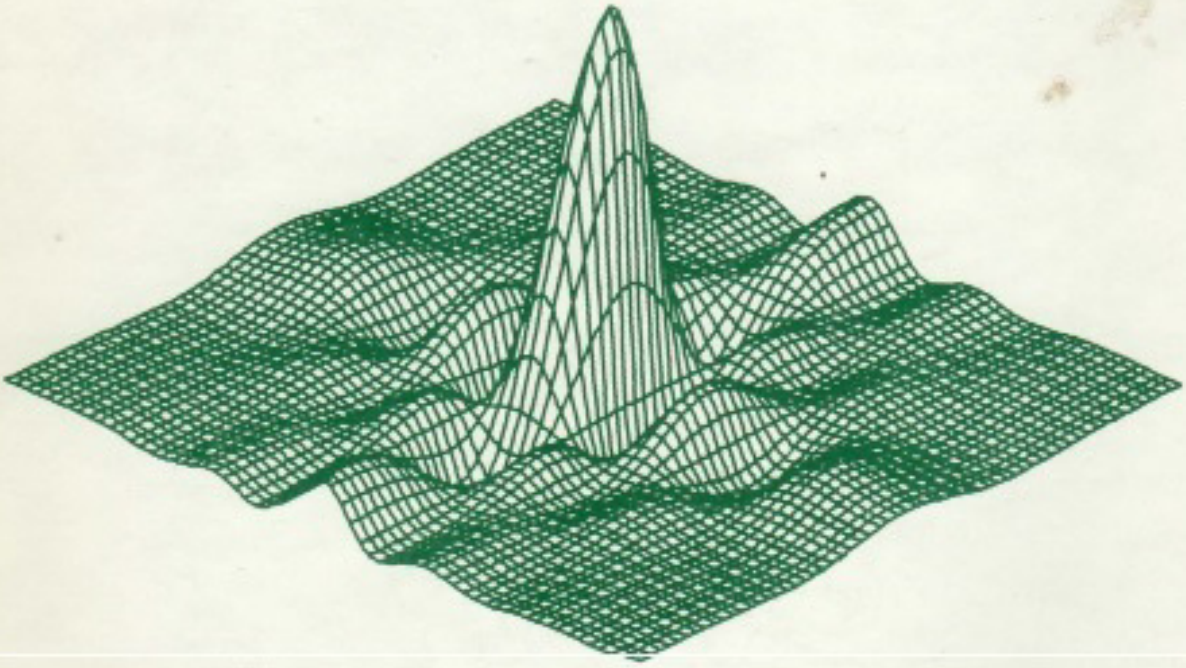
Publicacions del Departament d'Estadística no. 6

PREDICCIÓN MULTIVARIANTE BASADA EN DISTANCIAS

Carles M. Cuadras Sonia Salvo-Garrido

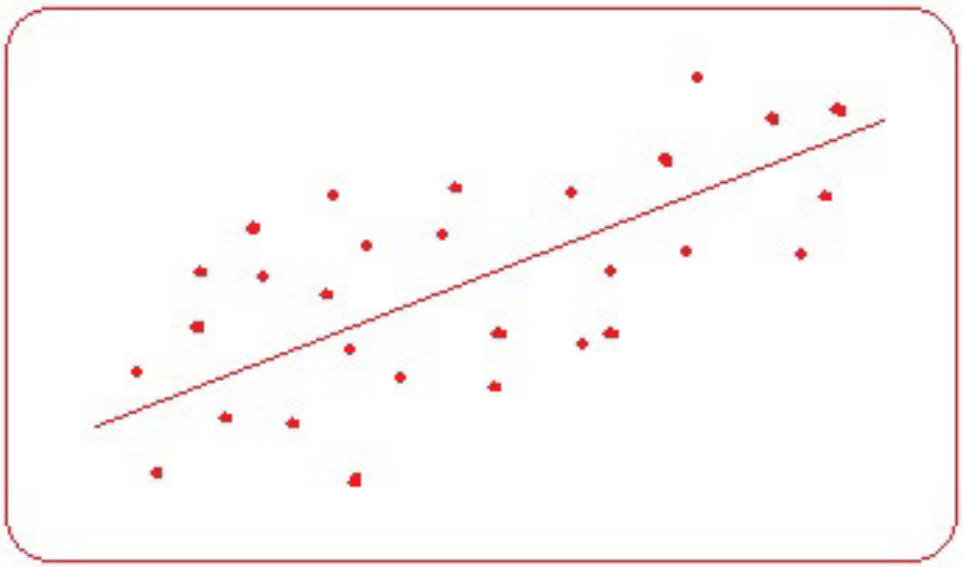


Publicacions del
DEPARTAMENT D'ESTADISTICA

UNIVERSITAT DE
BARCELONA

Predicción Multivariante Basada en Distancias

Curso monográfico de doctorado



PREDICCIÓN MULTIVARIANTE BASADA EN DISTANCIAS

Carles .M. Cuadras, Sonia Salvo

Octubre 2017

Índice

Prólogo

Este curso monográfico de doctorado trata de diversos métodos de Analisis Multivariante y Regresion, basados en el concepto y las propiedades de las distancias estadísticas. Una versión previa fue impartida en la Universidad Carlos III de Madrid durante varios cirsos académicos, en la década de 1990.

La monografía contiene un amplio resumen de la teoría y una colección de ejercicios, combinando los de carácter teórico con los aplicados, basados en datos reales.

Se constata a lo largo del curso, que la predicción de una variable sobre nuevos individuos, puede mejorarse, en general, utilizando dimensiones predictoras y funciones discriminantes, derivadas del análisis de ciertas matrices de distancias, que dando la mayoría de métodos clásicos como casos particualres. La distancia elegida juega el papel de modelo de predicción.

Los cálculos requeridos en los ejercicios, pueden realizarse mediante el programa MULTICUA (versión DOS). Contáctese con el primer autor para obtener este programa. Los datos utilizados en los ejercicios se pueden bajar del enlace:

http://www.ub.edu/stat/personal/cuadras/ejercicios.txt

Dos versiones recientes de programas que realizan predicción basada en distancias son:

1) Regresión lineal y generalitzada:
Package **dbstats (**Distance-based statistics, 2017).
Autor de contacto: Eva Boj [evaboj@ub.edu]

2) Análisis discriminante
WeDiBaDis (Weighted distance based discriminant analysis, 2016).
Autor de contacto: Concepción Arenas [carenas@ub.edu]

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, clustering), 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, visualización de curvas hiperespectrales, y especialmente en regresión y análisis discriminante. En este curso aprenderemos, mediante resúmenes de teoría y una selección de ejercicios, cómo utilizar la metodología basada en distancias para abordar estas partes de la estadística. Véase Cuadras (1989a).

1.2 Distancias

Una distancia δ sobre un conjunto (finito o no) Ω 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. $\delta_{ij} = \delta_{ji}$
Si además, se cumple la desigualdad triangular
P.4. $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$
diremos que la distancia es **métrica**.
Si además, se cumple que existen puntos $P_1, P_2, \ldots, P_n$ de un cierto espacio Euclídeo R^p, tales que
P5. $\delta_{ij} = d_E(P_i, P_j)$
siendo d_E la distancia ordinaria (véase (1.6)), diremos que la distancia δ_{ij} es **Euclídea**.

Si Ω es un conjunto finito $\{\omega_1, \omega_2, ..., \omega_n\}$, que para abreviar indicaremos como $\Omega = \{1, 2, ..., n\}$, las distancias δ_{ij} se expresan mediante la matriz simétrica $\mathbf{\Delta}$, llamada **matriz de distancias** sobre Ω:

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

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

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

es decir, la ordenación de los pares (i, j) de Ω, de acuerdo con su proximidad. Un buen método de representación de datos conserva, lo mejor posible, la preordenación de Ω.

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 al conjumto {H, Ch, Go, O, Gi}.

Una matriz de distancias $\mathbf{\Delta} = (\delta_{ij})$ puede ser transformada de diversos modos. Por ejemplo:

$$\delta^*_{ij} = \begin{cases} 0 & i = j, \\ \delta_{ij} + c & i \neq j. \end{cases} \tag{1.1}$$

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

$$\tilde{\delta}^2_{ij} = \begin{cases} 0 & i = j, \\ \delta^2_{ij} + c & i \neq j. \end{cases} \tag{1.2}$$

que afecta el cuadrado de la distancia y se llama **q-aditiva**. Las transformaciones (1.1) y (1.2) son útiles para conseguir que la distancia transformada cumpla ciertas propiedades (ser métrica o Euclídea), que la distancia original no posee, pero conservando la preordenación, es decir, las relaciones de proximidad entre los individuoes de Ω.

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

1. *Prueba que las transformaciones (1.1) y (1.2) conservan la preordenación de Ω.*

2. *Suponiendo que $\mathbf{\Delta}$ no es métrica, prueba que la transformación aditiva tomando*

$$c = \max\{\delta_{ij} - \delta_{ik} - \delta_{jk}\}$$

para cada $i, j, k \in \Omega$, transforma $\mathbf{\Delta}$ en $\mathbf{\Delta}^* = \left(\delta^*_{ij}\right)$, matriz de distancias que sí tiene la propiedad métrica.

1.3 Similaridades en general

En muchas aplicaciones es conveniente trabajar con similaridades, concepto dual al de distancias. Una **similaridad** s sobre un conjunto Ω, 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:

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

S.2. $s_{ij} = s_{ji}.$

Cuando Ω es un conjunto finito, entonces la matriz

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

se denomina **matriz de similaridades** sobre Ω.

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

$$\delta_{ij} = 1 - s_{ij}, \tag{1.3}$$

así como

$$\delta_{ij} = \sqrt{1 - s_{ij}}. \tag{1.4}$$

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:

$$\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}. \tag{1.5}$$

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

1.4 Distancias para variables cuantitativas

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

a) la distancia Euclídea,

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

b) la distancia "ciudad"

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

c) la distancia "valor absoluto"

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

Cuando Ω es una población normal multivariante $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, con $\boldsymbol{\Sigma}$ no singular, expresando $\mathbf{x},\mathbf{y}$ como vectores columna, la distancia estadística (al cuadrado) más apropiada es

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

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

Ejercicio 1.3 *Comprueba que la distancia d_E (al cuadrado) 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 cual sería 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 (al cuadrado) de Mahalanobis de $\mathbf{x}$ a $\boldsymbol{\mu}_1$ y de $\mathbf{x}$ a $\boldsymbol{\mu}_2$.

1.5 Similaridades y distancias con variables binarias

Supongamos que tenemos p variables binarias $X_1, X_2, ...X_p$, donde cada X_i toma los valores 0 ó 1. Para cada par de individuos (i,j), son bien conocidos los coeficientes de similaridad:

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

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

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) y (1.5).

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*

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

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

1.6 Similaridad con variables mixtas

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

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

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 α es el número de coincidencias para las p_3 variables cualitativas. 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 $p_1 = p_3 = 0$. Este coeficiente fue propuesto por Gower (1971).

Otras versiones de distancias juntando dos conjuntos de variables mixtas han sido propuestas por Cuadras (1992b) y Cuadras y Fortiana (1997b).

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

	A	***B***	***C***	***D***	***E***	***F***
1	180	76	1	1	0	0
2	174	67	1	3	1	1
3	174	68	2	3	1	1
4	170	64	2	2	0	1
5	177	70	1	3	0	1

	A	***B***	***C***	***D***	***E***	***F***
6	181	72	3	2	1	0
7	171	60	3	2	0	0
8	162	58	1	3	1	1
9	170	66	2	2	0	1
10	179	81	1	1	1	0

A= talla, B= peso, 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 distancias

Si disponemos de una densidad de probabilidad $f(x_1, ..., x_p)$, podemos definir distancias entre poblaciones y entre observaciones siguiendo un camino menos heurístico.

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

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

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

$$\mathbf{F} = E\left(\mathbf{Z}\mathbf{Z}'\right).$$

Como $E(\mathbf{Z}) = \mathbf{0}$, vemos que $\mathbf{F}$ es la matriz de covarianzas del vector $\mathbf{Z}$.

