NUEVOS MÉTODOS DE ANÁLISIS MULTIVARIANTE

Carles M. Cuadras

February 2, 2007

CMC Editions
Manacor 30
08023 Barcelona, Spain

Índice

PRÓLOGO

El Análisis Multivariante es un conjunto de métodos estadísticos y matemáticos, destinados a describir e interpretar los datos que provienen de la observación de varias variables estadísticas, estudiadas conjuntamente.

Este libro es una presentación convencional de los principales modelos y métodos del Análisis Multivariante, con referencias a algunas contribuciones recientes.

La exposición mantiene un cierto rigor matemático, compensado con una clara orientación aplicada. Todos los métodos se ilustran con ejemplos, que justifican su aplicabilidad. Para examinar los datos y ver más ejemplos consúltese la página web

www.ub.edu/stat/cuadras/cuad.html

Esta obra tiene como precedentes la monografia "Métodos de Análisis Factorial" (Pub. no. 7, Laboratorio de Cálculo, Universidad de Barcelona, 1974), y el libro "Métodos de Análisis Multivariante" (EUNIBAR, 1981; PPU, 1991; EUB, 1996, Barcelona).

Cómo citar este libro:

C. M. Cuadras
Nuevos Métodos de Análisis Multivariante
CMC Editions
Barcelona, 2007

Capítulo 1

DATOS MULTIVARIANTES

1.1 Introducción

El análisis multivariante (AM) es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resulten de observar un número $p > 1$ de variables estadísticas sobre una muestra de n individuos. Las variables observables son homogéneas y correlacionadas, sin que alguna predomine sobre las demás. La información estadística en AM es de carácter multidimensional, por lo tanto la geometría, el cálculo matricial y las distribuciones multivariantes juegan un papel fundamental.

La información multivariante es una matriz de datos, pero a menudo, en AM la información de entrada consiste en matrices de distancias o similaridades, que miden el grado de discrepancia entre los individuos. Comenzaremos con las técnicas que se basan en matrices de datos.

1.2 Matrices de datos

Supongamos n individuos $\omega_1, \ldots, \omega_n$ y p variables $X_1, \ldots, X_p$. Sea $x_{ij} = X_j(\omega_i)$ la observación de la variable X_j sobre el individuo ω_i. La matriz de

datos multivariantes es

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Las filas de $\mathbf{X}$ se identifican con los individuos y las columnas de $\mathbf{X}$ con las variables. Indicaremos:

1. $\mathbf{x}_i$ la fila i-ésima de $\mathbf{X}$.
2. X_j la columna j-ésima de $\mathbf{X}$.
3. $\overline{\mathbf{x}} = (\overline{x}_1, \ldots, \overline{x}_j, \ldots, \overline{x}_p)'$ el vector (fila) de las medias de las variables, siendo
$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$
4. La matriz simétrica $p \times p$ de covarianzas muestrales
$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ & \cdots & \cdots & \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix},$$
siendo
$$s_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ij'} - \overline{x}_{j'})$$
la covarianza entre las variables j, j'. Naturalmente, $\overline{\mathbf{x}}$ y $\mathbf{S}$ son medidas multivariantes de tendencia central y dispersión.

1.3 La matriz de centrado

Si $\mathbf{1} = (1, \ldots, 1)'$ es el vector columna de unos de orden $n \times 1$, y $\mathbf{J} = \mathbf{1}\mathbf{1}'$ es la matriz $n \times n$ de unos, ciertas características multivariantes se expresan mejor a partir de la matriz de centrado $\mathbf{H}$, definida como

$$\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{J}$$

Propiedades:

- $\mathbf{H}' = \mathbf{H}$.
- $\mathbf{H}^2 = \mathbf{H}$.
- $\mathbf{H1} = \mathbf{1}'\mathbf{H} = \mathbf{0}$.
- rang($\mathbf{H}$) $=n-1$.
- Los valores propios de $\mathbf{H}$ son 0 ó 1.
- $\overline{\mathbf{X}} = \mathbf{HX}$ es la matriz de datos centrados (las columnnas de $\overline{\mathbf{X}}$ suman 0).

1.4 Medias, covarianzas y correlaciones

El vector de medias, la matriz de covarianzas, etc., tienen expresiones matriciales simples.

1. $\overline{\mathbf{x}}' = \frac{1}{n}\mathbf{1}'\mathbf{X}$.
2. Matriz de datos centrados:
$$\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}' = \mathbf{HX}.$$
3. Matriz de covarianzas:
$$\mathbf{S} = \frac{1}{n}\overline{\mathbf{X}}'\overline{\mathbf{X}} = \frac{1}{n}\mathbf{X}'\mathbf{HX}.$$

Además de la matriz de covarianzas interesa también la matriz de correlaciones

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ & \cdots & \cdots & \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix},$$

donde $r_{ij} = \text{cor}(X_i, X_j)$ es el coeficiente de correlación (muestral) entre las variables X_i, X_j, que verifica:

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}, \qquad \mathbf{S} = \mathbf{DRD}, \tag{1.1}$$

siendo $\mathbf{D}$ la matriz diagonal con las desviaciones típicas de las variables.

1.5 Variables compuestas

Algunos métodos de AM consisten en obtener e interpretar combinaciones lineales adecuadas de las variables observables. Una variable compuesta Y es una combinación lineal de las variables observables con coeficientes $\mathbf{a} = (a_1, \ldots, a_p)'$

$$Y = a_1 X_1 + \ldots + a_p X_p.$$

Si $\mathbf{X} = [X_1, \ldots, X_p]$ es la matriz de datos, también podemos escribir

$$Y = \mathbf{Xa}.$$

Si $Z = b_1 X_1 + \ldots + b_p X_p = \mathbf{Xb}$ es otra variable compuesta, se verifica:

1. $\overline{Y} = \overline{\mathbf{x}}'\mathbf{a}$, $\overline{Z} = \overline{\mathbf{x}}'\mathbf{b}$.
2. $\text{var}(Y) = \mathbf{a}'\mathbf{Sa}$, $\text{var}(Z) = \mathbf{b}'\mathbf{Sb}$.
3. $\text{cov}(Y, Z) = \mathbf{a}'\mathbf{Sb}$.

Ciertas variables compuestas reciben diferentes nombres según la técnica multivariante: componentes principales, variables canónicas, funciones discriminantes, etc. Uno de los objetivos del Análisis Multivariante es encontrar variables compuestas adecuadas que expliquen aspectos relevantes de los datos.

1.6 Transformaciones lineales

Sea $\mathbf{T}$ una matriz $p \times q$. Una transformación lineal de la matriz de datos es

$$\mathbf{Y} = \mathbf{XT}$$

Las columnas $Y_1, \ldots, Y_q$ de $\mathbf{Y}$ son las variables transformadas.

Propiedades:

1. $\overline{\mathbf{y}}' = \overline{\mathbf{x}}'\mathbf{T}$, donde $\overline{\mathbf{y}}$ es el vector de medias de $\mathbf{Y}$.
2. $\mathbf{S}_Y = \mathbf{T}'\mathbf{ST}$, donde $\mathbf{S}_Y$ es la matriz de covarianzas de $\mathbf{Y}$.

Demost.:
$\overline{\mathbf{y}}' = \frac{1}{n}\mathbf{1}'\mathbf{Y} = \frac{1}{n}\mathbf{1}'\mathbf{XT} = \overline{\mathbf{x}}'\mathbf{T}$. $\mathbf{S}_Y = \frac{1}{n}\mathbf{Y}'\mathbf{HY} = \frac{1}{n}\mathbf{T}'\mathbf{X}'\mathbf{HXT} = \mathbf{T}'\mathbf{ST}$.

1.7 Teorema de la dimensión

La matriz de covarianzas $\mathbf{S}$ es (semi)definida positiva, puesto que:

$$\mathbf{a}'\mathbf{S}\mathbf{a} = \frac{1}{n}\mathbf{a}'\mathbf{X}'\mathbf{H}\mathbf{X}\mathbf{a} = \frac{1}{n}\mathbf{a}'\mathbf{X}'\mathbf{H}\mathbf{H}\mathbf{X}\mathbf{a} = \mathbf{b}'\mathbf{b} \geq 0,$$

siendo $\mathbf{b} = n^{-1/2}\mathbf{H}\mathbf{X}\mathbf{a}$.

El rango $r = \text{rang}(\mathbf{S})$ determina la dimensión del espacio vectorial generado por las variables observables, es decir, el número de variables linealmente independientes es igual al rango de $\mathbf{S}$.

Teorema 1.7.1 *Si $r = rang(\mathbf{S}) \leq p$ hay r variables linealmente independientes y las otras $p - r$ son combinación lineal de estas r variables.*

Demost.: Podemos ordenar las p variables de manera que la matriz de covarianzas de $X_1, \ldots, X_r$ sea no singular

$$\begin{pmatrix} s_{11} & \cdots & s_{1r} \\ \vdots & \ddots & \vdots \\ s_{r1} & \cdots & s_{rr} \end{pmatrix}$$
$$\begin{matrix} s_{j1} & \cdots & s_{jr} \end{matrix}$$

Sea $X_j, j > r$. Las covarianzas entre X_j y $X_1, \ldots, X_r$ verifican:

$$s_{jj} = \sum_{i=1}^{r} a_i s_{ji}, \quad s_{ji} = \sum_{i'=1}^{r} a_{i'} s_{ii'}.$$

Entonces

$$\begin{aligned} var(X_j - \textstyle\sum_{i=1}^{r} a_i X_i) &= s_{jj} + \textstyle\sum_{i,i'=1}^{r} a_i a_{i'} s_{ii'} - 2\sum_{i=1}^{r} a_i s_{ji} \\ &= \textstyle\sum_{i=1}^{r} a_i s_{ji} + \sum_{i=1}^{r} a_i (\sum_{i'=1}^{r} a_{i'} s_{ii'}) - 2\sum_{i=1}^{r} a_i s_{ji} \\ &= \textstyle\sum_{i=1}^{r} a_i s_{ji} + \sum_{i=1}^{r} a_i s_{ji} - 2\sum_{i=1}^{r} a_i s_{ji} \\ &= 0. \end{aligned}$$

Por lo tanto

$$X_j - \sum_{i=1}^{r} a_i X_i = c \Longrightarrow X_j = c + \sum_{i=1}^{r} a_i X_i$$

donde c es una constante.

Corol.lari 1.7.2 *Si todas las variables tienen varianza positiva (es decir, ninguna se reduce a una constante) y* $r = rang(\mathbf{R}) \leq p$, *hay* r *variables linealmente independientes y las otras* $p - r$ *son combinación lineal de estas* r *variables.*

Demost.: De (1.1) deducimos que $r = \text{rang}(\mathbf{R}) = \text{rang}(\mathbf{S})$.

1.8 Medidas globales de variabilidad y dependencia

Una medida de la variabilidad global de las p variables debe ser función de la matriz de covarianzas $\mathbf{S}$. Sean $\lambda_1, \ldots, \lambda_p$ los valores propios de $\mathbf{S}$. Las siguientes medidas tienen especial interés en AM.

a) Varianza generalizada:

$$|\mathbf{S}| = \lambda_1 \times \cdots \times \lambda_p.$$

b) Variación total:

$$\text{tr}(\mathbf{S}) = \lambda_1 + \cdots + \lambda_p$$

Una medida de dependencia global debe ser función de la matriz de correlaciones $\mathbf{R}$. Un coeficiente de dependencia es

$$\eta^2 = 1 - |\mathbf{R}|,$$

que verifica:

1. $0 \leq \eta^2 \leq 1$.
2. $\eta^2 = 0$ si y sólo si las p variables estan incorrelacionadas.
3. $\eta^2 = 1$ si y sólo si hay relaciones lineales entre las variables.

Demost.:
1. Sean $\lambda_1, \ldots, \lambda_p$ los valores propios de $\mathbf{R}$. Si g y a son las medias geométrica y aritmética de p números positivos, se verifica $g \leq a$. Entonces, de $\text{tr}(\mathbf{R}) = p$

$$(|\mathbf{R}|)^{1/p} = (\lambda_1 \times \cdots \times \lambda_p)^{1/p} \leq (\lambda_1 + \cdots + \lambda_p)/p = 1$$

y por lo tanto $0 \leq \det(\mathbf{R}) \leq 1$.

2. $\mathbf{R} = \mathbf{I}$ (matriz identidad) si y sólo si las p variables están incorrelacionadas y entonces $1 - |\mathbf{I}| = 0$.

3. Si $\eta^2 = 1$, es decir, $|\mathbf{R}| = 0$, entonces $\text{rang}(\mathbf{R}) < p$ y por lo tanto hay combinaciones lineales entre las variables (Teorema 1.7.1).

1.9 Distancias

Algunos métodos de AM están basados en criterios geométricos y en la noción de distancia entre individuos y entre poblaciones. Si

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

es una matriz de datos, con matriz de covarianzas $\mathbf{S}$, las tres definiciones más importantes de distancia entre las filas $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip}), \mathbf{x}_j' = (x_{j1}, \ldots, x_{jp})$ de $\mathbf{X}$ son:

1. Distancia Euclídea:

$$d_E(i,j) = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2}. \tag{1.2}$$

2. Distancia de K. Pearson

$$d_P(i,j) = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2 / s_{hh}}, \tag{1.3}$$

donde s_{hh} es la covarianza de la variable X_h.

