, \mathbf{y}'_1, \dots, \mathbf{y}'_{n_2}$  muestras multivariantes, filas de las matrices de datos  $\mathbf{X}, \mathbf{Y}$ , con vectores de medias  $\bar{\mathbf{x}}, \bar{\mathbf{y}}$ . Supongamos que las distancias (al cuadrado) entre observaciones vienen dadas por

$$d^2(\mathbf{a}, \mathbf{b}) = (\mathbf{a} - \mathbf{b})' \mathbf{M}^{-1} (\mathbf{a} - \mathbf{b}), \quad (6.3)$$

donde  $\mathbf{a}, \mathbf{b} \in \{\mathbf{x}_1, \dots, \mathbf{x}_{n_1}, \mathbf{y}_1, \dots, \mathbf{y}_{n_2}\}$  y  $\mathbf{M}$  es una matriz definida simétrica definida positiva. Sea

$$\hat{\Delta}^2(\Omega_1, \Omega_2) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d^2(\mathbf{x}_i, \mathbf{y}_j) - \hat{V}_d(\mathbf{X}) - \hat{V}_d(\mathbf{Y})$$

una medida de distancia entre las dos poblaciones. Entonces:

$$\hat{\Delta}^2(\Omega_1, \Omega_2) = (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)' (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2) = (\bar{\mathbf{x}} - \bar{\mathbf{y}})' \mathbf{M}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}}).$$

**Ejercicio 6.1** Se pide.

1. Determina la distribución de  $T_0^2$  cuando  $\Omega_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$   $i = 1, 2$  y  $\mu_1 = \mu_2$ .
2. Demuestra (6.2).
3. Sugiere una alternativa para  $T_0^2$ .

### 6.3 Planteamiento mediante funciones de proximidad

Sean  $\phi_1^2, \phi_2^2$  las funciones de proximidad, véase (4.14). Bajo  $H_0 : \Omega_1 = \Omega_2$  se tiene que  $\phi_1^2 \equiv \phi_2^2$ . Por lo tanto bajo  $H_0$  se puede asignar un nuevo individuo a  $\Omega_1$  o a  $\Omega_2$  con la misma probabilidad. Sin embargo, las funciones de probabilidad estimadas serán diferentes, pero se tiene

$$P(\hat{\phi}_1^2(\mathbf{x}) < \hat{\phi}_2^2(\mathbf{x}) \mid \Omega_i) = \frac{1}{2} \quad i = 1, 2 \quad (6.4)$$

Supongamos que con muestras  $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}, \mathbf{y}_1, \dots, \mathbf{y}_{n_2}$  de  $\Omega_1, \Omega_2$ , respectivamente, hemos obtenido los siguientes valores de las funciones de proximidad:

Muestra	$\hat{\phi}_1^2$	$\hat{\phi}_2^2$	Muestra	$\hat{\phi}_1^2$	$\hat{\phi}_2^2$
$\mathbf{x}_1$	$a_{11}$	$a_{12}$	$\mathbf{y}_1$	$b_{11}$	$b_{12}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{n_1}$	$a_{n_11}$	$a_{n_12}$	$\mathbf{y}_{n_2}$	$b_{n_21}$	$b_{n_22}$

Proponemos dos métodos a partir de los valores  $a_{ij}, b_{ij}$ .

#### 6.3.1 Prueba ji-cuadrado

Las asignaciones son:

$$\begin{aligned} \mathbf{x}_i &\rightarrow \Omega_1 & \text{si } a_{i1} - a_{i2} < 0 \\ \mathbf{x}_i &\rightarrow \Omega_2 & \text{si } a_{i1} - a_{i2} > 0 \\ \mathbf{y}_i &\rightarrow \Omega_1 & \text{si } b_{i1} - b_{i2} < 0 \\ \mathbf{y}_i &\rightarrow \Omega_2 & \text{si } b_{i1} - b_{i2} > 0 \end{aligned}$$

y dado que, si  $\Omega_1 = \Omega_2$ ,

$$P(a_{i1} < a_{i2}) = 1/2, \quad P(b_{i1} < b_{i2}) = 1/2,$$

si  $n_{ij}$  es la frecuencia de individuos de  $\Omega_i$  asignados a  $\Omega_j$ , obtenemos la siguiente tabla de frecuencias observadas y esperadas