Una definición de distancia (al cuadrado) entre dos observaciones $\mathbf{x} = (x_1, ..., x_p)$, $\mathbf{y} = (y_1, ..., y_p)$, que generaliza (1.9), 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 (1989b) y Oller, (1989), depende del parámetro $\boldsymbol{\theta}$. Véase también Miñarro y Oller (1992).

Ejercicio 1.8 *Sea X una variable aleatoria. Encuentra d_R en los casos:*

1. *X es exponencial de parámetro α.*
2. *X es Poisson de parámetro λ.*
3. *$\mathbf{X}$ es un vector aleatorio $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$, donde $\boldsymbol{\Sigma}_0$ es una matriz conocida.*

1.8 Teorema de caracterización

Sea $\boldsymbol{\Delta} = (\delta_{ij})$ una matriz $n \times n$ de distancias Euclídeas. Es decir, existen n vectores fila $\mathbf{x}_1, ..., \mathbf{x}_n \in R^m$, 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 o no? 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 $a_{ij} = -\delta_{ij}^2/2$ y calculemos $\mathbf{B} = \mathbf{HAH}$. Recordemos que una matriz simétrica es semidefinida positiva si todos sus valores propios son no negativos.

Teorema 1 *La matriz de distancias $\boldsymbol{\Delta}$ 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 rang$(\mathbf{B}) = m \leq n - 1$.*

Hagamos explícito este teorema. Si $\boldsymbol{\Delta}$ es Euclídea, es posible obtener la descomposición espectral

$$\mathbf{B} = \mathbf{U}\Lambda\mathbf{U}' = \mathbf{X}\mathbf{X}', \tag{1.13}$$

donde $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}$contiene los m vectores propios de $\mathbf{B}$, $\boldsymbol{\Lambda}$ es una matriz diagonal que contiene los valores propios ordenados $\lambda_1 > \cdots > \lambda_m > 0$. La matriz $\mathbf{B}$ proporciona las coordenadas Euclídeas del conjunto $\Omega =$

$\{1, 2, ..., n\}$. Cada fila $\mathbf{x}_i$ de $\mathbf{X}$ contiene las coordenadas, llamadas **coordenadas principales** del individuo i.

Las coordenadas principales tienen interesantes propiedades:

- Las filas $\mathbf{x}_1, ..., \mathbf{x}_n$ 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 elementos δ_{ij} de $\mathbf{\Delta}$.
- Las columnas $X_1, \ldots, X_m$ de $\mathbf{X}$, entendidas como variables, tienen media 0.
- Cada columna X_j de $\mathbf{X}$ tiene varianza igual a λ_j/n.
- Las columnas de $\mathbf{X}$ son ortogonales (variables incorrelacionadas).
- Las columnas de $\mathbf{X}$ pueden ser interpretadas como componentes principales.
- La representación de los individuos $1, 2, \ldots, n$ utilizando las filas de $\mathbf{X}$ es óptima.

La cantidad

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

es máxima en dimensión reducida k. En (1.14) $\delta_{ij}(k)$ es la distancia utilizando las $k < m$ primeras coordenadas y $\lambda_1 > \cdots > \lambda_k$ son 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* $\mathbf{S} \geq 0$*, es decir,* $\mathbf{S}$ *es (semi) definida positiva, entonces la distancia* δ_{ij} *tal que*

   $$\delta_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij} \tag{1.15}$$

   es Euclídea.
4. *Observa que, excepto por un factor constante, (1.4) es un caso particular de (1.5), y comprueba que (1.3) en cambio, 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 Ω son Euclídeas, entonces la distancia $\delta(.,.)$ tal que*

$$\delta^2(.,.) = \delta_1^2(.,.) + \delta_2^2(.,.) \tag{1.16}$$

   también es Euclídea.

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

Cuando la matriz de distancias $\mathbf{\Delta}$ no es Euclídea, debe transformarse para proporcionar una distancia Euclídea, pero conservando la preordenación de Ω. La transformación puede ser no lineal (obtenida numéricamente) o algebraica. El siguiente teorema brinda dos transformaciones algebraicas.

Teorema 2 *Sea $\mathbf{\Delta} = (\delta_{ij})$ una matriz de distancias no Euclídeas. Entonces $\mathbf{B}$ tendrá valores propios positivos y negativos: $\lambda_1 > ... > \lambda_k > 0 > \lambda'_1 > ... > \lambda'_{k'}$, con $k + k' = n - 1$. Se verifica:*

1. *La transformación q-aditiva con $c \geq -2\lambda'_k$, convierte $\mathbf{\Delta}$ en $\widetilde{\mathbf{\Delta}}$ Euclídea.*

2. *La transformación aditiva con $c \geq \lambda$, donde λ es el mayor valor propio de la matriz no simétrica*

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

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

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

1. *El primer apartado del Teorema 2.*

2. *El segundo apartado del Teorema 2.*

3. *Que las transformaciones aditiva y q-aditiva conservan la preordenación.*

Resolver el apartado 2 presenta mayor dificultad. Véase Cailliez (1983).

1.9 La fórmula de añadir un punto

Supongamos que $\mathbf{\Delta} = (\delta_{ij})$ es una matriz de distancias Euclídeas. Indiquemos por $n+1$ un nuevo individuo. Supongamos conocidas las distancias entre $n+1$ y los individuos $1, 2, \ldots, n$:

$$\delta_j = \delta(j, n+1), \qquad j \in \Omega. \tag{1.17}$$

Si cada 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 $n+1$ viene representado por $\mathbf{x} \in R^m$, las coordenadas de $\mathbf{x}$ (expresadas como un vector columna) se dan a continuación.

Teorema 3 *Las coordenadas $\mathbf{x} \in R^m$ de $n+1$ en función de $\mathbf{X}$ y de las distancias (1.17) es*

$$\mathbf{x} = \tfrac{1}{2}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{b} - \mathbf{d}), \tag{1.18}$$

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

$$\mathbf{x} = \tfrac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{X}'(\mathbf{b} - \mathbf{d}). \tag{1.19}$$

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 tiene la propiedad de ser Euclídea.*
2. *Representa las 6 bebidas.*
3. *Una nueva bebida refrescante LC mantiene las siguientes distancias con las demás*

	PC	*CC*	*CCC*	*DPC*	*D7UP*	*7-UP*
LC	220	265	244	210	99	85

Sitúa LC, utilizando (1.19), dentro del gráfico la representación anterior de las 6 bebidas ya conocidas.

TABLA 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 $\Omega_1, ..., \Omega_g$, representadas por los vectores de medias $\boldsymbol{\mu}_1, ...\boldsymbol{\mu}_g$ 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 de este método resultaría algo extensa (véanse los libros Cuadras, 1991, 2014), 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

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

La representación, a lo largo de los llamados ejes canónicos, incluye regiones confidenciales para cada uno de los vectores de medias (Figura 1.1).

Ejercicio 1.13 *Utilizando los datos del fichero* ***ejercicios.txt****, 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 las p columnas (= variables) de $\mathbf{Y}$. Usualmente los individuos se representan como puntos y las variables como vectores. Véase una alternativa interesante en Galindo (1988). La presentación conjunta **Biplot** puede ser obtenida por una descomposición singular de $\mathbf{Y}$, pero Gower y

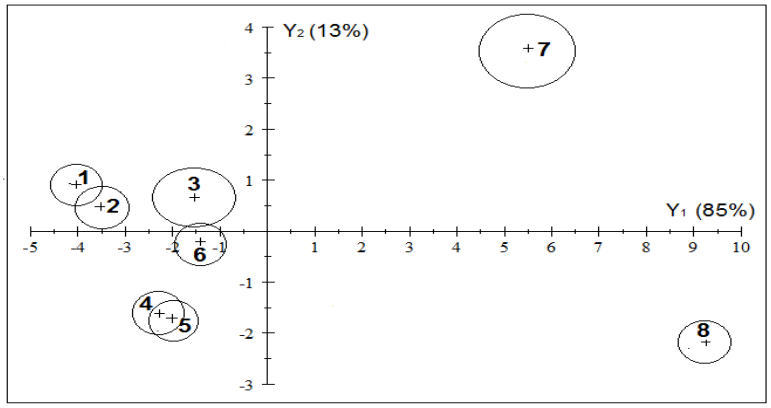


Figura 1.1: Representación canónica de 8 poblaciones conteniendo datos biométricos de 6 especies de coleópteros, encontrados en 8 localidades distintas.

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

$$\alpha_1 \mathbf{u}_1 = (\alpha_1, 0, ..., 0) \qquad \alpha_1 \in R_1,$$

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

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

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

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

Puesto que $\mathbf{X'1}_n = \mathbf{0}$, tenemos que las coordenadas representando al eje Y_1 son $\mathbf{x_1}(\alpha_1) = \alpha_1 \Lambda^{-1} \mathbf{X'Y}_1$, con $\alpha_1 \in R_1$, es decir, recorriendo el rango de Y_1.

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

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

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

	Kilómetro			
corredor	4	8	12	16
1	10	10	13	12
2	12	12	14	15
3	11	10	14	13
4	9	9	11	11
5	8	8	9	8
6	8	9	10	9
7	10	10	8	9
8	11	12	10	9
9	14	13	11	11
10	12	12	12	10
11	13	13	11	11
12	14	15	14	13

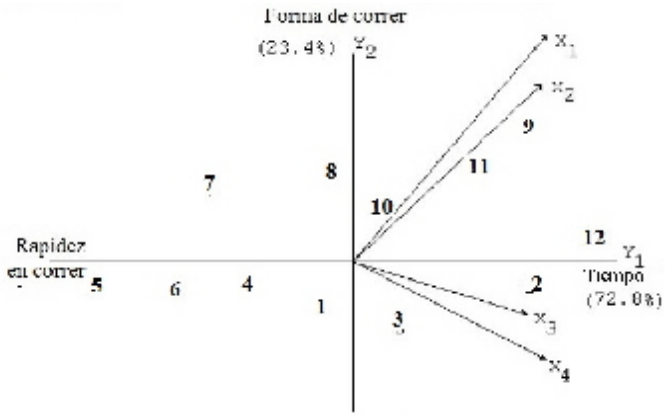


Figura 1.2: Gráfico biplot sobre los tiempos empleados por 12 corredores (puntos) en relación a 4 tiempos parciales (variables representadas mediante vectores). Las dos dimensiones principales representan la rapidez en correr y la forma de correr.

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

	An	*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	*4*	*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, representarla mediante el método biplot (ver un ejemplo de biplot en la Figura 1.2).

Capítulo 2

El modelo de regresión DB

2.1 El modelo clásico de regresión lineal

Consideremos el **modelo lineal**

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2.1}$$

donde $\mathbf{y}(n \times 1)$ es un vector (conocido) con n observaciones de una variable respuesta cuantitativa Y, $\mathbf{X}(n \times m)$ es la llamada matriz de diseño, es conocida y tiene rango rang$(\mathbf{X}) = m$, $\boldsymbol{\beta}(m \times 1)$ es un vector (desconcido) de parámetros y $\mathbf{e} = (e_1, ..., e_n)'$ es un vector aleatorio tal que

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

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

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

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}, \tag{2.2}$$

la suma de cuadrados residual es

$$R_0^2 = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \widehat{\mathbf{e}}'\widehat{\mathbf{e}}, \tag{2.3}$$

y una estimación insesgada de la varianza común σ^2 es

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

Ejercicio 2.1 *Consideremos el modelo lineal*

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

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

Escribe este modelo en la forma matricial (2.1), y estima los parámetros $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ *y* σ^2.