3. Distancia de Mahalanobis:

$$d_M(i,j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}. \tag{1.4}$$

Observaciones

Un cambio de escala de una variable X_j es una transformación $Y_j = \alpha X_j$, donde α es una constante. La distancia d_M es muy adecuada en AM debido a que verifica:

a) d_E supone implícitamente que las variables son incorrelacionadas y no es invariante por cambios de escala.

b) d_P también supone que las variables son incorrelacionades pero es invariante por cambios de escala.

c) d_M tiene en cuenta las correlaciones entre las variables y es invariante por transformaciones lineales no singulares de las variables, en particular cambios de escala.

Las distancias d_E y d_P son casos particulares de d_M cuando la matriz de covarianzas es la identidad $\mathbf{I}_p$ y diag($\mathbf{S}$), respectivamente. En efecto:

$$\begin{aligned} d_E(i,j)^2 &= (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j), \\ d_P(i,j)^2 &= (\mathbf{x}_i - \mathbf{x}_j)'[\text{diag}(\mathbf{S})]^{-1}(\mathbf{x}_i - \mathbf{x}_j). \end{aligned}$$

La distancia de Mahalanobis (al cuadrado) puede tener otras versiones:

1. Distancia de una observación $\mathbf{x}_i$ al vector de medias $\overline{\mathbf{x}}$ de $\mathbf{X}$:

$$(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$$

2. Distancia entre dos poblaciones representadas por dos matrices de datos $\mathbf{X}_{n_1 \times p}$, $\mathbf{Y}_{n_2 \times p}$:

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

donde $\overline{\mathbf{x}}, \overline{\mathbf{y}}$ son los vectores de medias y

$$\mathbf{S} = (n_1\mathbf{S}_1 + n_2\mathbf{S}_2)/(n_1 + n_2)$$

es la media ponderada de las correspondientes matrices de covarianzas.

N	E	S	W	N	E	S	W
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

Tabla 1.1: Depósitos de corcho (centigramos) de 28 alcornoques en las cuatro direcciones cardinales.

1.10 Un ejemplo

Exemple 1.10.1

La Tabla 1.1 contiene los datos de $n = 28$ alcornoques y $p = 4$ variables, que miden los depósitos de corcho (en centigramos) en cada uno de los cuatro puntos cardinales: N, E, S, W.

Medias, covarianzas y correlaciones

Vector de medias:

$$\overline{\mathbf{x}}'=(50.536, 46.179, 49.679, 45.179)$$

Matriz de covarianzas:

$$\mathbf{S} = \begin{pmatrix} 280.03 & 215.76 & 278.13 & 218.19 \\ & 212.07 & 220.88 & 165.25 \\ & & 337.50 & 250.27 \\ & & & 217.93 \end{pmatrix}$$

Matriz de correlaciones:

$$\mathbf{R} = \begin{pmatrix} 1 & 0.885 & 0.905 & 0.883 \\ & 1 & 0.826 & 0.769 \\ & & 1 & 0.923 \\ & & & 1 \end{pmatrix}$$

Variables compuestas

Las siguientes variables compuestas explican diferentes aspectos de la variabilidad de los datos:

			Media	Variancia:
Contraste eje N-S con eje E-W:	$Y_1 =$	$N + S - E - W$	8.857	124.1
Contraste N-S:	$Y_2 =$	$N - S$	0.857	61.27
Contraste E-W:	$Y_3 =$	$E - W$	1.000	99.5

Variables normalizadas

Una variable compuesta está normalizada si la suma de cuadrados de sus coeficientes es 1. La normalización evita que la varianza tome un valor arbitrario. La normalización de Y_1, Y_2, Y_3 dará:

		Media	Variancia:
$Z_1 =$	$(N + S - E - W)/2$	4.428	31.03
$Z_2 =$	$(N - S)/\sqrt{2}$	0.606	30.63
$Z_3 =$	$(E - W)/\sqrt{2}$	0.707	49.75

Interpretación

La normalización de las variables consigue que estas tengan varianzas más homogéneas. La principal dirección de variabilidad aparece al hacer la comparación del eje N-S con el eje E-W.

Visualización de datos

En los capítulos siguientes veremos métodos y técnicas de visualitzación de datos multivariantes. Como norma general es conveniente, antes de realizar el análisis, examinar y revisar los datos. La Figura 1.1 contiene un gráfico que permite visualizar la distribución de las 4 variables de la Tabla 1.1 y las relaciones lineales, o regresión lineal, entre cada par de variables.

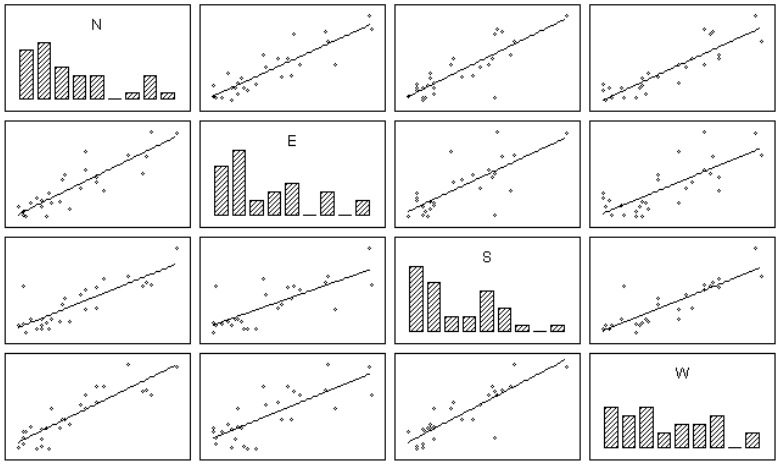

Figura 1.1: Distribución de las variables N, E, S, W y relaciones entre cada par de variables de la Tabla 1.1.

Capítulo 2

NORMALIDAD MULTIVARIANTE

2.1 Introducción

Los datos en AM suelen provenir de una población caracterizada por una distribución multivariante. Sea $\mathbf{X} = (X_1, \ldots, X_p)$ un vector aleatorio con distribución absolutamente continua y función de densidad $f(x_1, \ldots, x_p)$. Es decir, f verifica:

1) $f(x_1, \ldots, x_p) \geq 0$, para todo $(x_1, \ldots, x_p) \in R^p$.

2) $\int_{R^p} f(x_1, \ldots, x_p) dx_1 \cdots dx_p = 1$.

Conocida $f(x_1, \ldots, x_p)$ podemos encontrar la función de densitad de cada variable marginal X_j mediante la integral

$$f_j(x_j) = \int f(x_1, \ldots, x_j, \ldots, x_p) dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_p.$$

Como en el caso de una matriz de datos, es importante el vector de medias

$$\mu = (E(X_1), \ldots, E(X_p))',$$

donde $E(X_j)$ es la esperanza de la variable marginal X_j, y la matriz de covarianzas $\Sigma = (\sigma_{ij})$, siendo $\sigma_{ij} = \text{cov}(X_i, X_j)$, $\sigma_{ii} = \text{var}(X_i)$. Teniendo en cuenta que los elementos de la matriz $(\mathbf{X}-\mu)(\mathbf{X}-\mu)'$, de orden $p \times p$, son $(X_i - \mu_i)(X_j - \mu_j)$ y que $\text{cov}(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j)$, la matriz de covarianzas $\Sigma = (\sigma_{ij})$ es

$$\Sigma = E((\mathbf{X}-\mu)(\mathbf{X}-\mu)').$$

En este capítulo introducimos y estudiamos la distribución normal multivariante y tres distribuciones relacionadas con las muestras multivariantes: Wishart, Hotelling y Wilks.

2.2 Distribución normal multivariante

2.2.1 Definición

Sea X una variable aleatoria con distribución $N(\mu, \sigma^2)$, es decir, con media μ y varianza σ^2. La función de densidad de X es:

$$f(x;\, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} = \frac{(\sigma^2)^{-1/2}}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)} \tag{2.1}$$

Evidentemente se verifica:

$$X = \mu + \sigma Y \quad \text{donde} \quad Y \sim N(0,1). \tag{2.2}$$

Vamos a introducir la distribución normal mutivariante $N_p(\mu, \Sigma)$ como una generalización de la normal univariante. Por una parte, (2.1) sugiere definir la densidad de $\mathbf{X} = (X_1, \ldots, X_p)' \sim N_p(\mu, \Sigma)$ según:

$$f(\mathbf{x}; \mu, \Sigma) = \frac{|\Sigma|^{-1/2}}{(\sqrt{2\pi})^p} e^{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)}, \tag{2.3}$$

siendo $\mathbf{x} = (x_1, \ldots, x_p)', \mu = (\mu_1, \ldots, \mu_n)'$ y $\Sigma = (\sigma_{ij})$ una matriz definida positiva, que como veremos, es la matriz de covarianzas. Por otra parte, (2.2) sugiere definir la distribución $\mathbf{X} = (X_1, \ldots, X_p)' \sim N_p(\mu, \Sigma)$ como una combinación lineal de p variables $Y_1, \ldots, Y_p$ independientes con distribución $N(0,1)$.

$$\begin{array}{rl} X_1 & = \mu_1 + a_{11}Y_1 + \ldots + a_{1p}Y_p \\ \vdots & \qquad\qquad \vdots \\ X_p & = \mu_p + a_{p1}Y_1 + \ldots + a_{pp}Y_p \end{array} \tag{2.4}$$

que podemos escribir como

$$\mathbf{X} = \mu + \mathbf{A}\mathbf{Y} \tag{2.5}$$

donde $\mathbf{A} = (a_{ij})$ es una matriz $p \times q$ que verifica $\mathbf{A}\mathbf{A}' = \Sigma$.

Proposició 2.2.1 *Las dos definiciones (2.3) y (2.4) son equivalentes.*

Demost.: Según la fórmula del cambio de variable

$$f_X(x_1,\ldots,x_p) = f_Y(y_1(x),\ldots,y_p(x))\left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|$$

siendo $y_i = y_i(x_1,\ldots,x_p)$, $i = 1,\ldots,p$, el cambio y $J = \left|\frac{\partial y}{\partial x}\right|$ el jacobiano del cambio. De (2.5) tenemos

$$\mathbf{y} = \mathbf{A}^{-1}(\mathbf{x}-\mu) \Rightarrow \left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right| = |\mathbf{A}^{-1}|$$

y como las variables Y_i son $N(0,1)$ independientes:

$$f_X(x_1,\ldots,x_p) = (1/\sqrt{2\pi})^p e^{-\frac{1}{2}\sum_{i=1}^p y_i^2}|\mathbf{A}^{-1}|. \tag{2.6}$$

Pero $\Sigma^{-1} = (\mathbf{A}^{-1})'(\mathbf{A}^{-1})$ y por lo tanto

$$\mathbf{y}'\mathbf{y} = (\mathbf{x}-\mu)'(\mathbf{A}^{-1})'(\mathbf{A}^{-1})(\mathbf{x}-\mu) = (\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu). \tag{2.7}$$

Substituyendo (2.7) en (2.6) y de $|\mathbf{A}^{-1}|^2 = |\Sigma|^{-1}$ obtenemos (2.3).

2.2.2 Propiedades

1. De (2.5) es inmediato que $E(\mathbf{X}) = \mu$ y que la matriz de covarianzas es

$$E((\mathbf{X}-\mu)(\mathbf{X}-\mu)') = E(\mathbf{A}\mathbf{Y}\mathbf{Y}'\mathbf{A}') = \mathbf{A}\mathbf{I}_p\mathbf{A}' = \Sigma.$$

2. La distribución de cada variable marginal X_i es normal univariante:

$$X_i \sim N(\mu_i, \sigma_{ii}), \quad i = 1,\ldots,p.$$

   Es consecuencia de la definición (2.4).

3. Toda combinación lineal de las variables $X_1,\ldots,X_p$

$$Z = b_0 + b_1X_1 + \cdots + b_pX_p$$

   es también normal univariante. En efecto, de (2.4) resulta que Z es combinación lineal de $N(0,1)$ independientes.

4. Si Σ =diag$(\sigma_{11}, \ldots, \sigma_{pp})$ es matriz diagonal, es decir, $\sigma_{ij} = 0, i \neq j$, entonces las variables $(X_1, \ldots, X_p)$ son estocásticamente independientes. En efecto, la función de densidad conjunta resulta igual al producto de las funciones de densidad marginales:

$$f(x_1, \ldots, x_p; \mu, \Sigma) = f(x_1; \mu_1, \sigma_{11}) \times \cdots \times f(x_p; \mu_p, \sigma_{pp})$$

5. La distribución de la forma cuadrática

$$U = (\mathbf{x} - \mu)\Sigma^{-1}(\mathbf{x} - \mu)'$$

es ji-cuadrado con p grados de libertad. En efecto, de (2.5) $U = \mathbf{YY}' = \sum_{i=1}^{p} Y_i^2$ es suma de los cuadrados de p variables $N(0,1)$ independientes.

2.2.3 Caso bivariante

Cuando $p = 2$, la función de densidad se puede expresar en función de las medias y varianzas $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ y del coeficiente de correlación ρ =cor(X_1, X_2) :

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\frac{1}{1-\rho^2}\left\{\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right.\right.,$$

siendo $-1 < \rho < +1$. (Figura 2.1). Se verifica:

1. Hay independencia estocástica si y sólo si $\rho = 0$.
2. La distribución de la variable marginal X_i es $N(\mu_i, \sigma_i^2)$.
3. La función de densidad de X_2 condicionada a $X_1 = x$ es

$$f(x_2/x_1) = \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} \exp\left[\frac{-[(x_2 - \mu_2 - \rho(\sigma_2/\sigma_1)(x_1 - \mu_1)]^2}{2\sigma_2^2(1-\rho^2)}\right],$$

densidad de la distribución normal $N(\mu_2 + \rho(\sigma_2/\sigma_1)(x_1 - \mu_1), \sigma_2^2(1-\rho^2))$.