	Observadas		Esperadas	
	$\Omega_1$	$\Omega_2$	$\Omega_1$	$\Omega_2$
$\Omega_1$	$n_{11}$	$n_{12}$	$n_1/2$	$n_2/2$
$\Omega_2$	$n_{21}$	$n_{22}$	$n_1/2$	$n_2/2$

donde  $n_1 = n_{11} + n_{12}$ ,  $n_2 = n_{21} + n_{22}$ . Entonces el estadístico

$$\chi^2 = \frac{(n_{11} - n_1/2)^2}{n_1/2} + \frac{(n_{12} - n_1/2)^2}{n_1/2} + \frac{(n_{21} - n_2/2)^2}{n_2/2} + \frac{(n_{22} - n_2/2)^2}{n_2/2}$$

sigue una distribución ji-cuadrado con 2 g.l. y puede utilizarse para contrastar  $H_0$ .



### 6.3.2 Prueba de Mann-Whitney

Sean

$$X = a_1 - a_2, \quad Y = b_1 - b_2,$$

es decir,  $X, Y$  son las diferencias de los valores de las funciones de proximidad para la primera y segunda población. Entonces si

$$Z = Y - X$$

$$\text{Si } \Omega_1 = \Omega_2 \quad E(Z) = 0,$$

$$\text{Si } \Omega_1 \neq \Omega_2 \quad E(Z) > 0.$$

Podemos entonces utilizar el test de Mann-Whitney para comparar las dos muestras

$$x_i = a_{i1} - a_{i2}, \quad i = 1, \dots, n_1, \quad y_j = b_{j1} - b_{j2}, \quad j = 1, \dots, n_2.$$

**Ejercicio 6.2** *A partir de los datos mixtos sobre cáncer del ejercicio 3.2, compare los grupos de  $n_1 = 78$  (benignos) y  $n_2 = 59$  (malignos) aplicando:*

1. *Análisis de coordenadas principales.*
2. *Funciones de proximidad.*

## 6.4 Comparando muestras múltiples

Supongamos que  $\mathcal{D}_1, \dots, \mathcal{D}_g$  son  $g \geq 2$  conjuntos de datos independientes provenientes de las poblaciones  $\Pi_1, \dots, \Pi_g$ . Los datos pueden ser de tipo general (cuantitativos, cualitativos, nominales, mixtos). Estamos interesados en el test

$$H_0 : \Pi_1 = \dots = \Pi_g.$$

Bajo  $H_0$  todos los datos provienen de una misma distribución yacente.

Empecemos suponiendo que mediante una función de distancia entre observaciones, podemos obtener las matrices de intra-distancias  $\Delta_{11}, \dots, \Delta_{gg}$ , y la matrices de inter-distancias  $\Delta_{12}, \dots, \Delta_{g-1g}$ . De este modo tenemos la matriz  $n \times n$  de distancias

$$\Delta = \begin{bmatrix} \Delta_{11} & \cdots & \Delta_{1g} \\ \vdots & \ddots & \vdots \\ \Delta_{g1} & \cdots & \Delta_{gg} \end{bmatrix},$$

donde  $\Delta_{ij}$  es  $n_i \times n_j$ , siendo  $n = n_1 + \dots + n_g$ .

Seguidamente obtenemos, via análisis de coordenadas principales, las matrices  $\mathbf{G}$  y  $\mathbf{X}$  tal que  $\mathbf{G} = \mathbf{X}\mathbf{X}'$ . Expresamos la super-matriz  $\mathbf{X}$  como

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_g \end{bmatrix}.$$

Las distancias Euclideas entre las filas de  $\mathbf{X}_i$  y  $\mathbf{X}_{i'}$  proporcionan  $\Delta_{ii'}$ . Así las matrices  $\mathbf{X}_1, \dots, \mathbf{X}_g$  representan los  $g$  conjuntos de datos mixtos, que a continuación compararemos para contrastar  $H_0$ .

### 6.4.1 Partición de la variabilidad geométrica

Las filas  $\mathbf{x}_1, \dots, \mathbf{x}_n$  de cualquier matriz  $\mathbf{X}$  de datos multivariantes de orden  $N \times p$  verifican

$$\sum_{i=1}^N \sum_{i'=1}^N (\mathbf{x}_i - \mathbf{x}_{i'}) (\mathbf{x}_i - \mathbf{x}_{i'})' = 2N \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})', \quad (6.5)$$

donde  $\bar{\mathbf{x}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$  es el vector de medias.