Queremos ahora interpretar el modelo lineal desde la perspectiva de las distancias. Indiquemos por $\mathbf{x}_1, ..., \mathbf{x}_n$ las filas de $\mathbf{X}$. La distancia Euclídea (al cuadrado) entre cada par de observaciones i, j es

$$d_E^2(i,j) = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'. \tag{2.5}$$

Sobre la matriz de distancias Euclídeas $\mathbf{D} = (d_E(i,j))$, podemos aplicar el Teorema 1 de la Sección 1.8. Construimos $\mathbf{B} = \mathbf{HAH}$ y, a partir de la descomposición espectral de $\mathbf{B}$

$$\mathbf{B} = \mathbf{U}\Lambda\mathbf{U}' = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}', \tag{2.6}$$

resulta que las distancias entre las filas $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n$ de $\widetilde{\mathbf{X}}$, reproducen exactamente las distancias (2.5). Podemos, entonces, escribir el modelo (2.1) como

$$\mathbf{y} = \widetilde{\mathbf{X}}\boldsymbol{\gamma} + \mathbf{e}. \tag{2.7}$$

Se ha llevado a cabo una transformación $\beta \to \gamma$ de los parámetros.

Lo que más nos interesa aquí es que (2.7) ha sido construido utilizando *solamente distancias* (en este caso la distancia Euclídea), y le llamaremos **modelo DB** ("distance-based").

El vector proyección de $\mathbf{y}$ sobre el subespacio generado por el modelo (2.1) es

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.8}$$

Análogamente podríamos encontrar $\widehat{\widetilde{\mathbf{y}}}$ para el modelo (2.7). Sin embargo ambos modelos son esencialmente el mismo, pues se verifica

$$\widehat{\mathbf{y}} = \widehat{\widetilde{\mathbf{y}}}. \tag{2.9}$$

Como veremos enseguida, el modelo DB presenta ventajas utilizando otras distancias, lo que permitirá manejar modelos no lineales o modelos con variables explicactivas mixtas.

Ejercicio 2.2 *Se pide:*

1. *Encuentra la matriz* $\mathbf{T}$ *de la transformación lineal* $\boldsymbol{\gamma} = \mathbf{T}\boldsymbol{\beta}$.
2. *Demuestra la igualdad (2.9).*

La distancia Euclídea (2.5) se ha calculado utilizando las filas de la matriz de diseño $\mathbf{X}$, que es conocida. Ahora introduciremos el modelo de regresión DB ("distance based") en versión general.

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

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2.10}$$

donde $\mathbf{1}$ es el vector de unos, $\mathbf{X}$, $\boldsymbol{\beta}$ y $\mathbf{e}$ tienen el mismo significado que en el modelo (2.1). Obsérvese que, como $\mathbf{B1} = \mathbf{0}$, el vector $\mathbf{1}$ así como las columnas $X_1, ..., X_m$ de $\mathbf{X}$, son vectores propios de $\mathbf{B}$.

Podemos también escribir

$$\mathbf{y} = \beta_0 \mathbf{1} + \sum_{i=1}^{m} \beta_i X_i + \mathbf{e}, \tag{2.11}$$

donde m=rang ($\mathbf{B}$) y $X_1, ..., X_m$ (vectores columna $n \times 1$), juegan el papel de **variables predictoras.**

Las propiedades básicas del modelo de regresión DB son las siguientes:

- Las estimaciones de los parámetros de regresión son

$$\widehat{\beta}_0 = \overline{y} = \mathbf{y}'\mathbf{1}/n, \qquad \widehat{\beta}_i = X_i'\mathbf{y}/\lambda_i, \tag{2.12}$$

  donde λ_i es valor propio de $\mathbf{B}$.

- El vector predicción o proyección ortogonal de $\mathbf{y}$ sobre el subespacio generado por el modelo, es

$$\widehat{\mathbf{y}} = \overline{y}\mathbf{1} + \mathbf{X}\boldsymbol{\Lambda}^{-1}\mathbf{X}'\mathbf{y}. \tag{2.13}$$

- El coeficiente de correlación simple $r_i = r(\mathbf{y}, \mathbf{X}_i)$ es

$$r_i^2 = \left(\mathbf{y}'\mathbf{X}_i\right)/nS_y^2\lambda_i \,, \tag{2.14}$$

  donde S_y^2 es la varianza muestral de $\mathbf{y}$.

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

$$R^2 = \mathbf{y}'\mathbf{X}\mathbf{\Lambda}^{-1}\mathbf{X}'\mathbf{y}/nS_y^2 = \sum_{i=1}^{m} r_i^2, \tag{2.15}$$

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

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

2.2 El modelo DB en dimensión k

El modelo DB (2.10), (2.11), es el **modelo completo**. Sin embargo, para ciertas distancias, m =rang($\mathbf{B}$) crece con n. Incluso puede darse el caso de que $m = n - 1$. Entonces, el número de variables $X_1, ...,X_m$ (las columnas de $\mathbf{X}$) puede resultar excesivo y en consecuencia encontrarnos con un coeficiente de determinación R^2 arbitrariamente próximo a 1. Para evitar este problema, es conveniente partir $\mathbf{X}$ en dos partes

$$\mathbf{X} = \left(\mathbf{X}_{(k)}, \mathbf{Z}\right), \tag{2.16}$$

donde $\mathbf{X}_{(k)} = (X_1, ...,X_k)$ contiene k columnas adecuadas de $\mathbf{X}$, en el sentido de que son convenientes para predecir la variable respuesta Y, y la matriz $\mathbf{Z}$ contiene las restantes columnas. Las columnas en $\mathbf{X}_{(k)}$ no son necesariamente las k primeras de $\mathbf{X}$. Es decir, (2.16) es convencional.

Tomando estas k columnas, definimos el **modelo DB en dimensión** k, que se puede expresar de dos maneras equivalentes:

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}_{(k)}\boldsymbol{\beta}_{(k)} + \mathbf{e}_k, \qquad \mathbf{y} = \beta_0\mathbf{1} + \sum_{i=1}^{k} X_i\beta_i + \mathbf{e}_k. \tag{2.17}$$

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

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

$$r\left(\mathbf{y},X_1\right) > r\left(\mathbf{y},X_2\right) > \cdots > r\left(\mathbf{y},X_k\right). \tag{2.18}$$

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

Ejercicio 2.4 *Una revista de automóbiles publicó el consumo de gasolina Y en función de 10 características, sobre una muestra de 32 autos. Si 2 variables son binarias (b), 3 cualitativas (q) y el resto contínuas (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:*

TABLA 2.1 (véase **ejercicios.txt**)

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 *Encuentra la versión de las fórmulas (2.12)-(2.15), para el modelo DB en dimensión reducida k.*

2.3 Predicción DB sobre un nuevo individuo

Supongamos que tomando medidas sobre las variables (mixtas) observables, hemos obtenido la observación $\mathbf{w}_{n+1} = (w_1, ..., w_p)$ en un nuevo individuo $n+1$. Si tenemos definida una distancia entre individuos, entonces podremos calcular las distancias (al cuadrado)

$$d_i^2 = d^2\,(i, n+1) \qquad i = 1, ..., n, \tag{2.19}$$

entre $n+1$ y los otros n individuos cuyas observaciones conocemos para la variable respuesta Y. Queremos encontrar $y_{n+1} = Y(n+1)$, la predicción de Y en $n+1$. No es difícil obtener esta predicción, ya que, empleando la fórmula de añadir un punto (1.19), se pueden encontrar sus coordenadas a partir de las distancias. Estas coordenadas, utilizando el modelo completo, son

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

La predicción según el modelo (2.10) es

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

Substituyendo, se obtiene

$$\widehat{y}_{n+1} = \overline{y} + \mathbf{x}'\mathbf{\Lambda}^{-1}\mathbf{X}'\mathbf{y}. \tag{2.20}$$

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

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

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

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

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

$$\widehat{y}_{n+1}(k) = \overline{y} + \mathbf{x}'_{(k)}\mathbf{\Lambda}_k^{-1}\mathbf{X}'_k\mathbf{y}. \tag{2.21}$$

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

$$\widehat{y}_{n+1} = \overline{y} + \tfrac{1}{2}\left(\mathbf{b} - \mathbf{d}\right)' \mathbf{B}^{-} \mathbf{y},$$

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

2.4 Predicción con variables contínuas, categóricas y mixtas

El modelo DB se reduce al modelo clásico de regresión cuando la distancia utilizada es la Euclídea ordinaria (1.6) y las variables son cuantitativas y contínuas. 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, como veremos en la sección siguiente).

- Para p variables cualitativas $W_1, W_2, ..., W_p$, donde W_r tiene q_r estados $(1 \leq r \leq p)$, podemos tomar como distancia

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

  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 coinciden. Naturalmente los resultados son diferentes si consideramos otras distancias.

- La situación cambia bastante si las variables son mixtas, es decir, si constituyen una mezcla de contínuas, binarias y categóricas. Entonces la distancia no es Euclídea en el sentido de antes. Una buena posibilidad 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 (1989b), Cuadras y Arenas (1990), y continuado por Cuadras (1990), Cuadras y Fortiana (1993a), Cuadras, Arenas y Fortiana (1996). Pueden verse aplicaciones a la astronomia en Bartkowiak y Jakimiec (1994) y generalizaciones del modelo DB en Boj *et al.* (2010), así como aplicaciones a la química molecular en Robert *et al.* (1999), y teoría del riesgo en Boj *et al.* (2015), 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. *Para los siguientes datos ($n = 28$), la primera columna contiene la variable respuesta Y y las cinco columnas siguientes las variables explicativas. Comprueba que el método DB con la distancia de Gower, proporciona mejores resultados que el método clásico de regresión múltiple. (Nota: utiliza el archivo* ***ejercicios.txt*** *y 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 El método DB y la regresión no lineal

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

$$Y = f\left(X_1, ..., X_p; \boldsymbol{\theta}\right) + e, \tag{2.23}$$

donde $\boldsymbol{\theta}$ es el vector de parámetros y f es una función no lineal (posiblemente desconocida).

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 de tipo no lineal. Cuadras y Fortiana (1993b) demuestran que, si $p = 1$, la regresión DB con la distancia

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

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

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

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

$$d\left(i, j\right) = \sqrt{|i - j|}. \tag{2.25}$$

Demuestra que es Euclídea (compárala con Ejercicio 1.10) y, para $n = 8$*, demuestra que los 4 ejes principales obtenidos por análisis de coordenadas principales, es decir, utilizando (1.13), dan lugar a dimensiones de tipo lineal, cuadrática, cúbica y cuártica.*

Ejercicio 2.9 *En una determinada 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*

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

donde θ_1, θ_2*, son parámetros desconocidos.*

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

que siguen, completando las tres columnas de la derecha de la tabla:

y	x_1	x_2	$\widehat{y}$ *(con (2.26))*	$\widehat{y}$ *(DB* $k=4$*)*	$\widehat{y}$ *(DB* $k=5$*)*
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			
0.356	68	640			

2. *Compara las tres predicciones calculando la cantidad*

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

3. *En las predicciones DB, ¿ha sido necesario conocer el modelo de regresión no lineal (2.26)? ¿Qué ventajas le encuentras al método DB?*

Ejercicio 2.10 *La Tabla 2.3 procede de Cox y Cox (2001, p.91) y 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. *Interpretando la matriz binaria simétrica como una matriz de similaridades, y tomando convencionalmente el valor 3 en la diagonal principal, representa los 15 países por análisis de coordenadas principales.*

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 tomando 5 dimensiones.*

3. *Comenta los problemas encontrados con la regresión clásica y discute las ventajas del método DB para estos datos.*

4. *Prueba que el valor asignado a la diagonal principal (el 3 en nuestro caso) no altera la obtención de los ejes principales, es decir, la regresión DB no cambia si asignamos otro valor con tal de que la matriz de similaridaades resultante sea (semi) definida positiva.*

TABLA 2.3 (véase **ejercicios.txt**)

		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 2.4

	N	C	R	E	Z	Ei	J	G	M	precios
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 *Supongamos que* $1, 2, ..., n$ *representan* n *estímulos y que existe una escala de preferencias* $\theta_1, \theta_2, ..., \theta_n$*, donde* θ_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* $j \succ i$*, o también* $i \prec j$*. A fin de obtener la ordenación de los estímulos, los ordenamos según preferencias*

$$i_1 \prec i_2 \prec \cdots \prec i_n,$$

en el sentido de que se cumple

$$\theta_{i_1} < \theta_{i_2} < \cdots < \theta_{i_n}.$$

Para estimar los parámetros, se obtienen las proporciones

$$p_{ij} = P\left(j \succ i\right),$$

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

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

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

1. *Interpreta el modelo (2.27) comentando qué ocurriría en los casos:*

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

2. *Definimos la siguiente distancia entre cada par de estímulos*

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

   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 muestra 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:*

$$P(C \succ N) = 0.818 \qquad P(E \succ C) = 0.723 \qquad P(R \succ G) = 0.203.$$

Utilizando la distancia (2.28) se ha obtenido la matriz de distancias de la Tabla 2.4. Aplica un análisis de coordenadas principales y encuentra la escala de preferencias, identificando cada vegetal con las coordenadas de la primera dimensión principal.