4. La regresión es de tipo lineal, es decir, las curvas de regresión de la media

$$x_2 = E(X_2/X_1 = x_1), \qquad x_1 = E(X_1/X_2 = x_2),$$

son las rectas de regresión.

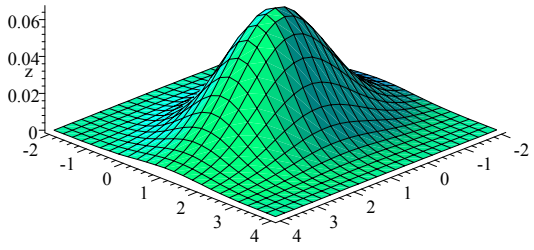

Figura 2.1: Función de densidad de una distribución normal bivariante de medias 1 y 1, desviaciones típicas 2 y 2, coeficiente de correlación 0.8.

2.3 Distribución de Wishart

La distribución de Wishart es la que sigue una matriz aleatoria simétrica definida positiva, generaliza la distribución ji-cuadrado y juega un papel importante en inferencia multivariante. Un ejemplo destacado lo constituye la distribución de la matriz de covarianzas $\mathbf{S}$, calculada a partir de una matriz de datos donde las filas son observaciones normales multivariantes.

Definición

Si las filas de la matriz $\mathbf{Z}_{n\times p}$ son independientes $N_p(0,\Sigma)$ entonces diremos que la matriz $\mathbf{Q}=\mathbf{Z}'\mathbf{Z}$ es Wishart $W_p(\Sigma,n)$, con parámetros Σ y n grados de libertad.

Textos avanzados prueban que cuando Σ es definida positiva y $n\geq p$, la densidad de $\mathbf{Q}$ es

$$f(\mathbf{Q}) = c|\mathbf{Q}|^{(n-p-1)}\exp(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\mathbf{Q})),$$

siendo

$$c^{-1} = 2^{np/2}\pi^{p(p-1)/4}|\Sigma|^{n/2}\prod_{i=1}^{p}\Gamma(\frac{1}{2}(n+1-i).$$

Propiedades:

1. Si $\mathbf{Q}_1,\mathbf{Q}_2$ son independientes Wishart $W_p(\Sigma,m), W_p(\Sigma,n)$, entonces la suma $\mathbf{Q}_1+\mathbf{Q}_2$ es también Wishart $W_p(\Sigma,m+n)$.

2. Si $\mathbf{Q}$ es Wishart $W_p(\Sigma, n)$, y separamos las variables en dos conjuntos y consideramos las particiones correspondientes de las matrices Σ y $\mathbf{Q}$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix},$$

Entonces $\mathbf{Q}_{11}$ es $W_p(\Sigma_{11}, n)$ y $\mathbf{Q}_{22}$ es $W_p(\Sigma_{22}, n)$.

3. Si $\mathbf{Q}$ es Wishart $W_p(\Sigma, n)$ y $\mathbf{T}$ es una matriz $p \times q$ de constantes, entonces $\mathbf{T}'\mathbf{Q}\mathbf{T}$ es $W_q(\mathbf{T}'\Sigma\mathbf{T}, n)$. En particular, si $\mathbf{t}$ es un vector, entonces

$$\frac{\mathbf{t}'\mathbf{Q}\mathbf{t}}{\mathbf{t}\Sigma\mathbf{t}} \quad \text{es} \quad \chi^2_n.$$

2.4 Distribución de Hotelling

Es una generalización multivariante de la distribución t de Student.

Definición

Si $\mathbf{y}$ es $N_p(\mathbf{0}, \mathbf{I})$, $\mathbf{Q}$ es Wishart $W_p(\mathbf{I}, m)$ y además $\mathbf{y}, \mathbf{Q}$ son independientes, entonces

$$T^2 = m\mathbf{y}'\mathbf{Q}^{-1}\mathbf{y}$$

sigue la distribución T^2 de Hotelling, que se indica por $T^2(p, m)$.

Propiedades:

1. Si $\mathbf{x}$ es $N_p(\mu,\Sigma)$ independiente de $\mathbf{M}$ que es $W_p(\Sigma, m)$, entonces

$$T^2 = m(\mathbf{x}-\mu)'\mathbf{M}^{-1}(\mathbf{x}-\mu) \sim T^2(p, m).$$

2. T^2 está directamente relacionada con la distribución de Fisher-Snedecor

$$T^2(p, m) \equiv \frac{mp}{m-p+1} F^p_{m-p+1}.$$

3. Si $\overline{\mathbf{x}}, \mathbf{S}$ son el vector de medias y la matriz de covarianzas de la matriz $\mathbf{X}_{n\times p}$ con filas independientes $N_p(\mu, \Sigma)$, entonces

$$(n-1)(\overline{\mathbf{x}}-\mu)'\mathbf{S}^{-1}(\overline{\mathbf{x}}-\mu) \sim T^2(p, n-1),$$

y por lo tanto

$$\frac{n-p}{p}(\overline{\mathbf{x}}-\mu)'\mathbf{S}^{-1}(\overline{\mathbf{x}}-\mu) \sim F^p_{n-p}.$$

4. Si $\overline{\mathbf{x}}, \mathbf{S}_1, \overline{\mathbf{y}}, \mathbf{S}_2$ son el vector de medias y la matriz de covarianzas de las matrices $\mathbf{X}_{n_1 \times p}, \mathbf{Y}_{n_2 \times p}$, respectivamente, con filas independientes $N_p(\mu, \Sigma)$, y consideramos la estimación conjunta centrada de Σ

$$\widetilde{\mathbf{S}} = (n_1\mathbf{S}_1 + n_2\mathbf{S}_2)/(n_1 + n_2 - 2),$$

entonces

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}} - \overline{\mathbf{y}})'\widetilde{\mathbf{S}}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \sim T^2(p, n_1 + n_2 - 2)$$

y por lo tanto

$$\frac{n_1 + n_2 - 1 - p}{(n_1 + n_2 - 2)p} T^2 \sim F^p_{n_1+n_2-1-p}.$$

2.5 Distribución de Wilks

La distribución F surge considerando el cociente

$$F = \frac{A/m}{B/n},$$

donde A, B són ji-cuadrados independients con m, n grados de libertad. Si consideramos la distribución

$$\Lambda = \frac{A}{A + B},$$

la relación entre Λ i F es

$$F = \frac{m}{n}\frac{\Lambda}{1 - \Lambda}.$$

La distribución de Wilks generaliza esta relación.

Definición

Si las matrices $\mathbf{A}, \mathbf{B}$ de orden $p \times p$ son independientes Wishart $W_p(\Sigma, m), W_p(\Sigma, n)$, respectivamente, la distribución del cociente de determinantes

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|}$$

es, por definición, la distribución lambda de Wilks, que indicaremos por $\Lambda(p, m, n)$.

Propiedades:

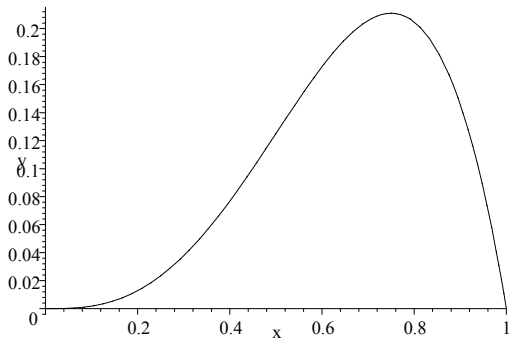

Figura 2.2: Un ejemplo de función de densidad lambda de Wilks.

1. $0 \leq \Lambda \leq 1$ y además Λ no depende de Σ. Por lo tanto, podemos estudiarla suponiendo $\Sigma = \mathbf{I}$.

2. Su distribución es equivalente a la del producto de n variables beta independientes:

$$\Lambda(p, m, n) \sim \prod_{i=1}^{n} U_i,$$

donde U_i es beta $B(\frac{1}{2}(m + i - p), \frac{1}{2}p)$.

3. Los parámetros se pueden permutar manteniendo la misma distribución. Concretamente:

$$\Lambda(p, m, n) \sim \Lambda(n, m + n - p, p).$$

4. Para valores 1 ó 2 de p, la distribución de Λ equivale a la F, según las fórmulas

$$\begin{array}{lll} \frac{1-\Lambda}{\Lambda}\frac{m}{n} & \sim F_m^n & (p = 1) \\ \frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{m-1}{n} & \sim F_{2(m-1)}^{2n} & (p = 2) \end{array} \tag{2.8}$$

5. En general, una transformación de Λ equivale, exacta o asintóticamente, a la distribución F.

2.6 Relaciones entre Wilks, Hotelling y F

A. Probemos la relación entre Λ y F cuando $p = 1$. Sean $A \sim \chi^2_m, B \sim \chi^2_n$ independientes. Entonces $\Lambda = A/(A+B) \sim \Lambda(1, m, n)$ y $F = (n/m)A/B = (n/m)\overline{F} \sim F^m_n$. Tenemos que $\Lambda = (A/B)/(A/B+1) = \overline{F}/(1+\overline{F})$, luego $\overline{F} = \Lambda/(1-\Lambda) \Rightarrow (n/m)\Lambda/(1-\Lambda) \sim F^m_n$. Mas si $F \sim F^m_n$ entonces $1/F \sim F^n_m$. Hemos demostrado que:

$$\frac{1-\Lambda(1,m,n)}{\Lambda(1,m,n)}\frac{m}{n} \sim F^n_m. \tag{2.9}$$

B. Recordemos que $\mathbf{y}$ es un vector columna y por lo tanto $\mathbf{yy'}$ es una matriz $p \times p$. Probemos la relación entre las distribuciones T^2 y F. Tenemos $T^2 = m\mathbf{y'Q^{-1}y}$, donde $\mathbf{Q}$ es $W_p(\mathbf{I},m)$, y $\mathbf{yy'}$ es $W_p(\mathbf{I},1)$. Se cumple

$$|\mathbf{Q}+\mathbf{yy'}| = |\mathbf{Q}||1+\mathbf{y'Q^{-1}y}|,$$

que implica

$$1+\mathbf{y'Q^{-1}y} = |\mathbf{Q}+\mathbf{yy'}|/|\mathbf{Q}| = 1/\Lambda,$$

donde $\Lambda = |\mathbf{Q}|/|\mathbf{Q}+\mathbf{yy'}| \sim \Lambda(p, m, 1) \sim \Lambda(1, m+1-p, p)$. Además $\mathbf{y'Q^{-1}y} = 1/\Lambda - 1 = (1-\Lambda)/\Lambda$. De (2.9) tenemos que $\mathbf{y'Q^{-1}y}(m+1-p)/p \sim F^p_{m+1-p}$ y por lo tanto

$$T^2 = m\mathbf{y'Q^{-1}y} \sim \frac{mp}{m+1-p}F^p_{m+1-p}.$$

2.7 Distribuciones con marginales dadas

Sea $F(x, y)$ la función de distribución de dos variables aleatorias (X, Y). Tenemos

$$H(x, y) = P(X \leq x, Y \leq y).$$

Consideremos las distribuciones marginales, es decir las distribuciones univariantes de X y de Y :

$$\begin{aligned} F(x) &= P(X \leq x) = H(x, \infty), \\ G(y) &= P(Y \leq y) = H(\infty, y). \end{aligned}$$

Un procedimiento para la obtención de modelos de distribuciones bivariantes consiste en encontrar H a partir de F, G y posiblemente algún parámetro.

Si suponemos X, Y independientes, una primera distribución es

$$H^0(x, y) = F(x)G(y).$$

M. Fréchet introdujo las distribuciones bivariantes

$$\begin{aligned} H^-(x, y) &= \max\{F(x) + G(y) - 1, 0\}, \\ H^+(x, y) &= \min\{F(x), G(y)\} \end{aligned}$$

y demostró la desigualdad

$$H^-(x, y) \leq H(x, y) \leq H^+(x, y).$$

Cuando la distribución es H^-, entonces se cumple la relación funcional entre X, Y

$$F(X) + G(Y) = 1.$$

y la correlación ρ^- es mínima. Cuando la distribución es H^+, entonces se cumple la relación funcional entre X, Y

$$F(X) = G(Y)$$

y la correlación ρ^+ es máxima. Previamente W. Hoeffding había probado la siguiente fórmula para la covarianza

$$\text{cov}(X, Y) = \int_{R^2} (H(x, y) - F(x)G(y))dxdy$$

y demostrado la desigualdad

$$\rho^- \leq \rho \leq \rho^+,$$

donde ρ^-, ρ y ρ^+ son las correlaciones entre X, Y cuando la distribución bivariante es H^-, H y H^+, respectivamente.