Supongamos ahora  $g$  matrices de datos  $\mathbf{X}_1, \dots, \mathbf{X}_g$ , donde cada  $\mathbf{X}_i$  es de orden  $n_i \times p$ . Recordemos la identidad  $\mathbf{T} = \mathbf{B} + \mathbf{W}$ , donde

$$\begin{aligned} \mathbf{B} &= \sum_{k=1}^g n_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}) (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})', \\ \mathbf{W} &= \sum_{k=1}^g \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k) (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)', \\ \mathbf{T} &= \sum_{k=1}^g \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}) (\mathbf{x}_{ki} - \bar{\mathbf{x}})', \end{aligned}$$

siendo  $n = n_1 + \dots + n_g$ . Aquí  $\mathbf{x}'_{ki}$  es una fila  $\mathbf{X}_k$  con media  $\bar{\mathbf{x}}_k$  y  $\bar{\mathbf{x}}$  es la media general.  $\mathbf{T}, \mathbf{B}, \mathbf{W}$  son las matrices total, entre muestras y dentro de muestras, respectivamente, utilizadas en MANOVA.

De (6.5) obtenemos:

$$\begin{aligned} \mathbf{T} &= \sum_{k,h=1}^g \sum_{i,i'=1}^{n_k, n_h} (\mathbf{x}_{ki} - \mathbf{x}_{hi'}) (\mathbf{x}_{ki} - \mathbf{x}_{hi'})' \\ &= 2n \sum_{k=1}^g \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}) (\mathbf{x}_{ki} - \bar{\mathbf{x}})', \\ \mathbf{B} &= \sum_{k,h=1}^g n_k n_h (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_h) (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_h)' \\ &= 2n \sum_{k=1}^g n_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}) (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})', \\ \mathbf{W}_k &= \sum_{i,i'=1}^{n_k} (\mathbf{x}_{ki} - \mathbf{x}_{ki'}) (\mathbf{x}_{ki} - \mathbf{x}_{ki'})' \\ &= 2n_k \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k) (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)', \end{aligned}$$

que prueba la siguiente identidad con matrices construidas con diferencias entre observaciones y entre medias:

$$\mathbf{T} = \mathbf{B} + n \sum_{k=1}^g n_k^{-1} \mathbf{W}_k, \quad (6.6)$$

donde  $\mathbf{T}, \mathbf{B}$  y  $\mathbf{W}_k$  son matrices  $p \times p$ .

Podemos descomponer la variabilidad de forma similar. Recordemos que la *variabilidad geométrica* de una matriz  $n \times n$  de distancias  $\Delta = (\delta_{ii'})$ , con matriz de productos internos  $\mathbf{G}$ , se define como

$$V_\delta = \frac{1}{2n^2} \sum_{i=1}^n \sum_{i'=1}^n \delta_{ii'}^2 = \text{tr}(\mathbf{G})/n,$$

siendo  $\delta_{ii'}^2 = (\mathbf{x}_i - \mathbf{x}_{i'})' (\mathbf{x}_i - \mathbf{x}_{i'})$ .  $V_\delta$  es la versión muestral de (4.7).

Tomando ahora trazas en (6.6) obtenemos

$$\text{tr}(\mathbb{T}) = \text{tr}(\mathbb{B}) + n \sum_{k=1}^g n_k^{-1} \text{tr}(\mathbb{W}_k).$$

Expresamos esta identidad como

$$V_\delta(\text{total}) = V_\delta(\text{entre}) + n^{-1} \sum_{k=1}^g n_k V_\delta(\text{dentro } k). \quad (6.7)$$

**Ejercicio 6.3** Probar las identidades (6.6) y (6.7).

### 6.4.2 Tests con coordenadas principales

Las identidades (6.6) y (6.7) sirven de base para construir un test que compare varias poblaciones. Dados  $g$  conjuntos de datos independientes  $\mathcal{D}_1, \dots, \mathcal{D}_g$ , podemos obtener la supermatriz de distancias  $\Delta$  y las coordenadas principales  $\mathbf{X}_1, \dots, \mathbf{X}_g$ . Seguidamente obtenemos  $\mathbb{B}$  y  $\mathbb{T}$  y calculamos dos estadísticos para contrastar  $H_0$ :

- a)  $\gamma_1 = \det(\mathbb{T} - \mathbb{B}) / \det(\mathbb{T})$ ,
- b)  $\gamma_2 = V_\delta(\text{entre}) / V_\delta(\text{total})$ .