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 2).*

2. *Los precios (por kilo de producto) de los 8 primeros vegetales son:*

   $$120 \quad 110 \quad 130 \quad 145 \quad 160 \quad 170 \quad 200 \quad 225$$

   Predice el precio del maíz.

Capítulo 3

Análisis discriminante DB

3.1 Introducción

Supongamos que Ω_1, Ω_2 son dos poblaciones, $X_1, ...,X_p$ son p variables observables y $\mathbf{x} = (x_1, ..., x_p)$ contiene las observaciones de las variables sobre un individuo $\omega \in \Omega_1 \cup \Omega_2$. Queremos asignar ω a una de las dos poblaciones.

Una **regla discriminante** es un criterio que permite asignar ω, y que a menudo se plantea en términos de una **función discriminante** $D(x_1, ..., x_p)$. Entonces la regla de clasificación es:

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

Las cuatro reglas mas importantes del análisis discriminante son:

- **Regla MV o de la máxima verosimilitud**

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

$$V(\mathbf{x}) = \log\ f_1(\mathbf{x}) - \log\ f_2(\mathbf{x}). \tag{3.1}$$

- **Regla B o regla de Bayes**

Supongamos que son conocidas las probabilidades "a priori":

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

La función discriminante es:

$$B(\mathbf{x}) = \log f_1(\mathbf{x}) - \log\ f_2(\mathbf{x}) + \log(q_1/q_2). \tag{3.2}$$

- **Regla M o de Matusita**

Supongamos que estamos en condiciones de calcular una distancia $\delta_i = \delta(\omega, \Omega_i)$, $i = 1, 2$, de ω 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 Ω_i, por ejemplo el vector de medias.

La regla M está basada en la función discriminante

$$M(\mathbf{x}) = \delta_2^2(\mathbf{x}) - \delta_1^2(\mathbf{x}). \tag{3.3}$$

- **Regla de Fisher del discriminador lineal**

Si Ω_i es $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, es decir, se puede identificar Ω_i mediante un vector de medias $\boldsymbol{\mu}_i$ y una matriz de covariancias $\boldsymbol{\Sigma}$ (común a las dos poblaciones), entonces

$$L(\mathbf{x}) = \left[\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{3.4}$$

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

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

  $$\delta_j^2 = \left(\mathbf{x} - \boldsymbol{\mu}_j\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_j\right) \tag{3.5}$$

  es la distancia (al cuadrado) de Mahalanobis de $\mathbf{x}$ a $\boldsymbol{\mu}_j$.

Todas esas reglas coinciden esencialmente 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 Ω_i es:

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

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

Ejercicio 3.1 *Supóngase que Ω_i es una población $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$, se pide:*

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 llamado discriminador cuadrático*

$$\begin{aligned} Q(\mathbf{x}) &= \tfrac{1}{2}\mathbf{x}'\left(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\mathbf{x} + \mathbf{x}'\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) \\ &\quad + \tfrac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \tfrac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_2| - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_1|. \end{aligned} \tag{3.6}$$

3.2 La función de proximidad de un individuo a una población

La distancia (3.5) proporciona una medida de la proximidad o distancia de ña observación multivariante $\mathbf{x}$ a la población de la cual procede, Queremos generalizar esta medida haciendo uso solamente de **distancias entre observaciones**, sin necesidad de trabajar con vectores de medias, que son parámetros relativos a la población.

Sea δ una distancia sobre Ω, $\mathbf{x} = (x_1, ..., x_p)$ una observación de un vector aleatorio $\mathbf{X}$, con densidad $f(x_1, ..., x_p)$ y soporte S. Llamaremos **variabilidad geométrica** relativa a la distancia δ, a la cantidad

$$V_\delta(\mathbf{X}) = \tfrac{1}{2}\int_{S\times S} \delta^2(\mathbf{x}, \mathbf{y})\, f(\mathbf{x})\, f(\mathbf{y})\, d\mathbf{x}d\mathbf{y}. \tag{3.7}$$

$V_\delta(\mathbf{X})$ es el valor esperado de todas las interdistancias (al cuadrado). Podemos entender $V_\delta(\mathbf{X})$ como una varianza generalizada. Cuando la distancia tiene la propiedad Euclídea, entonces $V_\delta(\mathbf{X})$ coincide con el concepto de **inercia**.

Ejercicio 3.2 *Demuestra que:*

1. *Si X es una variable aleatoria con varianza finita y $\delta(x, y) = |x - y|$, entonces*
$$V_\delta(X) = \mathrm{var}(X).$$
2. *Si $\mathbf{X}$ es un vector aleatorio con matriz de covarianzas $\boldsymbol{\Sigma}$, y $\delta = d_E$ es la distancia Euclídea (1.6), entonces*
$$V_{d_E}(\mathbf{X}) = \mathrm{tra}(\boldsymbol{\Sigma}).$$

Sea ω un individuo de Ω, y $\mathbf{x} = (x_1, ..., x_p)$ las observaciones de $\mathbf{X}$ sobre ω. Dada una distancia δ sobre Ω, definimos la **función de proximidad** (o simplemente **proximidad**) de ω a Ω en relación con $\mathbf{X}$ a la función

$$\phi_\delta^2(\mathbf{x}) = E\left[\delta^2(\mathbf{x}, \mathbf{Y})\right] - V_\delta(\mathbf{X}) = \int_S \delta^2(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y} - V_\delta(\mathbf{X}). \quad (3.8)$$

$\phi_\delta^2(\mathbf{x})$ es el valor esperado de la distancia (al cuadrado) de $\mathbf{x}$ (que es fijo) a $\mathbf{y}$ (que es aleatorio), menos la variabilidad geométrica. Como veremos en el Teorema 4, $\phi_\delta^2(\mathbf{x})$ puede expresarse como una distancia (al cuadrado) entre $\mathbf{x}$ y un cierto vector medio asociado a la población.

Ejercicio 3.3 *Se pide:*

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

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

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

$$\phi_\delta^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}). \quad (3.9)$$

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

$$\begin{aligned} \mathbf{x} &\to \psi(\mathbf{x}) \\ \delta(\mathbf{x}, \mathbf{y}) &= d_e(\psi(\mathbf{x}), \psi(\mathbf{y})), \end{aligned} \quad (3.10)$$

siendo d_e la distancia Euclídea (1.6). Diremos que ψ proporciona una representación Euclídea de δ.

El siguiente Teorema generaliza (3.9), de manera que la **función de proximidad** se puede interpretar como una distancia del individuo a la población.

Teorema 4 *Supongamos que existe una representación Euclídea de (Ω, δ) en un espacio L (Euclídeo o de Hilbert separable) con un producto escalar $< ., . >$ y una norma $\|\mathbf{z}\|^2 = < \mathbf{z}, \mathbf{z} >$, tal que*

$$\delta^2(\mathbf{x}, \mathbf{y}) = \|\psi(\mathbf{x}) - \psi(\mathbf{y})\|^2, \quad (3.11)$$

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

$$\phi_\delta^2(\mathbf{x}_0) = \|\psi(\mathbf{x}_0) - E[\psi(\mathbf{X})]\|,^2, \quad (3.12)$$

donde $\mathbf{x}_0$ contiene las coordenadas de un individo fijo ω.

3.3 La regla discriminante DB

Sean ahora Ω_1, Ω_2 dos poblaciones y δ una función de distancia. δ es la misma en cada población (en el sentido de que se expresa igual), pero puede tener versiones ligeramente diferentes, digamos δ_1, δ_2, cuando estemos en Ω_1, Ω_2, respectivamente. Por ejemplo, si las poblaciones son normal N_p $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, y consideramos las distancias de Mahalanobis

$$\delta_1^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})' \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \mathbf{y}),$$

$$\delta_2^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})' \boldsymbol{\Sigma}_2^{-1} (\mathbf{x} - \mathbf{y}),$$

entonces lo único que cambia es la matriz de covarianzas. Lo que debe quedar bien claro es que δ depende del vector aleatorio $\mathbf{X}$, que en general tendrá diferente distribución en Ω_1 y Ω_2.

Seguidamente, mediante (3.8), calcularemos las funciones de proximidad ϕ_1^2, ϕ_2^2, correspondientes a Ω_1, Ω_2. Sea ω un individuao para ser clasificado en Ω_1 o en Ω_2, con valores $\mathbf{x}_0 = \mathbf{X}(\omega)$.

La regla de clasificación DB es:

$$\begin{array}{l} \text{Asignar } \omega \text{ a } \Omega_1 \text{ si } \phi_1^2(\mathbf{x}_0) \leq \phi_2^2(\mathbf{x}_0), \\ \text{en otro caso, asignar } \omega \text{ a } \Omega_2. \end{array} \tag{3.13}$$

Teniendo en cuenta que, si existe la aplicación (3.10), se cumple

$$\begin{array}{l} \phi_1^2(\mathbf{x}_0) = \|\psi(\mathbf{x}_0) - E_{\Omega_1}[\psi(\mathbf{X})]\|^2, \\ \phi_2^2(\mathbf{x}_0) = \|\psi(\mathbf{x}_0) - E_{\Omega_2}[\psi(\mathbf{X})]\|^2, \end{array} \tag{3.14}$$

vemos que la regla DB asigna ω a la población más próxima. Por lo tanto, la regla DB es una regla del tipo M, pero que depende solamente de las distancias entre individuos.

3.4 Propiedades de la función de proximidad

Transformando la distancia δ también se transforma la proximidad ϕ^2. Se cumple:

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

$$\widetilde{\phi}^2 = c\phi^2. \tag{3.15}$$

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

$$\widetilde{\phi}^2 = \phi^2 + b/2. \tag{3.16}$$

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

$$\phi^2 = \phi_1^2 + \phi_2^2. \tag{3.17}$$

Ejercicio 3.4 *Se pide:*

1. *Demuestra las propiedades (3.15)-(3.17).*
2. *Razona que la aditividad es consistente con la independencia, es decir, la propiedad aditiva (3.17) resulta apropiada cuando juntamos dos variables independientes X, Y, con δ_1 asociada a X y δ_2 asociada a Y.*

3.5 La regla DB comparada con algunas reglas clásicas

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

$$L(\mathbf{x}_0) = \tfrac{1}{2}\left[\phi_2^2(\mathbf{x}_0) - \phi_1^2(\mathbf{x}_0)\right], \tag{3.18}$$

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

Análogamente si consideramos la distancia

$$\begin{array}{ll} \delta_j^2(\mathbf{x},\mathbf{y}) & = (\mathbf{x}-\mathbf{y})' \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\mathbf{y}) + \frac{1}{2}\log|\boldsymbol{\Sigma}_j| \qquad \mathbf{x} \neq \mathbf{y}, \\ & = 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \mathbf{x} = \mathbf{y}, \end{array} \tag{3.19}$$

entonces el discriminador cuadrático (3.6) es

$$Q(\mathbf{x}_0) = \tfrac{1}{2}\left[\phi_2^2(\mathbf{x}_0) - \phi_1^2(\mathbf{x}_0)\right]. \tag{3.20}$$

Finalmente, el llamado **discriminador Euclídeo**

$$E(\mathbf{x}_0) = \left[\mathbf{x}_0 - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{3.21}$$

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 hace falta calcular la inversa de $\boldsymbol{\Sigma}$.