Posteriormente, diversos autores han propuesto distribuciones bivariantes paramétricas a partir de las marginales F, G, que en algunos casos contienen a H^-, H^0 y H^+. Escribiendo F, G, H para indicar $F(x), G(y), H(x, y)$, algunas familias son:

1. Farlie-Gumbel-Morgenstern:

$$H_\theta = FG[1 + \theta(1 - F)(1 - G)], \quad -1 \leq \theta \leq 1.$$

2. Clayton-Oakes:

$$H_\alpha = [F^{-\alpha} + G^{-\alpha} - 1]^{-1/\alpha}, \quad -1 \leq \alpha < \infty.$$

3. Ali-Mikhail-Haq:

$$H_\theta = FG/[1 - \theta(1 - F)(1 - G)] \quad -1 \leq \theta \leq 1.$$

4. Cuadras-Augé:

$$H_\theta = (\min\{F, G\})^\theta (FG)^{1-\theta}, \quad -1 \leq \theta \leq 1.$$

5. Familia de correlación:

$$H_\theta(x, y) = \theta F(\min\{x, y\}) + (1 - \theta)F(x)J(y), \quad -1 \leq \theta \leq 1,$$

siendo $J(y) = [G(y) - \theta F(y))/(1 - \theta)$ una función de distribución univariante.

2.8 Complementos

La distribución normal multivariante es, con diferencia, la más utilizada en análisis multivariante. Textos como Anderson (1956), Rao (1973), se basan, casi exclusivamente, en la suposición de normalidad. Más recientemente se han estudiado generalizaciones, como las distribuciones elípticas, cuya densidad es de la forma

$$f(\mathbf{x}) = |\Sigma|^{-1/2} g((\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)),$$

donde g es una función positiva creciente. Otras distribuciones importantes son la multinomial y la Dirichlet.

Cuando se estudiaron muestras normales multivariantes, pronto se planteó la necesidad de encontrar la distribución de la matriz de covarianzas, y de algunos estadísticos apropiados para realizar tests multivariantes. Así fue como J. Wishart, H. Hotelling y S. S. Wilks propusieron las distribuciones que llevan sus nombres, en los años 1928, 1931 y 1932, respectivamente.

El estudio de las distribuciones con marginales dadas proporciona un método de construcción de distribuciones univariantes y multivariantes. Algunas referencias son: Hutchinson y Lai (1990), Cuadras y Augé (1981),

Cuadras (1992), Cuadras (2006). La fórmula de Hoeffding admite la siguiente generalización

$$\mathrm{cov}(\alpha(X), \beta(Y)) = \int_{R^2} (H(x,y) - F(x)G(y)) d\alpha(x) d\beta(y)$$

(Cuadras, 2002).

Capítulo 3

INFERENCIA MULTIVARIANTE

3.1 Conceptos básicos

Sea $f(\mathbf{x}, \boldsymbol{\theta})$ un modelo estadístico. La función "score" se define como

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

Una muestra multivariante está formada por las n filas $\mathbf{x}_1', \ldots, \mathbf{x}_p'$ independientes de una matriz de datos $\mathbf{X}_{n \times p}$. La función de verosimilitud es

$$L(\mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i, \boldsymbol{\theta}).$$

La función "score" de la muestra es

$$z(\mathbf{X}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{x}_i, \boldsymbol{\theta}).$$

La matriz de información de Fisher $F(\boldsymbol{\theta})$ es la matriz de covarianzas de $z(\mathbf{X}, \boldsymbol{\theta})$. Cuando un modelo estadístico es regular se verifica:

a) $E(z(\mathbf{X}, \boldsymbol{\theta})) = \mathbf{0}$.

b) $F(\boldsymbol{\theta}) = E(z(\mathbf{X}, \boldsymbol{\theta}) z(\mathbf{X}, \boldsymbol{\theta})')$.

Un estimador $t(\mathbf{X})$ de $\boldsymbol{\theta}$ es insesgado si $E(t(\mathbf{X})) = \boldsymbol{\theta}$. La desigualdad de Cramér-Rao dice que si $\text{cov}(t(\mathbf{X}))$ es la matriz de covarianzas de $t(\mathbf{X})$, entonces

$$\text{cov}(t(\mathbf{X})) \geq F(\boldsymbol{\theta})^{-1},$$

en el sentido de que la diferencia $\text{cov}(t(\mathbf{X}))-F(\boldsymbol{\theta})^{-1}$ es una matriz semidefinida positiva.

Un estimador $\widehat{\boldsymbol{\theta}}$ del parámetro desconocido $\boldsymbol{\theta}$ es máximo verosímil si maximiza la función $L(\mathbf{X},\boldsymbol{\theta})$. En condiciones de regularidad, podemos obtener $\widehat{\boldsymbol{\theta}}$ resolviendo la ecuación

$$\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{x}_i,\boldsymbol{\theta})=\mathbf{0}.$$

Entonces el estimador máximo verosímil $\widehat{\boldsymbol{\theta}}_n$ obtenido a partir de una muestra de tamaño n satisface:

a) Es asintóticamente normal con vector de medias $\boldsymbol{\theta}$ y matriz de covarianzas $(nF_1(\boldsymbol{\theta}))^{-1}$, donde $F_1(\boldsymbol{\theta})$ es la matriz de información de Fisher para una sola observación.

b) Si $t(\mathbf{X})$ es estimador insesgado de $\boldsymbol{\theta}$ tal que $\text{cov}(t(\mathbf{X}))=(nF_1(\boldsymbol{\theta}))^{-1}$, entonces $\widehat{\boldsymbol{\theta}}_n=t(\mathbf{X})$.

c) $\widehat{\boldsymbol{\theta}}_n$ converge en probabilidad a $\boldsymbol{\theta}$.

3.2 Estimación de medias y covarianzas

Si las n filas $\mathbf{x}_1',\ldots,\mathbf{x}_n'$ de $\mathbf{X}_{n\times p}$ son independientes $N_p(\mu,\Sigma)$ la función de verosimilitud es

$$L(\mathbf{X},\mu,\Sigma)=\det(2\pi\Sigma)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mu)\Sigma^{-1}(\mathbf{x}_i-\mu)'\right\}$$

Se verifica

$$\begin{aligned}\textstyle\sum_{i=1}^{n}(\mathbf{x}_i-\mu)'\Sigma^{-1}(\mathbf{x}_i-\mu) &= \textstyle\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})'\Sigma^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})+n(\overline{\mathbf{x}}-\mu)'\Sigma^{-1}(\overline{\mathbf{x}}-\mu)\\ &= tr\{\Sigma^{-1}\textstyle\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})'\}\\ &\quad +n(\overline{\mathbf{x}}-\mu)'\Sigma^{-1}(\overline{\mathbf{x}}-\mu)\end{aligned}$$

y por lo tanto el logaritmo de L se puede expresar como

$$\log L(\mathbf{X},\mu,\Sigma)=-\frac{n}{2}\log\det(2\pi\Sigma)-\frac{n}{2}tr(\Sigma^{-1}\mathbf{S})-\frac{n}{2}(\overline{\mathbf{x}}-\mu)'\Sigma^{-1}(\overline{\mathbf{x}}-\mu).$$

Derivando matricialmente respecto de μ y de Σ^{-1} tenemos

$$\begin{aligned}\tfrac{\partial}{\partial\mu}\log L &= n\Sigma^{-1}(\overline{\mathbf{x}}-\mu)=0,\\ \tfrac{\partial}{\partial\Sigma^{-1}}\log L &= \tfrac{n}{2}\left[\Sigma-S-(\overline{\mathbf{x}}-\mu)(\overline{\mathbf{x}}-\mu)'\right]=0.\end{aligned}$$

Las estimaciones máximo-verosímiles de μ, Σ son pues

$$\widehat{\mu} = \overline{\mathbf{x}}, \quad \widehat{\Sigma} = \mathbf{S}.$$

Si sólo μ es desconocido, la matriz de información de Fisher es

$$F(\mu) = E(n\Sigma^{-1}(\overline{\mathbf{x}} - \mu)n\Sigma^{-1}(\overline{\mathbf{x}} - \mu)') = n\Sigma^{-1}$$

y como $\text{cov}(\overline{\mathbf{x}}) = \Sigma/n$, tenemos $\overline{\mathbf{x}}$ que alcanza laa cota de Cramér-Rao.

Probaremos más adelante que:

1. $\overline{\mathbf{x}}$ es $N_p(\mu, \Sigma/n)$.
2. $\overline{\mathbf{x}}$ y $\mathbf{S}$ son estocásticamente independientes.
3. $n\mathbf{S}$ sigue la distribución de Wishart.

3.3 Tests multivariantes

Un primer método para construir tests sobre los parámetros de una población normal, se basa en las propiedades anteriores, que dan lugar a estadísticos con distribución conocida (ji-cuadrado, F).

3.3.1 Test sobre la media: una población

Supongamos que las filas de $\mathbf{X}_{n\times p}$ son independientes $N_p(\mu, \Sigma)$. Sea μ_0 un vector de medias conocido. Queremos realizar un test sobre la hipótesis

$$H_0 : \mu = \mu_0$$

1. Si Σ es conocida, como $\overline{\mathbf{x}}$ es $N_p(\mu, \Sigma/n)$, el estadístico de contraste es

$$n(\overline{\mathbf{x}}-\mu_0)'\Sigma^{-1}(\overline{\mathbf{x}}-\mu_0) \sim \chi^2_p.$$

2. Si Σ es desconocida, como $(n-1)(\overline{\mathbf{x}}-\mu)\mathbf{S}^{-1}(\overline{\mathbf{x}}-\mu) \sim T^2(p, n-1)$, el estadístico de contraste es

$$\frac{n-p}{p}(\overline{\mathbf{x}}-\mu_0)'\mathbf{S}^{-1}(\overline{\mathbf{x}}-\mu_0) \sim F^p_{n-p}.$$

En ambos casos se rechaza H_0 para valores grandes significativos del estadístico.

3.3.2 Test sobre la media: dos poblaciones

Supongamos ahora que tenemos dos matrices de datos independientes $\mathbf{X}_{n_1\times p}$, $\mathbf{Y}_{n_2\times p}$ que provienen de distribuciones $N_p(\mu_1,\Sigma)$, $N_p(\mu_2,\Sigma)$. Queremos construir un test sobre la hipótesis

$$H_0 : \mu_1 = \mu_2.$$

1. Si Σ es conocida, como $(\overline{\mathbf{x}}-\overline{\mathbf{y}})$ es $N_p(\mu_1-\mu_2,(1/n_1+1/n_2)\Sigma)$ el estadístico de contraste es

$$\frac{n_1n_2}{n_1+n_2}(\overline{\mathbf{x}}-\overline{\mathbf{y}})'\Sigma^{-1}(\overline{\mathbf{x}}-\overline{\mathbf{y}})\sim\chi^2_p.$$

2. Si Σ es desconocida, el estadístico de contraste es

$$\frac{n_1+n_2-1-p}{(n_1+n_2-2)p}\frac{n_1n_2}{n_1+n_2}(\overline{\mathbf{x}}-\overline{\mathbf{y}})'\widetilde{\mathbf{S}}^{-1}(\overline{\mathbf{x}}-\overline{\mathbf{y}})\sim F^p_{n_1+n_2-1-p}.$$

3.3.3 Comparación de medias

Supongamos que las filas de g matrices de datos son independientes, y que provienen de la observación de g poblaciones normales multivariantes:

matriz	orden	medias	covarianzas	distribución
$\mathbf{X}_1$	$n_1\times p$	$\overline{\mathbf{x}}_1$	$\mathbf{S}_1$	$N_p(\mu_1,\Sigma)$
$\mathbf{X}_2$	$n_2\times p$	$\overline{\mathbf{x}}_2$	$\mathbf{S}_2$	$N_p(\mu_2,\Sigma)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{X}_g$	$n_g\times p$	$\overline{\mathbf{x}}_g$	$\mathbf{S}_g$	$N_p(\mu_g,\Sigma)$

(3.1)

El vector de medias generales y la estimación centrada de la matriz de covarianzas común Σ son

$$\overline{\mathbf{x}}=\frac{1}{n}\sum_{i=1}^{g}n_i\overline{\mathbf{x}}_i,\quad \mathbf{S}=\frac{1}{n-g}\sum_{i=1}^{g}n_i\mathbf{S}_i,$$

siendo $\mathbf{S}_i=n_i^{-1}\mathbf{X}_i'\mathbf{H}\mathbf{X}_i$, $n=\sum_{i=1}^g n_i$.

Deseamos construir un test para decidir si podemos aceptar la hipótesis de igualdad de medias

$$H_0 : \mu_1=\mu_2=\ldots=\mu_g.$$

Introducimos las siguientes matrices:

$$\begin{array}{lll} \mathbf{B} & = \sum_{i=1}^{g} n_i(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})' & \text{(dispersión entre grupos)} \\ \mathbf{W} & = \sum_{i=1}^{g} \sum_{\alpha=1}^{n_i} (\overline{\mathbf{x}}_{i\alpha} - \overline{\mathbf{x}}_i)(\overline{\mathbf{x}}_{i\alpha} - \overline{\mathbf{x}}_i)' & \text{(dispersión dentro grupos)} \\ \mathbf{T} & = \sum_{i=1}^{g} \sum_{\alpha=1}^{n_i} (\overline{\mathbf{x}}_{i\alpha} - \overline{\mathbf{x}})(\overline{\mathbf{x}}_{i\alpha} - \overline{\mathbf{x}})' & \text{(dispersión total)} \end{array}$$

Se verifica que $\mathbf{W} = (n-g)\mathbf{S}$ y la relación:

$$\mathbf{T} = \mathbf{B} + \mathbf{W}.$$

Si la hipótesis nula es cierta, se verifica además

$$\mathbf{B} \sim W_p(\Sigma, g-1),\ \mathbf{W} \sim W_p(\Sigma, n-g),\ \mathbf{T} \sim W_p(\Sigma, n-1),$$

$\mathbf{B}, \mathbf{W}$ son estocásticamente independientes,

por lo tanto, si H_0 es cierta

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|} \sim \Lambda(p, n-g, g-1).$$

Rechazaremos H_0 si Λ es pequeña y significativa, o si la transformación a una F es grande y significativa.