Ambos estadísticos toman valores entre 0 y 1. Valores pequeños de  $\gamma_1$  y valores grandes de  $\gamma_2$ , respectivamente, proporcionan evidencias en favor de la hipótesis alternativa. Obsérvese que  $\gamma_2 = \text{tr}(\mathbb{B}) / \text{tr}(\mathbb{T})$  es un estadístico basado en la llamada entropía cuadrática, siendo utilizado en ANOQE (analysis of quadratic entropy), una generalización de ANOVA (analysis of variance). ANOQE fue propuesto por C. R. Rao en varios artículos. Obsérvese también que la distribución de  $\gamma_1$  es Wilks si las poblaciones son normales multivariantes con la misma matriz de covarianzas y tomando la distancia Euclídea.

Excepto para datos multinormales y algunas otras distribuciones, la distribución muestral de  $\gamma_1$  y  $\gamma_2$  es desconocida. La distribución asintótica implica secuencias de parámetros auxiliares, que han sido hallados para distribuciones muy específicas (véase Cuadras y Fortiana, 1995; Cuadras y Lahlou, 2000; Cuadras *et al.*, 2006). Liu y Rao (1997) derivan la distribución bootstrap de  $V_\delta(\text{entre})$ . Otra opción consiste en el uso de métodos de remuestreo, como el test de permutaciones, según se ha visto en la Sección 6.2.

### 6.4.3 Tests con funciones de proximidad

Otro test, que evita el uso del remuestreo, se obtiene mediante funciones de proximidad y estadística noparamétrica. Fijémonos primero que, con datos cuantitativos y la distancia de Mahalanobis, las funciones de proximidad son

$$\phi_\delta^2(\mathbf{x}, \Pi_k) = (\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k), \quad k = 1, \dots, g.$$

Tales funciones son la misma bajo  $H_0 : \boldsymbol{\mu}_1 = \dots = \boldsymbol{\mu}_g$ .

Supongamos en general que  $\mathbf{x}_{11}, \dots, \mathbf{x}_{1n_1}$  representan las  $n_1$  observaciones de  $\Pi_1$  y que  $\omega$  es un nuevo individuo con coordenadas  $\mathbf{x}$ . La versión muestral de la función de proximidad (4.8) es

$$\hat{\phi}_1^2(\mathbf{x}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta^2(\mathbf{x}, \mathbf{x}_{1i}) - V_\delta(\text{dentro } 1),$$

deonde  $\delta(\omega, \omega_{1i}) = \delta(\mathbf{x}, \mathbf{x}_{1i})$ . Nótese que no necesitamos encontrar los vectores  $\mathbf{x}$  con las coordenadas.

Similarmente se obtienen  $\hat{\phi}_2^2(\mathbf{x})$ , etc. No obstante, bajo  $H_0$  todas las funciones de proximidad poblacionales son la misma función. Así, podemos trabajar con la función de proximidad global

$$\hat{\phi}^2(\mathbf{x}) = \frac{1}{n} \sum_{g=1}^g \sum_{i=1}^{n_g} \delta^2(\mathbf{x}, \mathbf{x}_{gi}) - V_\delta(\text{total}).$$

If  $a_{gi} = \hat{\phi}^2(\mathbf{x}_{gi}) = \|\mathbf{x}_{gi} - \bar{\mathbf{x}}\|^2$ , donde  $\mathbf{x}_{gi}$  proviene de  $\Pi_g$ , obtenemos los valores de proximidad (calculados *utilizando sólo* distancias) para cada población:

$$\Pi_1 : a_{11}, \dots, a_{1n_1}; \dots; \Pi_g : a_{g1}, \dots, a_{gn_g}.$$

Bajo  $H_0$  los valores  $a_{gi}$  siguen (aproximadamente) la misma distribución. Podemos entonces aplicar el test no paramétrico de Kruskal-Wallis para aceptar o rechazar  $H_0$ . Este test es basado en el estadístico

$$H = \left( \frac{12}{n(n+1)} \right) \sum_{k=1}^g \frac{R_k^2}{n_k} - 3(n+1),$$

donde los  $n$  valores  $a_{gi}$  se ordenan y  $R_g$  es la suma de los rangos correspondientes a los valores en  $\Pi_g$ . El estadístico  $H$  es asintóticamente ji-cuadrado con  $g-1$  g.l. Véase Cuadras (2000).

**Ejercicio 6.4** Con los datos *Iris* de R. A. Fisher, donde  $g = 3$  y  $p = 4$ , comparar las tres poblaciones utilizando:

1. La distancia euclídea.
2. La raíz cuadrada de la distancia “valor absoluto”, véase (1.8).

## Capítulo 7

# Distintividad

Se entiende por distintividad, el problema discriminante que consiste en averiguar si un individuo pertenece a una de entre dos poblaciones conocidas (o a una combinación lineal convexa de ambas), o por el contrario pertenece a una tercera población, posiblemente desconocida. Por ejemplo, en un experimento agrícola en el que se obtiene una nueva variedad de planta, interesa saber si pertenece a una especie conocida o en cambio debe considerarse una nueva especie.