Ejercicio 3.5 *Se pide:*

1. *¿Cuál es la distancia adecuada a fin de hallar (3.18)?*
2. *En el cálculo de (3.21), ¿cuál es la matriz de covariancias común a* Ω_1, Ω_2*?*
3. *Demuestra (3.20).*
4. *¿Qué le ocurriría a la distancia (3.19) si* $|\mathbf{\Sigma}_i| < 1$*? ¿Cómo podemos superar este aparente inconveniente?*

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

Supongamos ahora que Ω_1, Ω_2 son dos poblaciones multinomiales y que, respecto a unas características cualitativas $A_1, ..., A_m$ (sucesos excluyentes), sus probabilidades son

$$\mathbf{p}_1 = (p_{11}, ..., p_{1m}), \quad p_{1k} \geq 0, \quad \sum p_{1k} = 1,$$
$$\mathbf{p}_2 = (p_{21}, ..., p_{2m}), \quad p_{2k} \geq 0, \quad \sum p_{2k} = 1.$$

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

$$\text{Asignar } \omega \text{ a } \Omega_i \text{ si } p_{ik} = \max\{p_{1k}, p_{2k}\}. \tag{3.22}$$

Consideremos ahora la distancia entre individuos de un mismo Ω_i

$$\delta_i^2(\omega, \omega') = \begin{cases} 0 & \text{si} \quad \omega, \omega' \in A_k, \\ \left(p_{ik}^{-1} + p_{ik'}^{-1}\right) & \text{si} \quad \omega \in A_k,\ \omega' \in A_{k'}, \end{cases} \tag{3.23}$$

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

Los valores que toman las funciones de proximidad para un individuo ω tal que presenta la característica A_k, son:

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

Se verifica que:

$$\min\left\{\phi_1^2(k), \phi_2^2(k)\right\} = \phi_i^2(k) \Longleftrightarrow p_{ik} = \max\{p_{1k}, p_{2k}\}, \tag{3.24}$$

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

Ejercicio 3.6 *Se pide:*

1. *Demuestra que en una población multinomial con m estados excluyentes, la variabilidad geométrica de la distancia (3.23) es igual a* $(m-1)$.

2. *Demuestra la implicación (3.24).*

Ejercicio 3.7 *Supongamos dos poblaciones normales*

$$\Omega_1 \text{ es } N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \qquad \Omega_2 \text{ es } N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

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*

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

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

$$pce = \tfrac{1}{2} P(\mathbf{x} \to \Omega_1 | \Omega_2) + \tfrac{1}{2} P(\mathbf{x} \to \Omega_2 | \Omega_1).$$

   Demuestra que es igual a

$$pce = \Phi(-M/2) \tag{3.25}$$

   donde Φ *es la función de distribución* $N(0,1)$ *y*

$$M^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{3.26}$$

   es la distancia de Mahalanobis entre Ω_1 *y* Ω_2.

4. *Calcula (3.25) bajo las condiciones del apartado 1.*

3.6 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 $X_1, ..., X_p$. Indiquemos (las representaciones Euclídeas de las muestras) por

$$\begin{array}{ll} \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{n_1} & \text{muestra de } \Omega_1, \\ \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{n_2} & \text{muestra de } \Omega_2. \end{array} \tag{3.27}$$

Considerando primeramente solo una población. Dada una distancia δ, y una muestra $\mathbf{x}_1, ..., \mathbf{x}_n$, una estimación de la variabilidad geométrica (3.7) es

$$\widehat{V}_\delta(\mathbf{X}) = \frac{1}{2n^2} \sum_{i,j=1}^{n} \delta^2(\mathbf{x}_i, \mathbf{x}_j). \tag{3.28}$$

Conviene observar que este estimador es un ejemplo de U-estadístico, y por lo tanto es el mejor estimador no-paramétrico de la variabilidad geométrica poblacional.

Si ω es un individuo, una estimación de la función de proximidad (3.8) es

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

siendo $\delta^2(\mathbf{x}_0, \mathbf{x}_i)$, $i = 1, ..., n$, las distancias (al cuadrado) de ω a las observaciones de la muestra, donde ω está representado por la imagen $\mathbf{x}_0$, véase (3.10), aunque esta suposición no es necesaria, ya que las distancias las podemos calcular igualmente partiendo de los datos originales.

Volviendo al caso de dos poblaciones, si ω es un individuo a clasificar, las estimaciones de las dos funciones de proximidad, en base a las muestras (3.27), son (ver (3.8)):

$$\begin{array}{ll} \widehat{\phi}^2_1(\mathbf{x}_0) & = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta^2(\mathbf{x}_0, \mathbf{x}_i) - \frac{1}{2n_1^2} \sum_{i,j=1}^{n_1} \delta^2(\mathbf{x}_i, \mathbf{x}_j), \\ \widehat{\phi}^2_2(\mathbf{y}_0) & = \frac{1}{n_2} \sum_{i=1}^{n_2} \delta^2(\mathbf{y}_0, \mathbf{y}_i) - \frac{1}{2n_2^2} \sum_{i,j=1}^{n_2} \delta^2(\mathbf{y}_i, \mathbf{y}_j), \end{array} \tag{3.29}$$

donde $\mathbf{x}_0, \mathbf{y}_0$ son las dos imágenes de ω, véase (3.10), según consideremos que ω pertenezca a Ω_1 o a Ω_2. Sin embargo, recordemos que el conocimiento de $\mathbf{x}_0, \mathbf{y}_0$, no es necesario, puesto que solamente necesitamos las distancias entre observaciones, a fin de obtener las funciones de proximidad $\widehat{\phi}^2_1$, $\widehat{\phi}^2_2$.

La regla de clasificación DB es entonces:

$$\text{Asignar } \omega \text{ a } \Omega_i \quad \text{si} \quad \widehat{\phi}_i^2 = \min\left\{\widehat{\phi}_1^2, \widehat{\phi}_2^2\right\}. \tag{3.30}$$

El siguiente teorema nos demuestra que podemos entender (3.30) como una regla M, cuando la citada representación existe.

Teorema 5 *Supongamos que en dos espacios Euclídeos (posiblemente diferentes) podemos representar ω y las dos muestras como*

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

respectivamente. Entonces se cumple que

$$\widehat{\phi}_1^2 = d_E^2\left(\mathbf{x}_0, \overline{\mathbf{x}}\right), \qquad \widehat{\phi}_2^2 = d_E^2\left(\mathbf{y}_0, \overline{\mathbf{y}}\right), \tag{3.31}$$

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

Como consecuencia del Teorema 5, podemos formular la regla DB como sigue:

$$\begin{array}{l} \text{Asignar } \omega \text{ a } \Omega_1 \text{ si } d_E^2\left(\mathbf{x}_0, \overline{\mathbf{x}}\right) < d_E^2\left(\mathbf{y}_0, \overline{\mathbf{y}}\right), \\ \text{en otro caso asignar } \omega \text{ a } \Omega_2. \end{array} \tag{3.32}$$

Por lo tanto, asignamos ω a la población que tiene más próxima.

Ejercicio 3.8 *Sea $\mathbf{\Delta}_1 = (\delta_{ij})$ la matriz de distancias $n_1 \times n_1$ entre las muestras de la primera población. Sean $\delta_1, \delta_2, ..., \delta_{n_1}$ las distancias de un individo ω a cada uno de los elementos de la muestra. Se pide:*

1. *¿Cómo podemos encontrar una representación en los términos del Teorema 5?*

2. *Demuestra que si podemos expresar*

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

entonces

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

3. *Utiliza este resultado para probar que* $\widehat{\phi}_1^2(x_0) = d_E^2(\mathbf{x}_0, \overline{\mathbf{x}})$.

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

Ejercicio 3.9 *Tenemos dos poblaciones* Ω_1, Ω_2 *y una muestra de tamaño 5 de cada población. respecto a tres variables mixtas. Las matrices de distancias entre los 5 individuos de la muestra 1 y los 5 de la muestra 2, de cada una de las poblaciones son, respectivamente:*

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

Las distancias de un individuo ω *a los individuos de cada una de las muestras son*

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

Clasifica ω *mediante la regla discriminante DB.*

3.7 Ventajas del método DB

El método de discriminación DB ("distance-based") no necesita conocer las funciones de densidad, que a menudo son difíciles o imposibles 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 similaridad de Gower.

Observa que una vez hemos calculado las distancias, la estimación de las funciones de proximidad (3.29) es muy sencilla.

En resumen, el método DB:

- Es equivalente al discriminador lineal (3.4) si tomamos la distancia de Mahalanobis. Es fácil ver que una variante de esta distancia proporciona el discriminador cuadrático (3.6).
- Puede abordar variables binarias utilizando (por ejemplo) el coeficiente de Jaccard.
- Puede abordar variables nominales (secuencias de ADN, palabras de un idioma, etc.) 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 (1989b, 1991) e ilustrado con ejemplos por Cuadras (1992a). Los Teoremas 4 y 5 están demostrados en Cuadras, Fortiana y Oliva (1997), artículo que contiene otros ejemplos con datos reales. Previamente, Krzanowski (1980) propuso un análisis discriminante para variables muixtas basado en el "location model". Pero el método DB clasifica mejor.

Ejercicio 3.10 *En un estudio de cáncer que afectó a 137 mujeres, se consideraron 11 variables (7 contínuas, 2 binarias y 3 categóricas) capaces de clasificar un tumor en benigno o maligno. En un primer grupo de $n_1 = 78$ mujeres el tumor resultó benigno, mientras que en otro grupo de $n_2 = 59$ mujeres el tumor era maligno. Con los datos del archivo* ***ejercicios.txt****, calcula el número de errores de clasificación, utilizando los discriminadores lineal (3.4), cuadrático (3.6) y Euclídeo (3.21), como así también la regla DB con la distancia de Gower. ¿Qué discriminador clasifica mejor?*

Ejercicio 3.11 *Supongamos que la matriz de covarianzas $\mathbf{\Sigma}$ es singular, es decir, existen combinaciones lineales entre las variables.*

1. *Razona entonces que (3.21) puede ser un discriminador adecuado.*
2. *Explica que, en general, el método DB sería aplicable en esta situación de singularidad.*

3.8 Discriminación en el caso de varias poblaciones

Supongamos que disponemos de $g > 2$ poblaciones $\Omega_1, ..., \Omega_g$ y muestras de tamaños $n_1, ..., n_g$ para cada población. Supongamos que en relación a p variables (posiblemente mixtas) y a una distancia δ, hemos calculado las funciones de proximidad $\widehat{\phi}_1^2(\omega), ..., \widehat{\phi}_g^2(\omega)$ correspondientes a un individuo ω a clasificar (véase (3.29)). La generalización de (3.30) es inmediata. La regla DB para $g > 2$ poblaciones es:

$$\text{Asignar } \omega \text{ a } \Omega_i \quad \text{si} \quad \widehat{\phi}_i^2(\omega) = \min\left\{\widehat{\phi}_1^2(\omega), ..., \widehat{\phi}_g^2(\omega)\right\}. \tag{3.33}$$

Se puede probar un resultado similar al Teorema 5. Por lo tanto la regla DB clasifica ω a la población que tiene más próxima.

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

$$\mathbf{\Delta}^2(k, k') = \frac{1}{n_k n_{k'}} \sum_{i,j=1}^{n} \delta^2 \left(\mathbf{x}(k)_i, \mathbf{x}(k')_j \right) - V_\delta(k) - V_\delta(k'), \qquad (3.34)$$

donde $V_\delta(k)$ es la variabilidad geométrica de Ω_k, tenemos la posibilidad de representar las g poblaciones por análisis de coordenadas principales sobre la matriz de orden $g \times g$, conteniendo las distancias $(\mathbf{\Delta}(k, k'))$. 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. Véase Cuadras (1991), Krzanowski (1994), Cuadras, Fortiana y Oliva (1996, 1997).