3.4 Teorema de Cochran

Algunos resultados de la sección anterior son una consecuencia del teorema de Cochran.

Lema 3.4.1 *Sea* $\mathbf{X}(n \times p)$ *una matriz de datos* $N_p(\mu, \Sigma)$ *y* $\mathbf{u}, \mathbf{v}$ *dos vectores* $n \times 1$ *tales que* $\mathbf{u}'\mathbf{u} = \mathbf{v}'\mathbf{v} = 1, \mathbf{u}'\mathbf{v} = 0$.

1. *Si* $\mu = 0$ *entonces* $\mathbf{y}' = \mathbf{u}'\mathbf{X}$ *es* $N_p(0, \Sigma)$.
2. $\mathbf{y}' = \mathbf{u}'\mathbf{X}$ *es independiente de* $\mathbf{z}' = \mathbf{v}'\mathbf{X}$.

Demost.: Sean $\mathbf{x}_1', \ldots, \mathbf{x}_n'$ las filas (independientes) de $\mathbf{X}$. Si $\mathbf{u} = (u_1, \ldots, u_n)'$ entonces $\mathbf{y}' = \mathbf{u}'\mathbf{X} = \sum_{i=1}^{n} u_i \mathbf{x}_i$ es normal multivariante con $\mu = 0$ y matriz de covarianzas

$$\begin{aligned} E(\mathbf{y}\mathbf{y}') &= E(\textstyle\sum_{i=1}^{n} u_i \mathbf{x}_i)(\sum_{i=1}^{n} u_i \mathbf{x}_i)' = E(\sum_{i,j=1}^{n} u_i u_j \mathbf{x}_i \mathbf{x}_j') \\ &= \textstyle\sum_{i,j=1}^{n} u_i u_j E(\mathbf{x}_i \mathbf{x}_j') = \sum_{i=1}^{n} u_i^2 E(\mathbf{x}_i \mathbf{x}_i') \\ &= \textstyle\sum_{i=1}^{n} u_i^2 \Sigma = \Sigma. \end{aligned}$$

Análogamente, si $\mathbf{v} = (v_1, \ldots, v_n)'$, $\mathbf{z}' = \mathbf{v}'\mathbf{X}$ es también normal y suponiendo $\mu = 0$,

$$E(\mathbf{y}\mathbf{z}') = \sum_{i=1}^{n} u_i v_j E(\mathbf{x}_i \mathbf{x}_j') = \sum_{i=1}^{n} u_i v_i E(\mathbf{x}_i \mathbf{x}_i') = \mathbf{u}'\mathbf{v}\Sigma = \mathbf{0},$$

que prueba la independencia entre $\mathbf{y}, \mathbf{z}$. Este resultado no depende de μ.□

Teorema 3.4.2 *Sea $\mathbf{X}(n \times p)$ una matriz de datos $N_p(0, \Sigma)$ y sea $\mathbf{C}(n \times n)$ una matriz simétrica.*

1. $\mathbf{X}'\mathbf{C}\mathbf{X}$ *tiene la misma distribución que una suma ponderada de matrices $W_p(\Sigma, 1)$, donde los pesos son valores propios de $\mathbf{C}$.*

2. $\mathbf{X}'\mathbf{C}\mathbf{X}$ *es Wishart $W_p(\Sigma, r)$ si y sólo si $\mathbf{C}$ es idempotente y $r(\mathbf{C}) = r$.*

Demost.: Sea

$$\mathbf{C} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i'$$

la descomposición espectral de $\mathbf{C}$, es decir, $\mathbf{C}\mathbf{u}_i = \lambda_i \mathbf{u}_i$. Entonces

$$\mathbf{X}'\mathbf{C}\mathbf{X} = \sum \lambda_i \mathbf{y}_i' \mathbf{y}_i$$

Por el Lema 3.4.1 anterior, las filas $\mathbf{y}_i'$ de la matriz

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1'\mathbf{X} \\ \vdots \\ \mathbf{u}_n'\mathbf{X} \end{pmatrix},$$

son también independientes $N_p(0, \Sigma)$ y cada $\mathbf{y}_i \mathbf{y}_i'$ es $W_p(\Sigma, 1)$.

Si $\mathbf{C}^2 = \mathbf{C}$ entonces $\mathbf{C}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ siendo $\lambda_i = 0$ ó 1. Por lo tanto $r = \text{tr}(\mathbf{C})$ y

$$\mathbf{X}'\mathbf{C}\mathbf{X} = \sum_{i=1}^{r} \mathbf{y}_i \mathbf{y}_i' \sim W_p(\Sigma, r).\square$$

El siguiente resultado se conoce como teorema de Craig, y junto con el teorema de Cochran, permite construir tests sobre vectores de medias.

Teorema 3.4.3 *Sea* $\mathbf{X}(n\times p)$ *una matriz de datos* $N_p(\mu,\Sigma)$ *y sean* $\mathbf{C}_1(n\times n)$, $\mathbf{C}_2(n\times n)$ *matrices simétricas. Entonces* $\mathbf{X}'\mathbf{C}_1\mathbf{X}$ *es independiente de* $\mathbf{X}'\mathbf{C}_2\mathbf{X}$ *si* $\mathbf{C}_1\mathbf{C}_2=\mathbf{0}$.

Demost.:

$$\begin{array}{ll}\mathbf{C}_1=\sum_{i=1}^n \lambda_i(1)\mathbf{u}_i\mathbf{u}_i', & \mathbf{X}'\mathbf{C}_1\mathbf{X}=\sum \lambda_i(1)\mathbf{y}_i\mathbf{y}_i',\\ \mathbf{C}_2=\sum_{j=1}^n \lambda_j(2)\mathbf{v}_j\mathbf{v}_j', & \mathbf{X}'\mathbf{C}_2\mathbf{X}=\sum \lambda_j(2)\mathbf{z}_j\mathbf{z}_j',\end{array}$$

siendo $\mathbf{y}_i'=\mathbf{u}_i'\mathbf{X}, \mathbf{z}_j'=\mathbf{v}_j'\mathbf{X}$. Por otra parte

$$\begin{aligned}\mathbf{C}_1\mathbf{C}_2 &= \sum \lambda_i(1)\lambda_j(2)\mathbf{u}_i\mathbf{u}_i'\mathbf{v}_j\mathbf{v}_j',\\ \mathbf{C}_1\mathbf{C}_2 &= \mathbf{0}\Rightarrow\lambda_i(1)\lambda_j(2)\mathbf{u}_i'\mathbf{v}_j=0, \quad \forall i,j.\end{aligned}$$

Si suponemos $\lambda_i(1)\lambda_j(2)\neq 0$, entonces por el Lema 3.4.1 $\mathbf{y}_i'(1\times p)=\mathbf{u}_i'\mathbf{X}$ es independiente de $\mathbf{z}_j'(1\times p)=\mathbf{v}_j'\mathbf{X}$. Así $\mathbf{X}'\mathbf{C}_1\mathbf{X}$ es independiente de $\mathbf{X}'\mathbf{C}_2\mathbf{X}$.□

Una primera consecuencia del Teorema anterior es la independencia entre vectores de medias y matrices de covarianzas muestrales.

Teorema 3.4.4 *Sea* $\mathbf{X}(n\times p)$ *una matriz de datos* $N_p(\mu,\Sigma)$. *Entonces :*

1. *La media* $\overline{\mathbf{x}}$ *es* $N_p(\mu,\Sigma/n)$.
2. *La matriz de covarianzas* $\mathbf{S}=\mathbf{X}'\mathbf{H}\mathbf{X}/n$ *verifica* $n\mathbf{S}\sim W_p(\Sigma,n-1)$.
3. $\overline{\mathbf{x}}$ *y* $\mathbf{S}$ *son estocásticamente independientes.*

Demost.: Consideremos $\mathbf{C}_1=n^{-1}\mathbf{1}\mathbf{1}'$. Tenemos $\text{rang}(\mathbf{C}_1)=1$, $\mathbf{X}'\mathbf{C}_1\mathbf{X}=\overline{\mathbf{x}}\overline{\mathbf{x}}'$. Consideremos también $\mathbf{C}_2=\mathbf{H}$. Como $\mathbf{C}_1\mathbf{C}_2=\mathbf{0}$ deducimos que $\overline{\mathbf{x}}$ es independiente de $\mathbf{S}$.

Por otra parte, como $\mathbf{H}^2=\mathbf{H}$, $\mathbf{H}\mathbf{1}=\mathbf{0}$, $\text{rang}(\mathbf{H})=n-1$, $\mathbf{H}$ tiene el valor propio 1 con multiplicidad $n-1$. Así $\mathbf{u}_i$, vector propio de valor propio 1, es ortogonal a $\mathbf{1}$, resultando que $\mathbf{y}_i'=\mathbf{u}_i'\mathbf{X}$ verifica $E(\mathbf{y}_i')=(\sum_{\alpha=1}^n u_{i\alpha})\mu=(\mathbf{u}_i'\mathbf{1})\mu=0\mu=\mathbf{0}$. Si $\mathbf{u}_j$ es otro vector propio, $\mathbf{y}_i,\mathbf{y}_j$ son independientes (Lema 3.4.1). Tenemos que $n\mathbf{S}=\sum_{i=1}^{n-1}\mathbf{y}_i\mathbf{y}_i'$, donde los $\mathbf{y}_i\mathbf{y}_i'$ son $W_p(\Sigma,1)$ independientes.□

Teorema 3.4.5 *Sean* $\mathbf{X}_i$, *matrices de datos independientes de orden* $n_i\times p$ *con distribución* $N_p(\mu_i,\Sigma)$, $i=1,\ldots g$, $n=\sum_{i=1}^g n_i$. *Si la hipótesis nula*

$$H_0:\mu_1=\mu_2=\ldots=\mu_g$$

es cierta, entonces $\mathbf{B},\mathbf{W}$ *son independientes con distribuciones Wishart:*

$$\mathbf{B}\sim W_p(\Sigma,g-1), \quad \mathbf{W}\sim W_p(\Sigma,n-g).$$

Demost.: Escribimos las matrices de datos como una única matriz

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

Sean

$$\begin{array}{ll} \mathbf{1}_1 & = (1,\ldots,1,0,\ldots,0),\ldots,\mathbf{1}_g = (0,\ldots 0,1,\ldots 1), \\ \mathbf{1} & = \sum_{i=1}^{g} \mathbf{1}_i = (1,\ldots,1,\ldots,1,\ldots,1), \end{array}$$

donde $\mathbf{1}_1$ tiene n_1 unos y el resto ceros, etc. Sean también

$$\begin{array}{ll} \mathbf{I}_i & = \text{diag}(\mathbf{1}_i), \quad \mathbf{I} = \sum_{i=1}^{g} \mathbf{I}_i, \\ \mathbf{H}_i & = \mathbf{I}_i - n_i^{-1}\mathbf{1}_i\mathbf{1}_i' \\ \mathbf{C}_1 & = \sum_{i=1}^{g} \mathbf{H}_i, \quad \mathbf{C}_2 = \sum_{i=1}^{g} n_i^{-1}\mathbf{1}_i\mathbf{1}_i' - n^{-1}\mathbf{1}\mathbf{1}'. \end{array}$$

Entonces

$$\begin{array}{rlrll} \mathbf{C}_1^2 & = \mathbf{C}_1, & \mathbf{C}_2^2 & = \mathbf{C}_2, & \mathbf{C}_1\mathbf{C}_2 = \mathbf{0}, \\ rang(\mathbf{C}_1) & = n-k, & rang(\mathbf{C}_2) & = g-1, & \\ \mathbf{W} & = \mathbf{X}'\mathbf{C}_1\mathbf{X}, & \mathbf{B} & = \mathbf{X}'\mathbf{C}_2\mathbf{X}. & \end{array}$$

El resultado es consecuencia de los Teoremas 3.4.4 y 3.4.5.□

3.5 Construcción de tests multivariantes

3.5.1 Razón de verosimilitud

Supongamos que la función de densidad de $(X_1,\ldots,X_p)$ es $f(\mathbf{x},\boldsymbol{\theta})$, donde $\mathbf{x} \in R^p$ y $\theta \in \boldsymbol{\Theta}$, siendo $\boldsymbol{\Theta}$ una región paramétrica de dimensión geométrica r. Sea $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ una subregión paramétrica de dimensión s, y planteamos el test de hipótesis

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \quad \text{vs} \quad H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta} - \boldsymbol{\Theta}_0.$$

Sea $\mathbf{x}_1,\ldots,\mathbf{x}_n$ una muestra de valores independientes de $\mathbf{X}$, consideremos la función de verosimilitud

$$L(\mathbf{x}_1,\ldots,\mathbf{x}_n;\theta) = \prod_{i=1}^{n} f(\mathbf{x},\boldsymbol{\theta})$$

y sea $\widehat{\boldsymbol{\theta}}$ el estimador máximo verosímil de $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Consideremos análogamente $\widehat{\boldsymbol{\theta}}_0$, el estimador de máxima verosimilitud de $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Tenemos que $\widehat{\boldsymbol{\theta}}$ maximiza L sin restricciones y $\widehat{\boldsymbol{\theta}}_0$ maximiza L cuando se impone la condición de que pertenezca a $\boldsymbol{\Theta}_0$. La razón de verosimilitud es el estadístico

$$\lambda_R = \frac{L(\mathbf{x}_1, \ldots, \mathbf{x}_n; \widehat{\boldsymbol{\theta}}_0)}{L(\mathbf{x}_1, \ldots, \mathbf{x}_n; \widehat{\boldsymbol{\theta}})},$$

que satisface $0 \leq \lambda_R \leq 1$. Aceptamos la hipótesis H_0 si λ_R es próxima a 1 y aceptamos la alternativa H_1 si λ_R es significativamente próximo a 0.