### 7.1 Planteamiento bajo normalidad

Supongamos que nuestros datos pueden provenir de tres poblaciones  $\Omega_i, i = 1, 2, 3$ . En relación a un vector aleatorio  $\mathbf{X}$ , cada población es normal:  $\Omega_i \sim N_p(\mu_i, \Sigma)$ . Se necesita decidir si  $\mathbf{X}$  está relacionada con  $\Omega_1, \Omega_2$  (dos poblaciones conocidas) o con  $\Omega_3$  (una nueva población desconocida).

Suponiendo que  $\mathbf{x}$  es una observación de  $\mathbf{X}$ . Se desean contrastar las hipótesis:

$$\begin{aligned} H_0 &: \mathbf{x} \text{ proviene de } N_p(\alpha\mu_1 + \beta\mu_2, \Sigma), \quad \alpha + \beta = 1 \\ H_1 &: \mathbf{x} \text{ proviene de } N_p(\mu_3, \Sigma). \end{aligned} \quad (7.1)$$

Rechazar  $H_0$  significa que  $\mathbf{x}$  proviene de una nueva población  $\Omega_3$  no relacionada con  $\Omega_1, \Omega_2$ , las poblaciones de las que poseemos información.

Bajo  $H_0(\alpha = 1), H'_0(\beta = 1), H''_0(\alpha + \beta = 1)$ , se cumple que:

$$U_1(\mathbf{x}) = \frac{((\mathbf{x} - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1))^2}{(\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1)} \sim \chi_1^2 \quad (7.2)$$

$$U_2(\mathbf{x}) = \frac{((\mathbf{x} - \mu_2)' \Sigma^{-1} (\mu_2 - \mu_1))^2}{(\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1)} \sim \chi_1^2 \quad (7.3)$$

$$W_1(\mathbf{x}) = (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1) - U_1(\mathbf{x}) \sim \chi_{p-1}^2 \quad (7.4)$$

Si  $W_1(\mathbf{x})$  es significativa, entonces  $\mathbf{x}$  puede provenir de la población diferente  $\Omega_3$ . En otro caso,  $\mathbf{x}$  proviene de la combinación lineal convexa

$$\alpha\Omega_1 + (1 - \alpha)\Omega_2 \quad 0 \leq \alpha \leq 1$$

Véase Rao (1973, pp. 577-579), Bar-Hend y Daudin (1997).

## 7.2 Planteamiento DB

En primer lugar, obsérvese que:

$$\begin{aligned} 2(\mu_2 - \mu_1)' \Sigma^{-1}(\mathbf{x} - \mu_1) = \\ (\mu_2 - \mu_1)' \Sigma^{-1}(\mu_2 - \mu_1) + (\mathbf{x} - \mu_1)' \Sigma^{-1}(\mathbf{x} - \mu_1) - (\mathbf{x} - \mu_2)' \Sigma^{-1}(\mathbf{x} - \mu_2) \geq 0 \end{aligned}$$

Definiendo

$$P_1(\mathbf{x}) = (\mu_2 - \mu_1)' \Sigma^{-1}(\mathbf{x} - \mu_1),$$

se tiene entonces

$$P_1(\mathbf{x}) = \frac{1}{2} [\delta_M^2(\mu_1, \mu_2) + \delta_M^2(\mathbf{x}, \mu_1) - \delta_M^2(\mathbf{x}, \mu_2)] \geq 0 \quad (7.5)$$

donde  $\delta_M^2$  es la distancia de Mahalanobis. La versión DB para esta distancia entre  $\Omega_1$  y  $\Omega_2$  es

$$\Delta^2(\Omega_1, \Omega_2) = \int \delta^2(\mathbf{x}, \mathbf{y}) f_1(\mathbf{x}) f_2(\mathbf{y}) d\mathbf{x} d\mathbf{y} - V_1(\mathbf{X}) - V_2(\mathbf{Y})$$