Finalmente, debemos 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 3.12 *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 distancia de Mahalanobis, es decir, prueba que (3.34) se reduce a (1.20).*

2. *Con los datos del archivo* ***ejercicios.txt****, 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.*

Capítulo 4

Aspectos computacionales en regresión DB

4.1 Selección de variables en el modelo 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 por orden decreciente de los valores propios, es un método para obtener las dimensiones principales que permitan realizar la mejor representación en dimensión k, de Ω, de modo que se maximice la variabilidad geométrica en la dimensión k elegida:

$$\frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2(k) = \frac{1}{n}\sum_{i=1}^{k}\lambda_i. \tag{4.1}$$

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

$$r^2(\mathbf{y},X_{i_1}) > \cdots > r^2(\mathbf{y},X_{i_k}). \tag{4.2}$$

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

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

La variabilidad geométrica DB condicional, se define como:

$$\frac{1}{2n^2}\sum_{i,j=1}^{n}(\widehat{y}_i-\widehat{y}_j)^2. \tag{4.3}$$

Es fácil probar que

$$\frac{1}{2n^2}\sum_{i,j=1}^{n}(\widehat{y}_i-\widehat{y}_j)^2=\frac{1}{n}\sum_{\alpha=1}^{k}\widehat{\beta}_{i_\alpha}\lambda_{i_\alpha}=R^2(k)S_y^2, \tag{4.4}$$

donde S_y^2 es la varianza muestral de Y. Esta ecuación puede ser interpretada como la versión DB de (4.1). Nótese que el criterio (4.2) maximiza $R^2(k)$, es decir brinda el subespacio en dimensión k, generado por las dimensiones principales que tienen máxima correlación múltiple con Y.

Ejercicio 4.1 *Sea* $\overline{\mathbf{X}}$ *la matriz de datos centrados y* $\mathbf{S}=n^{-1}\overline{\mathbf{X}}'\overline{\mathbf{X}}$ *la matriz de covarianzas de las variables columna de* $\overline{\mathbf{X}}$*. La transformación de componentes principales es* $\mathbf{Y}=\overline{\mathbf{X}}\mathbf{T}$*, siendo* $\mathbf{S}=\mathbf{TDT}'$ *la descomposición espectral de* $\mathbf{S}$*. Relacionando* $\mathbf{S}$ *con* $\mathbf{B}=\mathbf{XX}'$ *(ver 1.13), demuestra que las coordenadas principales (columnas de* $\mathbf{X}$*) pueden ser interpretadas como componentes principales.*

Ejercicio 4.2 *Determina la distancia implícita en la definición (4.3) y demuestra (4.4).*

4.2 Selección para un número grande de individuos

Los dos métodos de la Sección 4.1 se basan en maximizar (4.3) ó (4.4) y necesitan calcular los vectores propios de la matriz $\mathbf{B}$. Cuando n es muy grande, este cálculo puede ser muy costoso.

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 a sus valores propios.

Se define la secuencia $c(i)$:

$$c(0)=0,\quad c(i)=\mathbf{y}'\mathbf{B}_{(i)}\mathbf{y}/\mathbf{y}'\mathbf{B}\mathbf{y},\qquad i=1,2,\ldots,m, \tag{4.5}$$

siendo m =rang($\mathbf{B}$) y $\mathbf{B}_{(i)} = \mathbf{X}_{(i)}\mathbf{X}_i'$. Este coeficiente verifica:

$$c(i) = \frac{\sum_{\alpha=1}^{i} r_\alpha^2 \lambda_\alpha}{\sum_{\alpha=1}^{m} r_\alpha^2 \lambda_\alpha}, \tag{4.6}$$

donde r_α =corr($\mathbf{y}, \mathbf{X}_\alpha$). Cada $c(i)$ mide la predictibilidad de las primeras i dimensiones, ponderadas por los correspondientes valores propios. $c(0) = 0$, puede ser interpretadao como la falta de predicitibilidad de $\mathbf{1}$, el vector de unos, que es también un vector propio de $\mathbf{B}$.

Esta secuencia satisface:

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

La selección debe ser realizada representando en un diagrama los puntos

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

donde m^* es tal que $1 - c(i)$ toma un valor muy próximo a 0. Esto nos permite decidir el corte en m^*, de modo que después de m^*, el gráfico lineal se sitúa próximo al eje horizontal, indicando qué las dimensiones superiores deben descartarse. 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 dimesión i es aceptada o rechazada en función de que r_i^2 o λ_i sean grandes o pequeños.

Finalmente, destaquemos que este procedimiento solamente requiere el cálculo de los primeros m^* vectores propios.

Ejercicio 4.3 *Se pide:*

1. *Construye el gráfico de los puntos $(i, 1 - c(i))$ unidos con segmentos rectos (ver Figura 4.1), para los datos de automóbiles del Ejercicio 2.4.*
2. *Prueba (4.6) e interpreta una dimensión problemática, es decir, una dimensión tal que λ_i sea grande y r_i^2 sea pequeño, o recíprocamente, λ_i sea pequeño pero r_i^2 sea grande.*

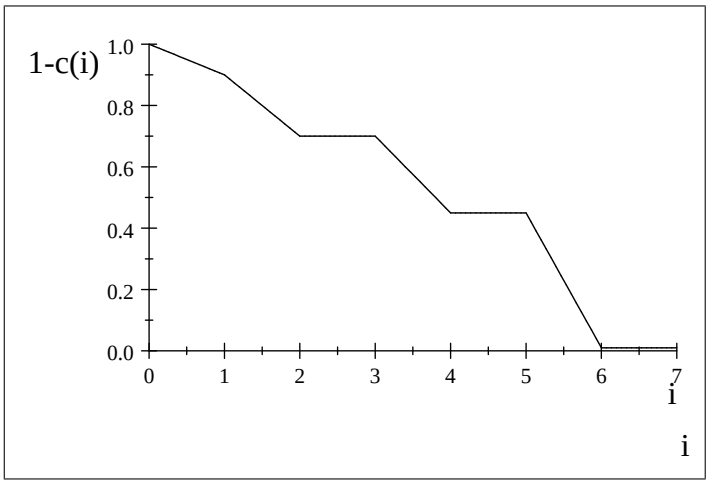


Figura 4.1: Gráfico para seleccionar las dimensiones predictivas en la regresión DB. Las dimensiones 1, 2, 4 y 6 serían predictivas.

4.3 Valores propios negativos o muy pequeños

Supongamos que la matriz $n \times n$ de distancias observadas, $\mathbf{\Delta} = (\delta_{ij})$, no es Euclídea. Puede ocurrir, por ejemplo, que se emplee el coeficiente de similaridad de Gower con algunos datos faltantes. Entonces la matriz $\mathbf{B}$ tendrá algunos valores propios negativos (Teoremas 1 y 2), por lo tanto tendremos algunas variables (columnas de $\mathbf{X}$) con "varianzas negativas". Consecuentemente el modelo de regresión DB no funcionaría.

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

$$\widetilde{\delta}_{ij}^2 = \begin{cases} 0 & i = j, \\ \delta_{ij}^2 + 2a & i \neq j, \end{cases} \tag{4.7}$$

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

$$\widehat{\mathbf{B}} = \mathbf{B} + a\mathbf{H}, \tag{4.8}$$

siendo $\mathbf{H}$ la matriz de centrado. Los vectores propios de $\widehat{\mathbf{B}}$ son los mismos y los valores propios son $\lambda_i + a$

$$\widehat{\mathbf{B}} X_i = (\lambda_i + a) X_i. \tag{4.9}$$

Si $\lambda_i < 0$ para algún i, podemos escoger $a > 0$ con tal que $\lambda_i + a > 0, i = 1, \ldots, m$, lo que nos permite conseguir la matriz de distancias Euclídeas $\widetilde{\mathbf{\Delta}} = (\widetilde{\delta}_{ij})$.

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

$$\widehat{\mathbf{y}}_{n+1} = \overline{\mathbf{y}} + \tfrac{1}{2}(\mathbf{b} - \mathbf{d})' \mathbf{U} \mathbf{\Lambda}^{-1} \mathbf{U}' \mathbf{y}, \tag{4.10}$$

la transformación q-aditiva (4.7), cambia el modelo DB a

$$\widehat{\mathbf{y}}_{n+1}(a) = \overline{\mathbf{y}} + \tfrac{1}{2}(\mathbf{b} - \mathbf{d})' \mathbf{U} (\mathbf{\Lambda} + a\mathbf{I})^{-1} \mathbf{U}' \mathbf{y}, \tag{4.11}$$

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

Ejercicio 4.4 *Se pide:*

1. *Demuestra las ecuaciones (4.9)-(4.11).*
2. *Propógase un criterio para seleccionar la mejor constante a en (4.11).*
3. *Escribe una versión apropiada de (4.11) para el modelo DB en dimensión k.*

Supongamos ahora que $\mathbf{\Delta} = (\delta_{ij})$ es una matriz de distancias Euclídea, pero algunos valores propios de $\mathbf{B}$ son muy pequeños. Entonces el ajuste de (4.10) puede quedar alterado por motivos numéricos, pues se invierten cantidades muy pequeñas.

Una solución al problema pasa también por emplear (4.11), que requiere la inversa de la matriz $(\mathbf{\Lambda} + a\mathbf{I})$, cuya inversa es más asequible, pudiendo mejorar el ajuste. Este inconveniente también se presenta para la estimación de los parámetros de regresión, (véase 2.12):

$$\widehat{\beta}_i(a) = \mathbf{X}_i' \mathbf{y} / (\lambda_i + a). \tag{4.12}$$

Existe una clara analogía entre (4.11) y (4.12) y la **regresión cresta** ("ridge regression", Hoerl y Kennard, 1970). La transformación (4.7), puede ser interpretada como una extensión de este procedimiento de regresión para una distancia general.

Ejercicio 4.5 *Se pide:*

1. *Construye un ejemplo donde el ajuste con el modelo DB proporcione una correlación múltiple $R^2 = 0$ y otro tal que $R^2 = 1$.*
2. *Describe la regresión cresta ("ridge regression") y comenta la versión DB.*

4.4 Relacionando dos conjuntos de variables

Supongamos que en un mismo conjunto finito Ω, se tienen dos conjuntos de variables $X_1, \ldots, X_p$ y las variables $Y_1, \ldots, Y_q$,posiblemente mixtas. Nuestro propósito es relacionar ambos conjuntos a través de la definición de una medida apropiada de asociación.

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

	$\mathbf{X}$	$\mathbf{Y}$
$\mathbf{X}$	$\mathbf{R}_{11}$	$\mathbf{R}_{12}$
$\mathbf{Y}$	$\mathbf{R}_{21}$	$\mathbf{R}_{22}$

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

$$\begin{array}{ll} \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{a}_i = r_i^2\mathbf{R}_{11}\mathbf{a}_i & i = 1, \ldots, m, \\ \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{b}_i = r_i^2\mathbf{R}_{22}\mathbf{b}_i & m = \min(p, q). \end{array} \tag{4.13}$$

Estas ecuaciones también pueden ser escritas empleando las matrices de covarianzas, con los mismos resultados en cuanto a valores propios.

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

$$U_i = \mathbf{a}_i'\mathbf{X}, \qquad V_i = \mathbf{b}_i'\mathbf{Y}, \qquad i = 1, \ldots, m$$

son las **variables canónicas**, cuyas correlaciones

$$r_i = \text{corr}(U_i, V_i), \quad i = 1, \ldots, m,$$

son las raíces cuadradas positivas de los valores propios, que se suponen ordenadas

$$r_1 \geq r_2 \geq \cdots \geq r_m, \tag{4.14}$$

y se denominan **correlaciones canónicas**. La primera 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 $U_1, \ldots, U_m$ son incorrelacionadas.

Medidas globales de asociación pueden basarse en (4.14), por ejemplo:

$$\eta^2 = \prod_{i=1}^{m} r_i^2, \qquad \theta^2 = 1 - \prod_{i=1}^{m}(1 - r_i^2). \tag{4.15}$$

Como veremos enseguida, es preferible utilizar la segunda medida.

Ejercicio 4.6 *Se pide:*

1. *Demuestra que con las dos ecuaciones (4.13) se obtienen los mismos valores propios.*

2. *Comprueba que* $0 \leq \eta^2 \leq 1$ *y* $0 \leq \theta^2 \leq 1$. *Discute los casos* 0 *y* 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, 2016).