El test basado en λ_R tiene muchas aplicaciones en AM, pero en la mayoría de los casos su distribución es desconocida. Existe un importante resultado (atribuido a Wilks), que dice que la distribución de -2 veces el logaritmo de λ_R es ji-cuadrado con $r - s$ g.l. cuando el tamaño de la muestra n es grande.

Teorema 3.5.1 *Bajo ciertas condiciones de regularidad, se verifica:*

$$-2 \log \lambda_R \quad \textit{es asintóticamente } \chi^2_{r-s},$$

donde $s = \dim(\boldsymbol{\Theta}_0) < r = \dim(\boldsymbol{\Theta})$.

Entonces rechazamos la hipótesis H_0 cuando $-2 \log \lambda_R$ sea grande y significativo. Veamos dos ejemplos.

Test de independencia

Si $(X_1, \ldots, X_p)$ es $N(\mu, \Sigma)$, y queremos hacer un test sobre la independencia estocástica de las variables, entonces

$$\begin{array}{rcl} \boldsymbol{\Theta}_0 & = & \{(\mu, \Sigma_0)\}, \quad s = 2p, \\ \boldsymbol{\Theta} & = & \{(\mu, \Sigma)\}, \quad r = p + p(p+1)/2, \end{array}$$

donde Σ_0 es diagonal. $\boldsymbol{\Theta}_0$ contiene las p medias de las variables y las p varianzas. Σ es cualquier matriz definida positiva. Se demuestra (Sección 5.4.2) que

$$-2 \log \lambda_R = -n \log |\mathbf{R}|,$$

donde $\mathbf{R}$ es la matriz de correlaciones. El estadístico $-n \log |\mathbf{R}|$ es asintóticamente ji-cuadrado con

$$q = p + p(p+1)/2 - 2p = p(p-1)/2 \quad \text{g.l.}$$

Si las variables son independientes, tendremos que $\mathbf{R} \approx \mathbf{I}$, $-n \log |\mathbf{R}| \approx 0$, y es probable que $\chi^2_q = -n \log |\mathbf{R}|$ no sea significativo.

Test de comparación de medias

Consideremos el test de comparación de medias planteado en la Sección 3.3.3. Ahora

$$\begin{array}{rcll} \boldsymbol{\Theta}_0 & = & \{(\mu, \Sigma)\}, & s = p + p(p+1)/2, \\ \boldsymbol{\Theta} & = & \{(\mu_1, \ldots, \mu_g), \Sigma)\}, & r = gp + p(p+1)/2, \end{array}$$

donde Σ es matriz definida positiva y μ (vector) es la media común cuando H_0 es cierta. Hay $gp + p(p+1)/2$ parámetros bajo H_1, y $p + p(p+1)/2$ bajo H_0. Se demuestra la relación

$$\lambda_R = \Lambda^{n/2},$$

donde $\Lambda = |\mathbf{W}|/|\mathbf{T}|$ es la lambda de Wilks y $n = n_1 + \ldots + n_g$. Por lo tanto $-n \log \Lambda$ es asintóticamente ji-cuadrado con $r - s = (g-1)p$ g.l. cuando la hipótesis H_0 es cierta.

3.5.2 Principio de unión-intersección

Es un principio general que permite construir tests multivariantes a partir de tests univariantes y se aplica a muchos tests. Como ejemplo, planteemos la hipótesis nula multivariante $H_0 : \mu{=}\mu_0$ como un test univariante. Sea $X_a = \mathbf{Xa}$ una variable compuesta con media $\mu(a) {=}\mu\mathbf{a}$. El test univariante $H_0(a) : \mu(a) {=}\mu_0(a)$ contra la alternativa $H_1(a) : \mu(a) {\neq}\mu_0(a)$ se resuelve mediante la t de Student

$$t(a) = \sqrt{n-1}\frac{\overline{x}(a) - \mu_0(a)}{s(a)} \sim t_{n-1}$$

donde $\overline{x}(a) = \overline{\mathbf{x}}'a$ es la media muestral de X_a y $s^2(a) = \mathbf{a}'\mathbf{S}\mathbf{a}$ es la varianza. Aceptaremos $H_0 : \mu{=}\mu_0$ si aceptamos todas las hipótesis univariantes $H_0(a)$, y nos decidiremos por la alternativa $H_1 : \mu \neq \mu_0$ si aceptamos una sola de las alternativas $H_1(a)$, es decir, formalmente (principio de unión-intersección):

$$H_0 = \bigcap_a H_0(a), \quad H_1 = \bigcup_a H_1(a).$$

Así rechazaremos H_0 si la máxima $t(a)$ resulta significativa. Pues bien, la T^2 de Hotelling (Sección 3.3.1) es precisamente el cuadrado de esta máxima t de Student.

Teorema 3.5.2 *En el test sobre el vector de medias, la T^2 de Hotelling y la t de Student están relacionadas por*

$$T^2 = \max_a t^2(a).$$

Demost.: $(\overline{\mathbf{x}} - \mu_0)$ es un vector columna y podemos escribir $t^2(a)$ como

$$t^2(a) = (n-1)\frac{\mathbf{a}'(\overline{\mathbf{x}} - \mu_0)(\overline{\mathbf{x}} - \mu_0)'\mathbf{a}}{\mathbf{a}'\mathbf{S}\mathbf{a}}$$

Sea $\mathbf{A} = (\overline{\mathbf{x}} - \mu_0)(\overline{\mathbf{x}} - \mu_0)'$ matriz de orden $p \times p$ y rango 1. Si $\mathbf{v}_1$ satisface $\mathbf{A}\mathbf{v}_1 = \lambda_1 \mathbf{S}\mathbf{v}_1$ entonces

$$\lambda_1 = \max_v \frac{\mathbf{v}'\mathbf{A}\mathbf{v}}{\mathbf{v}'\mathbf{S}\mathbf{v}}.$$

De $(\overline{\mathbf{x}} - \mu_0)(\overline{\mathbf{x}} - \mu_0)'\mathbf{v}_1 = \lambda_1 \mathbf{S}\mathbf{v}_1$ resulta que $\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0)(\overline{\mathbf{x}} - \mu_0)'\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$ y de la identidad

$$\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0)(\overline{\mathbf{x}} - \mu_0)'(\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0)) = (\overline{\mathbf{x}} - \mu_0)'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0)(\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0))$$

vemos que $\lambda_1 = (\overline{\mathbf{x}} - \mu_0)'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0), \mathbf{v}_1 = \mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0)$. Por lo tanto

$$T^2 = \max_a t^2(a) = (n-1)(\overline{\mathbf{x}} - \mu_0)'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0). \square$$

3.6 Ejemplos

Exemple 3.6.1

Se desean comparar dos especies de moscas de agua: *Amerohelea fascinata, Amerohelea pseudofascinata.* En relación a las variables X_1 = long. antena, X_2 = long. ala (en mm), para dos muestras de tamaños $n_1 = 9$ y $n_2 = 6$, se han obtenido las matrices de datos de la Tabla 3.1.

Vectores de medias (valores multiplicados por 100):

$$\overline{\mathbf{x}} = (141.33, 180.44), \quad \overline{\mathbf{y}} = (122.67, 192.67).$$

Amerohelea fascinata		*A. pseudofascinata*	
$n_1 = 9$		$n_2 = 6$	
X_1	X_2	X_1	X_2
1.38	1.64	1.14	1.78
1.40	1.70	1.20	1.86
1.24	1.72	1.18	1.96
1.36	1.74	1.30	1.96
1.38	1.82	1.26	2.00
1.48	1.82	1.28	2.00
1.54	1.82		
1.38	1.90		
1.56	2.08		

Tabla 3.1: X_1 = long. antena, X_2 = long. ala (en mm), para dos muestras de tamaño $n_1 = 9$ y $n_2 = 6$,.

Matrices de covarianzas:

$$\mathbf{S}_1 = \begin{pmatrix} 98.00 & 80.83 \\ 80.83 & 167.78 \end{pmatrix} \quad \mathbf{S}_2 = \begin{pmatrix} 39.47 & 43.47 \\ 43.47 & 77.87 \end{pmatrix}.$$

Estimación centrada de la matriz de covarianzas común:

$$\widehat{\mathbf{S}} = \frac{1}{13}(8\mathbf{S}_1 + 5\mathbf{S}_2) = \begin{pmatrix} 75.49 & 66.46 \\ 66.46 & 133.81 \end{pmatrix}.$$

Distancia de Mahalanobis entre las dos muestras:

$$D^2 = (\overline{\mathbf{x}} - \overline{\mathbf{y}})\widehat{\mathbf{S}}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}})' = 15.52.$$

Estadístico T^2 :

$$T^2 = \frac{6 \times 9}{6 + 9} D^2 = 55.87$$

Estadístico F :

$$\frac{9 + 6 - 1 - 2}{2(9 + 6 - 2)} T^2 = 25.78 \sim F^2_{12}$$

Decisión: rechazamos la hipótesis de que las dos especies son iguales (Nivel de significación=0.001).

Exemple 3.6.2

Comparación de las especies *virginica, versicolor, setosa* de flores del género *Iris* (datos de R. A. Fisher, Tabla 3.2), respecto a las variables que miden longitud y ancho de sépalos y pétalos:

$$X_1, X_2 = \text{long., anch.(sépalos)}, X_3, X_4 = \text{long., anch.(pétalos)}.$$

Vectores de medias y tamaños mustrales:

$$\begin{array}{lll} \textit{I. setosa} & (5.006, 3.428, 1.462, 0.246) & n_1 = 50 \\ \textit{I. versicolor} & (5.936, 2.770, 4.260, 1.326) & n_2 = 50 \\ \textit{I. virginica} & (6.588, 2.974, 5.550, 2.026) & n_3 = 50 \end{array}$$

Matriz dispersión entre grupos:

$$\mathbf{B} = \begin{pmatrix} 63.212 & -19.953 & 165.17 & 71.278 \\ & 11.345 & -57.23 & -22.932 \\ & & 436.73 & 186.69 \\ & & & 80.413 \end{pmatrix}$$

Matriz dispersión dentro grupos:

$$\mathbf{W} = \begin{pmatrix} 38.956 & 12.630 & 24.703 & 5.645 \\ & 16.962 & 8.148 & 4.808 \\ & & 27.322 & 6.284 \\ & & & 6.156 \end{pmatrix}$$

Lambda de Wilks:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|} = 0.02344 \sim \Lambda(4, 147, 2)$$

Transformación a una F :

$$\Lambda \to F = 198.95 \sim F^8_{288}$$

Decisión: las diferencias entre las tres especies son muy significativas.

X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Tabla 3.2: Longitud y anchura de sépalos y pétalos de 3 especies del género Iris: Setosa, Versicolor, Virginica.

3.7 Complementos

C. Stein probó que la estimación $\widehat{\mu} = \overline{\mathbf{x}}$ de μ de la distribución $N_p(\mu, \Sigma)$ puede ser inadmisible si $p \geq 3$, en el sentido de que no minimiza

$$\sum_{i=1}^{p} (\widehat{\mu}_i - \mu_i)^2,$$

y propuso una mejora de aquel estimador. B. Efron y C. Morris explicaron esa peculiaridad desde una perspectiva bayesiana. S. M. Stigler dió una interesante explicación en términos de regresión, justificando porqué $p \geq 3$ (consultar Cuadras, 1991).

El principio de unión intersección es debido a S. N. Roy, pero no siempre es aplicable. El test de máxima-verosimilitud es atribuido a S. Wilks y es más general. Es interesante notar que $-2 \log \Lambda$ se puede interpretar como una distancia de Mahalanobis. Otros tests semejantes fueron propuestos por C. R. Rao y A. Wald. Consultar Cuadras y Fortiana (1993b), Rao (1973).

En general, es necesario corregir los tests multiplicando por una constante a fin de conseguir tests insesgados (la potencia del test será siempre más grande que el nivel de significación). Por ejemplo, es necesario hacer la modificación de G. E. P. Box sobre el test de Bartlett para comparar matrices de covarianzas (Sección 7.5.2).

Capítulo 4

ANALISIS DE CORRELACION CANONICA

4.1 Introducción

En este capítulo estudiamos la relación multivariante entre vectores aleatorios. Introducimos y estudiamos las correlaciones canónicas, que son generalizaciones de las correlaciones simple y múltiple.

Tenemos tres posibilidades para relacionar dos variables:

- La correlación simple si X, Y son dos v.a.
- La correlación múltiple si Y es una v.a. y $\mathbf{X} = (X_1, \ldots, X_p)$ es un vector aleatorio.
- La correlación canónica si $\mathbf{X} = (X_1, \ldots, X_p)$ e $\mathbf{Y} = (Y_1, \ldots, Y_q)$ son dos vectores aleatorios.