La versión DB para  $P_1(\mathbf{x})$  es entonces

$$P_1(\mathbf{x}) = \frac{1}{2} [\Delta^2(\Omega_1, \Omega_2) + \phi_1^2(\mathbf{x}) - \phi_2^2(\mathbf{x})] \quad (7.6)$$

donde  $\phi_1^2, \phi_2^2$  son las funciones de proximidad. De forma similar:

$$P_2(\mathbf{x}) = \frac{1}{2} [\Delta^2(\Omega_1, \Omega_2) + \phi_2^2(\mathbf{x}) - \phi_1^2(\mathbf{x})] \quad (7.7)$$

Entonces las versiones DB de (7.2-7.4) para probar

$$\begin{aligned} H_0 &: \mathbf{x} \text{ proviene de } \Omega_1 \quad (\alpha = 1), \\ H'_0 &: \mathbf{x} \text{ proviene de } \Omega_2 \quad (\alpha = 0), \\ H''_0 &: \mathbf{x} \text{ proviene de } \alpha\Omega_1 + (1 - \alpha)\Omega_2, \quad (0 \leq \alpha \leq 1), \end{aligned}$$

respectivamente, están basadas en los estadísticos

$$U_1(\mathbf{x}) = \frac{[P_1(\mathbf{x})]^2}{\Delta^2(\Omega_1, \Omega_2)} \quad (7.8)$$

$$U_2(\mathbf{x}) = \frac{[P_2(\mathbf{x})]^2}{\Delta^2(\Omega_1, \Omega_2)} \quad (7.9)$$

$$W_1(\mathbf{x}) = \phi_1^2(\mathbf{x}) - U_1(\mathbf{x}) \quad (7.10)$$

Las distribuciones de  $U_1(\mathbf{x}), U_2(\mathbf{x}), W_1(\mathbf{x})$  bajo la hipótesis nula, pueden ser obtenidas por remuestreo (véase Sección 6.2). Si  $W_1(\mathbf{x})$  es significativa, entonces  $\mathbf{x}$  puede provenir de una población *diferente*  $\Omega_3$ . Este planteamiento DB ha sido estudiado por Cuadras y Fortiana (2000). Véase también Rao (1973, p.579).

**Ejercicio 7.1** *Se pide:*

1. A partir de los datos mixtos sobre cáncer del Ejercicio 3.2, representa mediante un histograma de  $U_1(\mathbf{x}), U_2(\mathbf{x}), W_1(\mathbf{x})$  empleando solamente la muestra de los  $n_1 = 78$  (benignos).
2. Demuestra (7.2-7.4).
3. Define  $W_2(\mathbf{x}) = \phi_2^2(\mathbf{x}) - U_2(\mathbf{x})$ . Demuestra que  $W_1(\mathbf{x}) = W_2(\mathbf{x})$ .

### 7.3 Planteamiento mediante la razón de proximidades

Sea la función de proximidad  $\phi_i^2(\mathbf{x}), i = 1, 2$ , entre una observación  $\mathbf{x}$  y  $\Omega_i, i = 1, 2$ . Si introducimos la razón

$$r_i(\mathbf{x}) = \frac{\phi_i^2(\mathbf{x})}{\phi_1^2(\mathbf{x}) + \phi_2^2(\mathbf{x})} \quad i = 1, 2, \quad (7.11)$$

entonces si  $r_i(\mathbf{x}) \cong 0, \mathbf{x}$  provendrá de  $\Omega_i, i = 1, 2$ . Además, si  $\phi_1^2(\mathbf{x}), \phi_2^2(\mathbf{x})$  siguen (asintóticamente) la misma distribución, bajo  $H_0 : \Omega_1 = \Omega_2$  se cumple que

$$E(r_i(\mathbf{x})) = 1/2.$$

Considerando la hipótesis nula

$$H_0^1 : \mathbf{x} \text{ proviene de } \Omega_1$$

y suponiendo que  $r_i(\mathbf{x})$  sigue una distribución uniforme (0,1) (un argumento similar vale para otra distribución como la distribución beta). Entonces:

$$P(r_1(\mathbf{x}) \geq 1 - \alpha \mid H_0^1) = \alpha$$

y se puede aceptar  $H_0^1$  si

$$r_1(\mathbf{x}) \leq 1 - \alpha$$

para un  $\alpha$  nivel de significación dado.

Similarmente podemos considerar

$$H_0^2 : \mathbf{x} \text{ proviene de } \Omega_2$$

y aceptar esa hipótesis si

$$r_2(\mathbf{x}) \leq 1 - \alpha$$

Si ambas hipótesis  $H_0^1, H_0^2$  son rechazadas, entonces se puede decidir que  $\mathbf{x}$  proviene de una población *diferente*  $\Omega_3$ .