4.5 Relación DB entre variables mixtas

Supongamos que sobre una misma población $\Omega = \{1, 2, \ldots, n\}$ se dispone de dos conjuntos de datos mixtos M_1, M_2 que se desean relacionar y comparar. Una forma de lograrlo es haciendo una extensión de la regresión DB. Suponiendo que se dispone de una función de distancia δ, que sea apropiada y que permita calcular 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) y consideremos las descomposiciones espectrales de $\mathbf{B}_1$ y $\mathbf{B}_2$, que proporcionan las matrices de coordenadas principales

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

Entonces M_1, M_2 pueden representarse por las matrices cuantitativas $\mathbf{X}, \mathbf{Y}$. Empleando las filas de $\mathbf{X}$, coordenadas principales relativas a $\mathbf{\Delta}_1$, se puede representar Ω en un espcio Euclídeo de menor dimensión. Las filas de $\mathbf{U}_1$ se denominan **coordenadas estándar**. Similarmente, obtenemos las coordenadas principales $\mathbf{Y}$, que nos permiten representar Ω 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 $\mathbf{Y}$, matrices de órdenes $n \times p$ y $n \times q$, podemos plantear el modelo de regresión DB multivariante

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{4.16}$$

donde $\mathbf{B}$ es una matriz de parámetros de orden $p \times q$, y $\mathbf{E}$, es una matriz de errores aleatorios de orden $n \times q$. Si se tiene que $\mathbf{B} = \mathbf{0}$, entonces no hay relación entre los datos M_1 y M_2.

Para medir la relación entre M_1 y M_2, podemos adoptar la medida $\eta^2 = \det(\mathbf{U}_2'\mathbf{U}_1\mathbf{U}_1'\mathbf{U}_2)$, que en (4.15) se expresa en términos de correlaciones canónicas entre $\mathbf{X}, \mathbf{Y}$. Sin embargo, es más adecuado utilizar

$$\theta^2 = 1 - \det(\mathbf{I}_n - \mathbf{U}_2'\mathbf{U}_1\mathbf{U}_1'\mathbf{U}_2) = 1 - \prod_{i=1}^{m}(1 - r_i^2). \tag{4.17}$$

En efecto, al contrario de η^2, la medida de asociación generalizada θ^2 está poco influenciada por las correlaciones canónica muy pequeñas. Por ejemplo, si r_1 es grande pero r_m es muy pequeño, r_m^2 disminuye mucho el valor de η^2, mientras que $(1 - r_m^2)$ apenas influiría en θ^2. Este coeficiente ha sido utilizado por Cuadras *et al.* (2012) en el tratamiento de imágenes espectrales mediante distancias.

Suponiendo normalidad multivariante, se puede probar que aplicando el test de la razón de verosimilitud λ al modelo (4.16) para decidir si se acepta la hipótesis nula $H_0 : \mathbf{B} = \mathbf{0}$, entonces se obtiene $\lambda = 1 - \theta^2$, es decir, el coeficiente θ^2 guarda estrecha relación con la razón de verosimilitud. Véase Cuadras (2015).

Ejercicio 4.7 *Se pide:*

1. *Prueba la relación entre θ^2 y las correlaciones canónicas en (4.17).*
2. *Suponiendo $m = 3$ y $r_1^2 = 0.94$, $r_2^2 = 0.88, r_3^2 = 0.01$, calcula η^2 y θ^2.*
3. *Supongamos que en (4.13) y (4.15) es $q = 1$, $p > 1$, es decir, hay $m = 1$ correlaciones canónicas. Demuestra que θ es el coeficiente de correlación múltiple.*
4. *Calcula θ^2 para relacionar las dos matrices de distamcias del Ejercicio 3.9.*

Capítulo 5

Comparación DB de poblaciones y distintividad

5.1 Comparando conjuntos de datos mixtos

Sean dos poblaciones Ω_1, Ω_2 y $\boldsymbol{\xi}$ un vector aleatorio mixto con valores en el conjunto $\mathbf{S}_\xi$. Supongamos que se dispone de dos muestras de tamaños n_1 y n_2, provenientes de Ω_1 y Ω_2. Sobre la base de $\boldsymbol{\xi}$, ambas muestras proporcionan dos conjuntos de datos M_1,M_2. Sea δ una función de distancia en S_ξ, que nos brinda las distancias entre las observaciones originales, posiblemente diferentes según que los individuos provengan de Ω_1 o de Ω_2 y supongamos que se dispone de dos aplicaciones $\Psi_i : \Omega_i \to L$, $i = 1, 2$, siendo L un espacio Euclídeo (o de Hilbert separable, véanse (3.10) y las Secciones 3.2 y 3.3). Estas aplicaciones proporcionan unas coordenadas que satisface (3.11). Ψ_i depende de $\Omega_i, i = 1, 2$.

Nos interesa comparar, mediante un test estadístico, las distribuciones de $\boldsymbol{\xi}$ en Ω_1 y en Ω_2, sobre la base de M_1,M_2 y la distancia δ. Esta comparación se puede formular estableciendo la hipótesis nula

$$H_0 : E[\Psi_1(\xi)] = E[\Psi_2(\xi)]. \tag{5.1}$$

5.2 Comparación mediante coordenadas principales

Utilizando δ y calculando las distancias sobre la base de los datos M_1,M_2, se obtienen dos matrices $n_i \times n_i$ de interdistancias $\boldsymbol{\Delta}_{11}, \boldsymbol{\Delta}_{22}$, $i = 1, 2$, así como la matriz $n_1 \times n_2$ de interdistancias $\boldsymbol{\Delta}_{12}$. Si consideramos la supermatriz de

distancias

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

y obtenemos la matriz $\mathbf{Z}$ de coordenadas principales, particionada en

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix},$$

donde $\mathbf{X}(n_1 \times p)$ se relaciona con $\mathbf{\Delta}_{11}$, $\mathbf{Y}(n_2 \times p)$ se relaciona con $\mathbf{\Delta}_{22}$. Las distancias entre filas de $\mathbf{X}$ y filas de $\mathbf{Y}$, reproducen $\mathbf{\Delta}_{12}$. Se puede suponer que la dimensión efectiva p ha sido obtenida, empleando algún criterio de reducción de la dimensión.

$\mathbf{X}$ e $\mathbf{Y}$ pueden ser interpretadas como dos matrices de datos "independientes", relacionadas con M_1,M_2, con vectores de medias $\overline{\mathbf{x}}, \overline{\mathbf{y}}$ y matrices de covarianza $\mathbf{S}_x, \mathbf{S}_y$.

Sea

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

la supermatriz de producto interno relacionada con $\mathbf{\Delta}$. Entonces $\mathbf{B} = \mathbf{Z}\mathbf{Z}'$ y $\mathbf{X}, \mathbf{Y}$ cumplen las restricciones $n_1\overline{\mathbf{x}}+n_2\overline{\mathbf{y}} = 0$, siendo $\overline{\mathbf{x}}, \overline{\mathbf{y}}$ los vectores de medias, y además $\mathbf{X}\mathbf{Y}' = \mathbf{B}_{12}$.

Un test para decidir H_0 con nivel de significación α, puede basarse en el estadístico

$$T_0^2 = (\overline{\mathbf{x}} - \overline{\mathbf{y}})'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}),$$

siendo

$$\mathbf{S} = (n_1 S_1 + n_2 S_2)/(n_1 + n_2).$$

En caso de normalidad multivariante, la distribución de T_0^2 está relacionada con la distribución T^2 de Hotelling. En general, la distribución de T_0^2 es desconocida, pero podemos emplear técnicas de remuestro.

Un test de permutaciones, relativamente simple, consiste en obtener las $N = (n_1 + n_2)!/(n_1!n_2!)$ permutaciones de las filas de $\mathbf{Z}$, obteniendo los correspondientes valores T^2 y ordenándolos de menor a mayor

$$T_1^2 \leq \cdots \leq T_k^2 \leq \cdots \leq T_N^2,$$

donde k se relaciona con el nivel de significancia α mediante

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

Decidimos entonces:

a) Aceptar H_0 si $T_0^2 \leq T_k^2$. b) Rechazar H_0 si $T_0^2 > T_k^2$.

La comparación mediante distancias de dos o más poblaciones ha sido estudiada por Cuadras y Fortiana (2004), Cuadras (2008), y otros autores.

Ejercicio 5.1 *Se pide:*

1. *Determina la distribución de* T_0^2 *cuando* Ω_i *es* $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ $i = 1, 2$ *y* $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

2. *Demuestra que* $n_1\overline{\mathbf{x}}+n_2\overline{\mathbf{y}} = 0$.

3. *Sugiere alguna alternativa al estadístico* T_0^2 *para contrastar la hipótesis* H_0.

5.3 Distintividad

Se entiende por **distintividad** ("typicality"), el problema del análisis discriminante que consiste en averiguar si un individuo pertenece a una población de entre dos poblaciones conocidas (o a una combinación lineal convexa de ambas), o en cambio pertenece a una tercera población 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, por el contario, debe considerarse una nueva especie.

5.4 Distintividad suponiendo 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, es decir, Ω_i es $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Se pretende decidir si $\mathbf{X}$ está relacionada con Ω_1, Ω_2 (dos poblaciones conocidas) o con Ω_3 (una nueva población desconocida).

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

$$\begin{array}{lll} H_0 : \mathbf{x} & \text{proviene de} & N_p(\alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \quad 0 \leq \alpha \leq 1, \\ H_1 : \mathbf{x} & \text{proviene de} & N_p(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}). \end{array} \tag{5.2}$$

Rechazar H_0 significa que $\mathbf{x}$ proviene de una nueva población Ω_3 no relacionada con Ω_1, Ω_2, las dos poblaciones de las que se posee información controlada. Es decir, $\mathbf{x}$ es un "outlier".

Bajo $H_0(\alpha = 1), H_0'(\alpha = 0), H_0''(0 \leq \alpha \leq 1)$, se cumple que:

$$U_1(\mathbf{x}) = \frac{[(\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)]^2}{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)} \sim \chi_1^2, \tag{5.3}$$

$$U_2(\mathbf{x}) = \frac{[(\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)]^2}{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)} \sim \chi_1^2, \tag{5.4}$$

$$W_1(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) - U_1(\mathbf{x}) \sim \chi_{p-1}^2, \tag{5.5}$$

donde "$\sim$" indica que el estadístico sigue la distribución ji-cuadrado.

Si $W_1(\mathbf{x})$ es significativa, $\mathbf{x}$ podría provenir de la población diferente Ω_3. En otro caso, $\mathbf{x}$ proviene de la combinación lineal convexa

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

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

5.5 Distintividad: planteamiento DB

En primer lugar, obsérvese que:

$$2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) =$$

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + (\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \geq 0.$$

Definiendo

$$P_1(\mathbf{x}) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1),$$

se tiene entonces

$$P_1(\mathbf{x}) = \tfrac{1}{2}\left\lfloor \delta_M^2(\mu_1, \mu_2) + \delta_M^2(\mathbf{x}, \mu_1) - \delta_M^2(\mathbf{x}, \mu_2) \right\rfloor \geq 0, \tag{5.6}$$

donde δ_M^2 es la distancia de Mahalanobis.

La versión DB para la distancia δ_M^2 entre Ω_1 y Ω_2 es (véase (3.34)):

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

siendo S el soporte de la densidad.

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

$$P_1(\mathbf{x}) = \tfrac{1}{2}\left[\Delta^2(\Omega_1, \Omega_2) + \phi_1^2(\mathbf{x}) - \phi_2^2(\mathbf{x})\right], \tag{5.7}$$

donde ϕ_1^2, ϕ_2^2 son las funciones de proximidad. De forma similar:

$$P_2(\mathbf{x}) = \tfrac{1}{2}\left[\Delta^2(\Omega_1, \Omega_2) + \phi_2^2(\mathbf{x}) - \phi_1^2(\mathbf{x})\right]. \tag{5.8}$$

Entonces las versiones DB de (5.3-5.5) para probar

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

respectivamente, están basadas en los estadísticos

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

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

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

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 5.2). Si $W_1(\mathbf{x})$ es significativa, entonces $\mathbf{x}$ podría provenir de una población distinta Ω_3.