4.2 Correlación múltiple

Queremos relacionar una variable respuesta Y con p variables cuantitativas explicativas $X_1, \ldots, X_p$, que suponemos centradas. El modelo de regresión múltiple consiste en encontrar la combinación lineal

$$\widehat{Y} = \beta_1 X_1 + \ldots + \beta_p X_p$$

que mejor se ajuste a la variable Y. Sea Σ la matriz de covarianzas de $\mathbf{X}$ y $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)'$ el vector columna con las covarianzas $\delta_j = cov(Y, X_j), j = 1, \ldots, p$. El criterio de ajuste es el de los mínimos cuadrados.

Teorema 4.2.1 *Los coeficientes $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ que minimizan la cantidad $E(Y - \widehat{Y})^2$ verifican la ecuación*

$$\widehat{\boldsymbol{\beta}} = \Sigma^{-1}\boldsymbol{\delta}. \tag{4.1}$$

Demost.:

$$\begin{aligned} \phi(\boldsymbol{\beta}) &= E(Y - \widehat{Y})^2 \\ &= E(Y)^2 + E(\widehat{Y})^2 - 2E(Y\widehat{Y}) \\ &= \mathrm{var}(Y) + \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{\delta} \end{aligned}$$

Derivando vectorialmente respecto de $\boldsymbol{\beta}$ e igualando a 0

$$\frac{\partial}{\partial\beta}\phi(\boldsymbol{\beta}) = 2\Sigma\boldsymbol{\beta} - 2\boldsymbol{\delta} = 0.$$

La variable predicción es $\widehat{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\beta}_1 X_1 + \ldots + \widehat{\beta}_p X_p$. Si ponemos

$$Y = \widehat{Y} + \widetilde{Y},$$

entonces $\widetilde{Y}$ es la variable residual.

La correlación múltiple entre Y y $X_1, \ldots, X_p$ es, por definición, la correlación simple entre Y y la mejor predicción $\widehat{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. Se indica por

$$R = cor(Y, \widehat{Y}).$$

Se verifica:

1. $0 \leq R \leq 1$.
2. $R = 1$ si Y es combinación lineal de $X_1, \ldots, X_p$.
3. $R = 0$ si Y está incorrelacionada con cada una de las variables X_i.

Teorema 4.2.2 *La variable predicción $\widehat{Y}$, residual $\widetilde{Y}$ y la correlación múltiple R cumplen:*

1. *$\widehat{Y}$ e $\widetilde{Y}$ son variables incorrelacionadas.*

2. $var(Y) = var(\widehat{Y}) + var(\widetilde{Y})$.

3. $R^2 = var(\widehat{Y})/var(Y)$.

Demost.: 1) es consecuencia de $\Sigma\widehat{\boldsymbol{\beta}} = \boldsymbol{\delta}$. En efecto,

$$\begin{aligned}\mathrm{cov}(\widehat{Y}, \widetilde{Y}) &= E(\widehat{Y}\widetilde{Y}) = E(\widehat{\boldsymbol{\beta}}'\mathbf{X}'(Y - \widehat{\boldsymbol{\beta}}'\mathbf{X})) \\ &= \widehat{\boldsymbol{\beta}}'\boldsymbol{\delta} - \widehat{\boldsymbol{\beta}}'\Sigma\widehat{\boldsymbol{\beta}} = 0.\end{aligned}$$

2) es consecuencia inmediata de 1). Finalmente, de

$$\mathrm{cov}(Y, \widehat{Y}) = \mathrm{cov}(Y, \Sigma_{i=1}^p \widehat{\beta}_i X_i) = \Sigma_{i=1}^p \widehat{\beta}_i \delta_i = \widehat{\boldsymbol{\beta}}'\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}}'\Sigma\widehat{\boldsymbol{\beta}} = \mathrm{var}(\widehat{Y}),$$

obtenemos

$$R^2 = \frac{\mathrm{cov}^2(Y, \widehat{Y})}{\mathrm{var}(Y)\mathrm{var}(\widehat{Y})} = \frac{\mathrm{var}(\widehat{Y})}{\mathrm{var}(Y)}. \tag{4.2}$$

4.3 Correlación canónica

Sean $\mathbf{X} = (X_1, \ldots, X_p)$, $\mathbf{Y} = (Y_1, \ldots, Y_q)$ dos vectores aleatorios de dimensiones p y q. Planteemos el problema de encontrar dos variables compuestas

$$U = \mathbf{Xa} = a_1X_1 + \ldots + a_pX_p, \quad V = \mathbf{Yb} = b_1Y_1 + \ldots + b_pY_q,$$

siendo $\mathbf{a} = (a_1, \ldots, a_p)', \mathbf{b} = (b_1, \ldots, b_p)'$ tales que la correlación entre ambas

$$cor(U, V)$$

sea máxima. Indicamos por $\mathbf{S}_{11}, \mathbf{S}_{22}$ las matrices de covarianzas (muestrales) de las variables $\mathbf{X}, \mathbf{Y}$, respectivamente, y sea $\mathbf{S}_{12}$ la matriz $p \times q$ con las covarianzas de las variables $\mathbf{X}$ con las variables $\mathbf{Y}$. Es decir:

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

donde $\mathbf{S}_{21} = \mathbf{S}_{12}'$.

Podemos suponer

$$\mathrm{var}(U) = \mathbf{a}'\mathbf{S}_{11}\mathbf{a} = 1, \quad \mathrm{var}(V) = \mathbf{b}'\mathbf{S}_{22}\mathbf{b} = 1.$$

Así el problema se reduce a:

$$\text{maximizar} \quad \mathbf{a}'\mathbf{S}_{12}\mathbf{b} \quad \text{restringido a} \quad \mathbf{a}'\mathbf{S}_{11}\mathbf{a} = \mathbf{b}'\mathbf{S}_{22}\mathbf{b} = 1.$$

Los vectores de coeficientes $\mathbf{a}, \mathbf{b}$ que cumplen esta condición son los primeros vectores canónicos. La máxima correlación entre U, V es la primera correlación canónica r_1.

Teorema 4.3.1 *Los primeros vectores canónicos satisfacen las ecuaciones*

$$\begin{aligned} \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a} &= \lambda\mathbf{S}_{11}\mathbf{a}, \\ \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b} &= \lambda\mathbf{S}_{22}\mathbf{b}. \end{aligned} \tag{4.3}$$

Demost.: Consideremos la función

$$\phi(\mathbf{a}, \mathbf{b}) = \mathbf{a}'\mathbf{S}_{12}\mathbf{b} - \frac{\lambda}{2}(\mathbf{a}'\mathbf{S}_{11}\mathbf{a} - 1) - \frac{\mu}{2}(\mathbf{b}'\mathbf{S}_{22}\mathbf{b} - 1),$$

donde λ, μ son multiplicadores de Lagrange. Entonces de $\partial\phi/\partial\mathbf{a} = \partial\phi/\partial\mathbf{b} = \mathbf{0}$ obtenemos las dos ecuaciones

$$\mathbf{S}_{12}\mathbf{b} - \lambda\mathbf{S}_{11}\mathbf{a} = \mathbf{0}, \quad \mathbf{S}_{21}\mathbf{a} - \mu\mathbf{S}_{22}\mathbf{b} = \mathbf{0}. \tag{4.4}$$

Multiplicando la primera por $\mathbf{a}'$ y la segunda por $\mathbf{b}'$, tenemos

$$\mathbf{a}'\mathbf{S}_{12}\mathbf{b} = \lambda\mathbf{a}'\mathbf{S}_{11}\mathbf{a}, \quad \mathbf{b}'\mathbf{S}_{21}\mathbf{a} = \mu\mathbf{b}'\mathbf{S}_{22}\mathbf{b},$$

que implican $\lambda = \mu$. Así pues, de la segunda ecuación en (4.4), $\mathbf{b} = \lambda^{-1}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}$, y substituyendo en la primera obtenemos $\lambda^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a} - \lambda\mathbf{S}_{11}\mathbf{a} = \mathbf{0}$. Prescindiendo de λ^{-1}, pues es un factor multiplicativo arbitrario, y operando análogamente con la otra ecuación, obtenemos (4.3).

Teorema 4.3.2 *Los vectores canónicos normalizados por* $\mathbf{a}'\mathbf{S}_{11}\mathbf{a} = \mathbf{b}'\mathbf{S}_{22}\mathbf{b} = 1$, *están relacionados por*

$$\begin{aligned} \mathbf{a} &= \lambda^{-1/2}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}, \\ \mathbf{b} &= \lambda^{-1/2}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}, \end{aligned}$$

y la primera correlación canónica es $r_1 = \sqrt{\lambda_1}$, *donde* λ_1 *es el primer valor propio de* $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$.

Demost.: Tenemos de (4.4) que $\mathbf{a} = \alpha \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}$, donde α es una constante a determinar. Partimos de que $\mathbf{a}'\mathbf{S}_{11}\mathbf{a} = 1$ y para $\alpha = \lambda^{-1/2}$ resulta que:

$$\begin{aligned}\mathbf{a}'\mathbf{S}_{11}\mathbf{a} &= \lambda^{-1/2}\mathbf{a}'\mathbf{S}_{11}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b} \\ &= \lambda^{-1/2}\mathbf{a}'\mathbf{S}_{12}\mathbf{b} \\ &= \lambda^{-1/2}\lambda^{-1/2}\mathbf{a}'\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a} \\ &= \lambda^{-1}\lambda\mathbf{a}'\mathbf{S}_{11}\mathbf{a} \\ &= 1\end{aligned}$$

La correlación es $r_1 = \mathbf{a}'\mathbf{S}_{12}\mathbf{b}$ y como $1 = \lambda^{-1/2}\mathbf{a}'\mathbf{S}_{12}\mathbf{b}$ deducimos que $r_1^2 = \lambda_1$.

De hecho, las ecuaciones en valores y vectores propios tienen otras soluciones. Concretamente hay $m = \min\{p, q\}$ parejas de vectores canónicos $\mathbf{a}_1, \mathbf{b}_1, \ldots, \mathbf{a}_m, \mathbf{b}_m$, que proporcionan las variables y correlaciones canónicas

$$\begin{array}{lll} U_1 = \mathbf{X}\mathbf{a}_1, & V_1 = \mathbf{Y}\mathbf{b}_1, & r_1 = \mathrm{cor}(U_1, V_1), \\ U_2 = \mathbf{X}\mathbf{a}_2, , & V_2 = \mathbf{Y}\mathbf{b}_2, & r_2 = \mathrm{cor}(U_2, V_2), \\ \quad \vdots & \quad \vdots & \quad \vdots \\ U_m = \mathbf{X}\mathbf{a}_m, & V_m = \mathbf{Y}\mathbf{b}_m, & r_m = \mathrm{cor}(U_m, V_m). \end{array}$$

Teorema 4.3.3 *Supongamos $r_1 > r_2 > \ldots > r_m$. Entonces:*

1. *Tanto las variables canónicas $U_1, \ldots, U_m$ como las variables canónicas $V_1, \ldots, V_m$ están incorrelacionadas.*

2. *La primera correlación canónica $r_1 = cor(U_1, V_1)$ es la máxima correlación entre una combinación lineal de $\mathbf{X}$ y una combinación lineal de $\mathbf{Y}$.*

3. *La segunda correlación canónica $r_2 = cor(U_2, V_2)$ es la máxima correlación entre las combinaciones lineales de $\mathbf{X}$ incorrelacionadas con U_1 y las combinaciones lineales de $\mathbf{Y}$ incorrelacionadas con V_1.*

4. *$cor(U_i, V_j) = 0$ si $i \neq j$.*

Demost.: Sea $i \neq j$. Expresando (4.3) para $\mathbf{a}_k, \lambda_k$, $k = i, j$, y multiplicando por $\mathbf{a}_j'$ y por $\mathbf{a}_i'$ tenemos que

$$\begin{aligned}\mathbf{a}_j'\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_i &= \lambda_i\mathbf{a}_j'\mathbf{S}_{11}\mathbf{a}_i, \\ \mathbf{a}_i'\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_j &= \lambda_j\mathbf{a}_i'\mathbf{S}_{11}\mathbf{a}_j.\end{aligned}$$

Restando: $(\lambda_i - \lambda_j)\mathbf{a}_i'\mathbf{S}_{11}\mathbf{a}_j = 0 \Rightarrow \mathbf{a}_i'\mathbf{S}_{11}\mathbf{a}_j = 0 \Rightarrow cor(U_i, U_j) = 0$.

Por otra parte, expresando (4.3) como

$$\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a} = \lambda_i\mathbf{a}_i, \quad \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}_j = \lambda_j\mathbf{b}_j,$$

y multiplicando por $\mathbf{b}_j'\mathbf{S}_{21}$ y por $\mathbf{a}_i'\mathbf{S}_{12}$ llegamos a

$$\begin{aligned}\mathbf{b}_j'\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_i &= \lambda_i\mathbf{b}_j'\mathbf{S}_{21}\mathbf{a}_i,\\ \mathbf{a}_i'\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}_j &= \lambda_j\mathbf{a}_i'\mathbf{S}_{12}\mathbf{b}_j.\end{aligned}$$

Restando: $(\lambda_i - \lambda_j)\mathbf{a}_i'\mathbf{S}_{12}\mathbf{b}_j = 0 \Rightarrow \mathbf{a}_i'\mathbf{S}_{12}\mathbf{b}_j = 0 \Rightarrow cor(U_i, V_j) = 0$.