**Ejercicio 7.2** *Se pide:*

1. *A partir de los datos mixtos sobre cáncer del Ejercicio 4.10, representa mediante un histograma  $r_1(\mathbf{x}), r_2(\mathbf{x})$ , empleando ambas muestras independientemente.*
2. *Demuestra que, bajo la hipótesis nula  $H_0 : \Omega_1 = \Omega_2$  y empleando las funciones de proximidad estimadas, se tiene asintóticamente que,  $E(r_1(\mathbf{x})) = E(r_2(\mathbf{x})) = 1/2$ .*
3. *Propón algún criterio de prueba alternativa.*



# Bibliografia

- [1] Arenas, C., Cuadras, C. M. (2002) Recent statistical methods based on distances. *Contributions to Science*, **2**, 183–191.
- [2] Arenas, C. and Cuadras, C. M. (2004) Comparing two methods for joint representation of multivariate data. *Comm. Stat. Simul. Comp.* **33**, 415–430.
- [3] Arenas, C. and Turbón, D. (1999) The usefulness of discrimination based on distances on human evolution. *Qüestió*, **22**(3), 529–538.
- [4] Bar-Hend, A. and Daudin, J. J. (1997) A test of a special case of typicality. in linear discriminant analysis. *Biometrics*, **53**, 39–48.
- [5] Cox., T.F., Cox, M.A.(1994) *Multidimensional Scaling*. Chapman & Hall, London.
- [6] Cuadras, C.M. (1989) Distance analysis in discrimination and classification using both continuous and categorical variables. In: Y. Dodge (Ed.), *Statistical Data Analysis and Inference*, pp. 459–473. Elsevier Science Publishers B. V. (North–Holland), Amsterdam.
- [7] Cuadras, C. M. (1990) An eigenvector pattern arising in nonlinear regression. *Qüestió*, **14**, 89–95.
- [8] Cuadras, C. M. and C. Arenas (1990) A distance based regression model for prediction with mixed data. *Communications in Statistics A. Theory and Methods* **19**, 2261–2279.
- [9] Cuadras, C. M. (1991) *Métodos de Análisis Multivariante*. 2a edic., PPU, Barcelona.
- [10] Cuadras, C. M. (1992) Some examples of distance based discrimination. *Biometrical Letters*, **29**, 1–18.
- [11] Cuadras, C. M (1992) Probability distributions with given multivariate marginals and given dependence structure. *J. of Multivariate Analysis*, **42**, 51–66.

- [12] Cuadras, C. M. (1998) Multidimensional dependencies in ordination and classification. In: K. Fernández and E. Morinneau (Eds.), *Analyses Multidimensionnelles des Données*, pp.15-26, CISIA-Ceresta, Saint Mandé (France).
- [13] Cuadras, C.M. (2000). *Problemas de Probabilidades y Estadística. Vol. II. Inferencia Estadística*. EUB, Barcelona.
- [14] Cuadras, C.M. y Fortiana, J. (1993) Aplicaciones de las distancias en estadística. *Qüestió*, 17, 39-74.
- [15] Cuadras, C.M. y Fortiana, J. (1993) Continuous metric scaling and prediction. En: *Multivariate Analysis: Future Directions 2*. (C.M. Cuadras and C.R. Rao, eds.). Elsevier, Amsterdam, pp. 47-66.
- [16] Cuadras, C.M. (1993) Interpreting an Inequality in Multiple Regression. *Amer. Statistician*, 47, 256-258.
- [17] Cuadras, C.M. y Fortiana, J. (1994) Ascertaining the underlying distribution of data set. En: *Selected Topics on Stochastic Modelling*, (R. Gutierrez and M.J. Valderrama, eds.). World Scientific, Singapore, pp. 223-230.
- [18] Cuadras, C.M. y Fortiana, J. (1995) A continuous metric scaling solution for a random variable. *J. of Multivariate Analysis*, 52(1), 1-14.
- [19] Cuadras, C.M. y Fortiana, J. (1995) Representation of statistical structures, classification and prediction using multidimensional scaling. En: *Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organization* (W. Gaul and D. Pfeifer, eds.), Springer Verlag, Berlin, pp. 20-31.
- [20] Cuadras, C.M. (1996) Discussion of "The relation between theory and applications in statistics" by D.R. Cox. *TEST*, 4, 253-261.
- [21] Cuadras, C. M., C. Arenas and J. Fortiana (1996) Some computational aspects of a distance-based model for prediction. *Communications in Statistics. Simulation and Computation*, 25(3), 593-609.
- [22] Cuadras, C. M., Atkinson, R.A. Fortiana, J. (1997) Probability densities from distances and discriminant analysis. *Statistics and Probability Letters*, 33, 405-411
- [23] Cuadras, C. M., Cuadras, D. and Lahlou, Y. (2006) Principal directions of the general Pareto distribution with applications. *J. of Statistical Planning and Inference*, 136, 2572-2583.
- [24] Cuadras, C. M. and Lahlou, Y. (2000) Some orthogonal expansions for the logistic distribution. *Communications in Statistics, Theory and Methods*, 29, 2643-2663.