Este planteamiento DB de la distintividad ha sido estudiado por Cuadras y Fortiana (1998, 2000). Véase también Rao (1973, p.579). El caso de varias poblaciones ha sido estudiado por Bar-Hen (2001), Irigoien y Arenas (2008).

Ejercicio 5.2 *Se pide:*

1. *A partir de los datos mixtos sobre cáncer del Ejercicio 3.10, dibuja el histograma de* $U_1(\mathbf{x})$, $U_2(\mathbf{x})$, $W_1(\mathbf{x})$, *empleando solamente la muestra de los* $n_1 = 78$ *casos benignos.*
2. *Demuestra (5.3)-(5.5).*
3. *Define* $W_2(\mathbf{x}) = \phi_2^2(\mathbf{x}) - U_2(\mathbf{x})$. *Demuestra que* $W_1(\mathbf{x}) = W_2(\mathbf{x})$.

5.6 Distintividad mediante 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 de proximidades

$$r_i(\mathbf{x}) = \frac{\phi_i^2(\mathbf{x})}{\phi_1^2(\mathbf{x}) + \phi_2^2(\mathbf{x})}, \qquad i = 1, 2, \tag{5.9}$$

entonces se cumple que $r_i(\mathbf{x}) \cong 0$, si $\mathbf{x}$ 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 la hipótesis nula $H_0 : \Omega_1 = \Omega_2$ se cumple que

$$E[r_i(\mathbf{x})] = \tfrac{1}{2}.$$

Considerando la hipótesis nula

$$H_0^{(1)} : \mathbf{x} \quad \text{proviene de} \quad \Omega_1.$$

Supongamos que $r_i(\mathbf{x})$ sigue una distribución uniforme en el intervalo $(0, 1)$.

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 nivel de significación α estipulado.

Similarmente podemos considerar

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

y aceptar la hipótesis $H_0^{(2)}$ si

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

Si ambas hipótesis $H_0^{(1)}, H_0^{(2)}$ son rechazadas, entonces se puede afirmar que $\mathbf{x}$ procede de una población diferente Ω_3.

Un argumento parecido valdría comsiderando otra distribución de $r_1(\mathbf{x})$, como por ejemplo la distribución beta.

Ejercicio 5.3 *Se pide:*

1. *A partir de los datos mixtos sobre cáncer del Ejercicio 3.10, dibuja un histograma de $r_1(\mathbf{x})$ y de $r_2(\mathbf{x})$ 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. *Utilizando también funciones de proximidad, propóngase algún criterio alternativo para resolver la distintividad.*

5.7 Complementos

Las distancias estadísticas se han aplicado en otras áreas de la estadística y la probabilidad. En Anderson (2001, 2006), Cuadras y Fortiana (2004), Cuadras (2008), se plantea la comparación de poblaciones y el análisis multivariante de la varianza, mediante distancias y test de permutaciones.

En Cuadras y Fortianan (1993b, 1995, 1997a) las distancias se aplican para obtener desarrollos ortogonales de variables aleatorias, con aplicaciones a los test de bondad de ajuste, y para formular la asociación estocástica en teoría de cópulas (didtribuciones bivariantes con marginales uniformes). Véase también Cuadras (2014).

En Cuadras, Atkinson y Fortiana (1997) se propone un método para generar densidades a partir de distancias y funciones de proximidad. La construcción de estas funciones puede compararse con la transformación que da lugar a la función característica. Para otras transformaciones basadas en distancias, véase Székely *et al.* (2007).

Los siguientes tres ejercicios son de recapitulación.

Ejercicio 5.4 *Para el modelo multivariante DB expresado como* $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, *donde* $\mathbf{X}, \mathbf{Y}$ *se han obtenido mediante análisis de coordenadas principales (vease (4.16)), hallar la predicción* $\mathbf{y} = (y_1, \ldots, y_q)$ *de un nuevo individuo conocidas las distamcias a los otros* n.

Ejercicio 5.5 *Sea* $f(x)$ *la función de densidad de una variable aleatoria* X *con soporte* S. *Mediante una distancia* δ *se calcula la variabilidad geométrica* V *y se construye la función de proximidad* $\phi^2(x)$. *Prueba que, transformando* δ *si es necesario, se puede suponer que* $f_\delta(x) = \exp[-\phi^2(x)]$. $x \in S$, *es una función de densidad. Prueba que* $V \geq H(f)$, *siendo* $H(f) = -\int_S f(x) \ln f(x) dx$ *la entropía de Shannon.*

Ejercicio 5.6 *En el método DB para resolver la distintividad, dibuja un triángulo cuyos vértices representen las dos poblaaciones* Ω_1, Ω_2 *y la observación* ω *a clasificar. Interpreta geométricamente* U_1, U_2 *y* W_1.

Referencias

[1] Anderson, M. J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.

[2] Anderson, M. J. (2006) Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, **62**, 245-253.

[3] Bar-Hen, A., Daudin, J.-J. (1997) A test of a special case of typicality in linear discriminant analysis. *Biometrics*, **53**, 39-48.

[4] Bar-Hen, A. (2001) Preliminary tests in linear discriminant analysis. Statistica, 61 (4), 585–594.

[5] Bartkowiak, A., Jakimiec, M. (1994) Distance-based regression in prediction of solar flare activity. *Qüestiió*, 18. 7-38.

[6] Boj, E., Delicado, P., Fortiana, J. (2010) Distance-based local linear regression for functional predictors. *Computational Statistics and Data Analysis*, 54, 429-437.

[7] Boj, E., Costa, T., Fortiana, J. (2015) Assessing the importance of risk factors in distance-based generalized linear models.*Methodology and Computing in Applied Probability,* 17, 951-962.

[8] Cailliez, F. (1983) The analytical solution of the additive constant problem. *Psychometrika,* 48, 305-308.

[9] Cox., T. F., Cox, M. A. (2001) *Multidimensional Scaling.* Chapman & Hall, London, 2a Ed.

[10] Cuadras, C. M. (1989a) Distancias estadisticas. *Estadistica Española*, 30 (119), 295-378.

[11] Cuadras, C. M. (1989b) 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.

[12] Cuadras, C. M. (1990) An eigenvector pattern arising in nonlinear regression. *Qüestiió*, 14, 89-95.

[13] Cuadras, C. M., Arenas, C. (1990) A distance based regression model for prediction with mixed data. *Communications in Statistics A. Theory and Methods,* 19, 2261–2279.

[14] Cuadras, C. M. (1991) *Métodos de Análisis Multivariante*. 2da edic., PPU, Barcelona.

[15] Cuadras, C. M. (1991) A distance-based approach for discriminant analysis and its properties. *Math. Preprint Series* 90, Univ. of Barcelona.

[16] Cuadras, C. M. (1992a) Some examples of distance based discrimination. *Biometrical Letters*, 29,1-18.

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

[18] Cuadras, C. M. (2008) Distance-based multisample tests for multivariate data. In: *Advances in Mathematical and Statistical Modeling*, (Arnold, B. C., Balakrishnan, N., Sarabia, J. M., Mínguez, R. Eds.), Birkhauser, Boston, pp. 61-71, 2008.

[19] Cuadras, C. M. (2014) *Nuevos étodos de Análisis Multivariante*. CMC Editions, Barcelona.

[20] Cuadras, C. M. (2015) El llegat de Galton, Pearson, Fréchet i altres: com mesurar i interpretar l'associació estadística. *Butlletí de la Societat Catalana de Matemàatiques*, 30 (1), 5-56.

[21] Cuadras, C. M. , Fortiana, J. (1993a) Aplicaciones de las distancias en estadística. *Qüestió*, 17(1), 39-74.

[22] Cuadras, C. M., Fortiana, J. (1993b) Continuos metric scaling and prediction. En: *Multivariate Analysis: Future Directions 2*. (C.M. Cuadras and C.R. Rao, eds.). Elsevier, Amsterdam, pp. 47-66.

[23] Cuadras, C. M. (1993) Interpreting an inequality in mutiple regression. *Amer. Statician*, 47, 256-258.

[24] Cuadras, C. M. , Fortiana, J. (1995) A continuos metric scaling solution for a random variable. *Journal of Multivariate Analysis*, 52 (1), 1-14.

[25] Cuadras, C. M., C. Arenas, Fortiana, J. (1996) Some computational aspects of a distance–based model for prediction. *Communications in Statistics. Simulation and Computation, 25(3)*, 593–609.

[26] Cuadras, C. M., Fortiana, J., Oliva, F. (1997) The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification*, 14, 117-136.

[27] Cuadras, C. M., Atkinson, R. A. Fortiana, J. (1997) Probability densities from distances and discriminant analysis. *Statistics and Probability Letters*, 33, 405-411.

[28] C uadras, C. M., Fortiana, J., Oliva, F. (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.

[29] Cuadras, C. M., Fortiana, J.(1997a) Continuous scaling on a bivariate copula. In: Viktor Benes and Josef Stepan, (Eds). *Distributions with given marginals and moment problems*, pp 137–142. Kluwer Academic Pub., Dordrecht.

[30] Cuadras, C. M., Fortiana, J. (1997b) 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.

[31] Cuadras, C.M., Fortiana, J. (1998) Typicality in discriminant analysis with mixed variables. *Data Science, Classification and Related Topics*, IFCS-98, Rome, pp. 82-85,

[32] Cuadras, C. M., Fortiana, J. (2000) The Importance of Geometry in Multivariate Analysis and some Applications. In: *Statistics for the 21st Century*, Ch. 4, C.R. Rao, G. Szekely, eds., Marcel Dekker, pp. 93-108.

[33] Cuadras, C. M., Fortiana, J. (2004) Distance-based multivariate two sample tests. In: *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life* (M. S.

Nikulin, N. Balakrishnan, M. Mesbah, N. Limnios,Eds.), Birkhauser, Boston, 273-290, 2004.

[34] Cuadras, C. M., Valero, S., Cuadras, D., Salembier, P., Chanussot, J. (2012) Distance-based measures of association with applications in relating hyperspectral images. *Comm. Stat., Theor.- Meth.*, **41**, 2342–2355.

[35] Cuadras, C. M. (2014) Nonlinear principal and canonical directions from continuous extensions of multidimensional scaling. *Open Journal of Statistics*, 4 (2), 132-149.

[36] Galindo Villardón, M. P. (1988) Una alternativa de representación simultánea: HJ-Biplot. *Qüestiió*, 10 (1), 13-23.

[37] Gabriel, K. R. (1971) The biplot graphic display of matrices with applications to principal component analysis. *Biometrika*, 58, 453-467.

[38] Gower, J. C. (1968) Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55, 582-585.

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

[40] Gower, J. C., Harding, S.A. (1988) Nonlinear biplots. *Biometrika*, 75, 445-455.

[41] Greenacre, M. J. (1984) *Theory and Applications of Correspondence Analysis.* Academic Press, London.

[42] Greenacre, M. J. (2016) *Correspondence Analysis in Practice.* 3rd Edit., Chapman&Hall, London.

[43] Hoerl, A. E., Kennard, R. W. (1970) Ridge regression. Biased regression for npn-ortoghonal problems. Technometrics, *Technometrics*, 12, 55-67.

[44] Irigoien, I., Arenas, C. (2008) INCA: New statistic for estimating the number of clusters and identifying atypical units. *Statistics in Medicine*, **27**, 2948-2973.

[45] Krzanowski, W. J. (1980) Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics, 36,* 493-499.

[46] Krzanowski, W. J. (1994) Ordination in the presence of group structure, for general multivariate data. *Journal of Classification*, 11, 195-207.

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

[48] Miñarro, A., Oller, J.M. (1992) Some remarks on the individuals-score distance and its applications to statistical infrence. *Qüestiió*, 16, 43-57.

[49] 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.

[50] Rao, C. R. (1973) *Linear Statistical Inference and Its Applications.* New York: Wiley.

[51] Robert, D„ Amat, L., Carbo-Dorca, R. (1999) Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: Prediction of the corticosteroid-binding globulin binding affinity for a steroid family, *J. Chem. Inf. Comput. Sci.*, 39 (2), 333–344.

[52] Székely, G. J., Rizzo, M. L., Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 36, 2769-2794.