- [25] Cuadras, C. M. and J. Fortiana (1996) Weighted continuous metric scaling. In: Gupta, A. K. and V. L. Girko (Eds.), *Multidimensional Statistical Analysis and Theory of Random Matrices*, pp. 27–40. VSP, Zeist, The Netherlands.
- [26] Cuadras, C. M., Fortiana, J. and F. Oliva (1996) Representation of statistical structures, classification and prediction using multidimensional scaling. In: W. Gaul and D. Pfeifer, *From Data to Knowledge*, pp. 20–31. Springer-Verlag, Berlin.
- [27] Cuadras, C. M., Fortiana, J. and F. Oliva (1997) The proximity of an individual to a population with applications in discriminant analysis. *J. of Classification*, 14, 117–136.
- [28] Cuadras, C. M., Fortiana, J. (1997) Continuous scaling on a bivariate copula. In: Viktor Benes and Josef Stepan, (Eds). *Distributions with éven marginals and moment problems*, pp 137–142. Kluwer Academic Pub., Dordrecht.
- [29] Cuadras, C. M., Fortiana, J. (1997) Visualizing categorical data with related metric scaling. In: J. Blasius and M. Greenacre, (Eds.) *Visualization of Categorical Data*, pp. 365–376, Academic Press, N. York.
- [30] Cuadras, C.M., Fortiana, J. (1998) Typicality in discriminant analysis with mixed variables. *Working paper*.
- [31] Cuadras, C. M., Fortiana, J. (2000) The Importance of Geometry in Multivariate Analysis and some Applications. In: *Statistics for the 21st Century*, C.R. Rao and G. Szekely, eds., Marcel Dekker, New York.
- [32] Cuadras, C. M., Fortiana, J. and M. J. Greenacre (2000) Continuous extensions of matrix formulations in correspondence analysis, with applications to the FGM family of distributions. In: R.D.H. Heijmans, D.S.G. Pollock and A. Satorra, (Eds.), *Innovations in Multivariate Statistical Analysis*, pp. 101–116. Kluwer Ac. Publ., Dordrecht.
- [33] Cuadras, C. M. and Fortiana, J. (2004) Distance-based multivariate two sample tests, In *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis and Quality of Life*, (Eds. M. S. Nikulin, N. Balakrishnan, M. Mesbah, N. Limnios), pp. 273–290. Birkhauser, Boston.
- [34] Escoufier, Y. (1973) Le traitement des variables vectorielles. *Biometrics*, 29, 751–760.
- [35] Gabriel, K.R. (1971) The biplot graphic display of matrices with applications to principal component analysis. *Biometrika*, 58, 453–467.
- [36] Gower, J. C. (1968) Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55, 582–585.

- [37] Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857-874.
- [38] Gower, J. C. and Harding, S.A. (1988) Nonlinear biplots. *Biometrika*, 75, 445-455.
- [39] Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression. *Technometrics*, 12, 55-67 and 69-82.
- [40] Krzanowski, W.J. (1980) Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 36, 493-499.
- [41] Krzanowski, W.J. (1994) Ordination in the presence of group structure, for general multivariate data. *Journal of Classification*, 11, 195-207.
- [42] Liu, Z. J. and Rao, C. R. (1995) Asymptotic distribution of statistics based on quadratic entropy and bootstrapping. *Journal of Statistical Planning and Inference*, 43, 1-18.
- [43] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press, London.
- [44] Miñarro, A. and Oller, J.M. (1992) Some remarks on the individuals-score distance and its applications to statistical inference. *Qüestió*, 16, 43-57.
- [45] Oller, J.M. (1989) Some geometrical aspects of data analysis and statistics. En: *Statistical Data Analysis and Inference*. (Y. Dodge, eds.) Elsevier, Amsterdam, pp. 41-58.
- [46] Rao, C.R. (1973) *Linear Statistical Inference and Its Applications*. New York: Wiley.