4.4 Correlación canónica y descomposición singular

Podemos formular una expresión conjunta para los vectores canónicos utilizando la descomposición singular de una matriz. Supongamos $p \geq q$, consideremos la matriz $p \times q$

$$\mathbf{Q} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}$$

y hallemos

$$\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}',$$

la descomposición singular de $\mathbf{Q}$, donde $\mathbf{U}$ es una matriz $p \times q$ con columnas ortonormales, $\mathbf{V}$ es una matriz $q \times q$ ortogonal, y $\mathbf{\Lambda}$ es una matriz diagonal con los valores singulares de $\mathbf{Q}$. Es decir, $\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \mathbf{V}'\mathbf{V} = \mathbf{V}'\mathbf{V} = \mathbf{I}_q$, $\mathbf{\Lambda}$ =diag$(\lambda_1, \ldots, \lambda_p)$.

Teorema 4.4.1 *Los vectores canónicos y correlaciones canónicas son*

$$\mathbf{a}_i = \mathbf{S}_{11}^{-1/2}\mathbf{u}_i, \quad \mathbf{b}_i = \mathbf{S}_{22}^{-1/2}\mathbf{v}_i, \quad r_i = \lambda_i.$$

Demost.:

$$\mathbf{Q}\mathbf{Q}' = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}\mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2} = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}'$$

y por lo tanto

$$\mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}\mathbf{u}_i = \lambda_i^2\mathbf{u}_i$$

Multiplicando por $\mathbf{S}_{11}^{-1/2}$

$$\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}(\mathbf{S}_{11}^{-1/2}\mathbf{u}_i) = \lambda_i^2(\mathbf{S}_{11}^{-1/2}\mathbf{u}_i)$$

y comparando con resultados anteriores, queda probado el teorema.□

4.5 Significación de las correlaciones canónicas

Hemos encontrado las variables y correlaciones canónicas a partir de las matrices de covarianzas y correlaciones muestrales, es decir, a partir de muestras de tamaño n. Naturalmente, todo lo que hemos dicho vale si sustituimos $\mathbf{S}_{11}, \mathbf{S}_{12}, \mathbf{S}_{22}$ por las versiones poblacionales $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}$. Sean

$$\rho_1 \geq \rho_2 \geq \cdots \geq \rho_m$$

las $m = \min\{p, q\}$ correlaciones canónicas obtenidas a partir de $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}$, soluciones de:

$$|\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \rho^2\Sigma_{11}| = 0.$$

Si queremos decidir cuáles son significativas, supongamos normalidad multivariante, indiquemos $\rho_0 = 1$ y planteemos el test

$$H_0^k : \rho_k > \rho_{k+1} = \cdots = \rho_m = 0, \quad (k = 0, 1, \ldots, m),$$

que equivale a $\text{rang}(\Sigma_{22}^{-1}\Sigma_{21}) = k$. El test de Bartlett-Lawley demuestra que si H_0^k es cierta, entonces

$$L_k = -[n - 1 - k - \frac{1}{2}(p + q + 1) + \sum_{i=1}^{k} r_i^{-2}] \log[\prod_{i=k+1}^{m} (1 - r_i^2)$$

es asintóticamente ji-cuadrado con $(m - k)(p - k)$ g.l. Este test se aplica secuencialmente: si L_i es significativo para $i = 0, 1, \ldots, k - 1$, pero L_k no es significativo, entonces se acepta H_0^k.

4.6 Test de independencia

Suponiendo normalidad, afirmar que $\mathbf{X}$ es independiente de $\mathbf{Y}$ consiste en plantear

$$H_0 : \Sigma_{12} = \mathbf{0}, \quad H_1 : \Sigma_{12} \neq \mathbf{0}.$$

Podemos resolver este test de hipótesis de dos maneras.

4.6.1 Razón de verosimilitud

Si la hipótesis es cierta, entonces el test de razón de verosimilitud (Sección 3.5.1) se reduce al estadístico

$$\Lambda = \frac{|\mathbf{S}|}{|\mathbf{S}_{11}||\mathbf{S}_{22}|} = \frac{|\mathbf{R}|}{|\mathbf{R}_{11}||\mathbf{R}_{22}|},$$

que sigue la distribución lambda de Wilks $\Lambda(p, n-1-q, q)$, equivalente a $\Lambda(q, n-1-p, q)$. Rechazaremos H_0 si Λ es pequeña y significativa (Mardia *et al.* 1979, Rencher, 1998).

Es fácil probar que Λ es función de las correlaciones canónicas

$$\Lambda = |\mathbf{I} - \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}| = \prod_{i=1}^{m}(1 - r_i^2).$$

4.6.2 Principio de unión intersección

Consideremos las variables $U = a_1X_1 + \ldots + a_pX_p$,$V = b_1Y_1 + \ldots + b_pY_q$. La correlación entre U, V es

$$\rho(U, V) = \frac{\mathbf{a}_{12}'\Sigma_{12}\mathbf{b}}{\sqrt{\mathbf{a}\Sigma_{11}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{22}\mathbf{b}}}$$

H_0 equivale a $\rho(U, V) = 0$ para todo U, V. La correlación muestral es

$$r(U, V) = \frac{\mathbf{a}'\mathbf{S}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\mathbf{S}_{11}\mathbf{a}}\sqrt{\mathbf{b}'\mathbf{S}_{22}\mathbf{b}}}.$$

Aplicando el principio de unión intersección (Sección 3.5.2), aceptaremos H_0 si $r(U, V)$ no es significativa para todo U, V, y aceptaremos H_1 si $r(U, V)$ es significativa para algún par U, V. Este criterio nos lleva a estudiar la significación de

$$r_1 = \max_{U,V} r(U, V)$$

es decir, de la primera correlación canónica. Por tanto, el test es:

$$H_0 : \rho_1 = 0, \quad H_1 : \rho_1 > 0.$$

Existen tablas especiales para decidir si r_1 es significativa (Morrison, 1976), pero también se puede aplicar el estadístico L_0 de Bartlett-Lawley.

4.7 Un ejemplo

Ejemplo 1. Se consideran $n = 25$ familias y las variables:

$$\begin{array}{llll} X_1 = & \text{long. cabeza primer hijo}, & X_2 = & \text{ancho cabeza primer hijo}, \\ Y_1 = & \text{long. cabeza segundo hijo}, & Y_2 = & \text{ancho cabeza segundo hijo}, \end{array}$$

La matriz de correlaciones es:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.7346 & 0.7108 & 0.7040 \\ 0.7346 & 1.0000 & 0.6932 & 0.8086 \\ 0.7108 & 0.6932 & 1.0000 & 0.8392 \\ 0.7040 & 0.8086 & 0.8392 & 1.0000 \end{pmatrix}$$

Entonces:

$$\mathbf{R}_{11} = \begin{pmatrix} 1.0000 & 0.7346 \\ 0.7346 & 1.0000 \end{pmatrix}, \quad \mathbf{R}_{12} = \begin{pmatrix} 0.7108 & 0.7040 \\ 0.6932 & 0.8086 \end{pmatrix},$$
$$\mathbf{R}_{22} = \begin{pmatrix} 1.0000 & 0.8392 \\ 0.8392 & 1.0000 \end{pmatrix}.$$

Las raíces de la ecuación:

$$|\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} - \lambda\mathbf{R}_{11}| = 0.460363\lambda^2 - 0.287596\lambda + 0.000830 = 0$$

son: $\lambda_1 = 0.6218$, $\lambda_2 = 0.0029$, y por tanto las correlaciones canónicas son:

$$r_1 = 0.7885, \quad r_2 = 0.0539.$$

Los vectores canónicos normalizados son:

$$\begin{array}{llll} \mathbf{a}_1 = & (0.0566, 0.0707)', & \mathbf{a}_2 = & (0.1400, -0.1870)', \\ \mathbf{b}_1 = & (0.0502, 0.0802)', & \mathbf{b}_2 = & (0.1760, -0.2619)'. \end{array}$$

Las variables canónicas con variaza 1 son:

$$\begin{array}{llll} U_1 = & 0.0566X_1 + 0.0707X_2, & V_1 = & 0.0502Y_1 + 0.0802Y_2, \quad (r_1 = 0.7885), \\ U_2 = & 0.1400X_1 - 0.1870X_2, & V_2 = & 0.1760Y_1 - 0.2619Y_2, \quad (r_2 = 0.0539). \end{array}$$

La dependencia entre (X_1, X_2) y (Y_1, Y_2) viene dada principalmente por la relación entre (U_1, V_1) con correlación 0.7885, más alta que cualquier correlación entre una variable X_i y una variable Y_j. Podemos interpretar las

primeras variables canónicas como un factor de "tamaño" de la cabeza y las segundas como un factor de "forma". Habría entonces una notable relación en el tamaño y una escasa relación en la forma de la cabeza.

El test de independencia entre (X_1, X_2) y (Y_1, Y_2) da

$$\Lambda = \frac{|\mathbf{R}|}{|\mathbf{R}_{11}||\mathbf{R}_{22}|} = 0.3771 \sim \Lambda(2, 22, 2)$$

que, según (2.8), transformamos con una F obteniendo 6.60 con 4 y 42 g.l. Rechazamos la hipótesis de independencia.

La prueba de significación de las correlaciones canónicas dá:

$$\begin{array}{ll} H_0^0 : \rho_0 = 1 > \rho_1 = \rho_2 = 0, & L_0 = 22.1 \quad (4 \text{ g.l.}), \\ H_0^1 : \rho_1 > \rho_2 = 0, & L_1 = 1.22 \quad (2 \text{ g.l.}). \end{array}$$

Podemos rechazar H_0^0 y aceptar H_0^1. Solamente la primera correlación canónica es significativa.

Ejemplo 2. Se consideran los resultados de unas elecciones celebradas en las 41 comarcas catalanas y para cada comarca se tabulan los valores de las siguientes variables:

$$\begin{array}{ll} X_1 = \log(\text{porcentaje de votos a CU}), & X_2 = \log(\text{porcentaje de votos a PSC}), \\ X_3 = \log(\text{porcentaje de votos a PP}), & X_4 = \log(\text{porcentaje de votos a ERC}), \\ Y_1 = \log(\text{cociente Juan/Joan}), & Y_2 = \log(\text{cociente Juana/Joana}), \end{array}$$

donde "cociente Juan/Joan" significa el resultado de dividir el número de hombres que se llaman Juan por el número de hombres que se llaman Joan. Valores positivos de las variables Y_1, Y_2 en una comarca indican predominio de los nombres en castellano sobre los nombres en catalán.

La matriz de correlaciones es:

	X_1	X_2	X_3	X_4	Y_1	Y_2
X_1	1	−.8520	−.6536	−.5478	−.6404	−.5907
X_2		1	.5127	−.7101	.7555	.6393
X_3			1	−.6265	.5912	.5146
X_4				1	−.7528	−.7448
Y_1					1	.8027
Y_2						1

Sólo hay 2 correlaciones canónicas:

$$r_1 = 0.8377, \; r_2 = 0.4125.$$

Las variables canónicas son:

$$\begin{array}{lll} U_1 = & +0.083X_1 - 0.372X_2 - 0.1130X_3 + 0.555X_4, & (r_1 = 0.8377), \\ V_1 = & +0.706Y_1 + 0.339Y_2, & \\ U_2 = & +1.928X_1 + 2.4031.546X_2 + 1.127X_3 + 1.546X_4, & (r_2 = 0.4125). \\ V_2 = & +1.521Y_1 - 1.642Y_2, & \end{array}$$

Las primeras variables canónicas U_1, V_1, que podemos escribir convencionalmente como

$$\begin{aligned} U_1 &= +0.083\text{CU} - 0.372\text{PSC} - 0.1130\text{PP} + 0.555\text{ERC}, \\ V_1 &= +0.706(\text{Juan/Joan}) + 0.339(\text{Juana/Joanna}), \end{aligned}$$

nos indican que las regione más catalanas, en el sentido de que los nombres castellanos Juan y Juana no predominan tanto sobre los catalanes Joan y Joanna, tienden a votar más a CU y ERC, que son partidos más nacionalistas. Las regiones que votan más al PSC y al PP, que son partidos más centralistas, están en general, más castellanizadas. Las segundas variables canónicas tienen una interpretación más dificil.

4.8 Complementos

El análisis de correlación canónica (ACC) fué introducido por H. Hotelling en 1935, que buscaba la relación entre tests mentales y medidas biométricas, a fin de estudiar el número y la naturaleza de las relaciones entre mente y cuerpo, que con un análisis de todas las correlaciones sería difícil de interpretar. Es un método de aplicación limitada, pero de gran interés teórico puesto que diversos métodos de AM se derivan del ACC.

Aplicaciones a la psicología se pueden encontrar en Cooley y Lohnes (1971), Cuadras y Sánchez (1975). En ecología se ha aplicado como un modelo para estudiar la relación entre presencia de especies y variables ambientales (Gittings, 1985).

La distribución de las correlaciones canónicas es bastante complicada. Solamente se conocen resultados asintóticos (Muirhead, 1982).

Si $f(x, y)$ es la densidad de dos v.a. X, Y, tiene interés en estadística el concepto de máxima correlación (propuesto por H. Gabelein) que se define como

$$\rho_1 = \sup_{\alpha,\beta} cor(\alpha(X), \beta(Y)),$$

donde $\alpha(X), \beta(Y)$ son funciones con varianza finita. Entonces $\rho_1 = 0$ si X, Y son variables independientes. Podemos ver a ρ_1 como la primera correlación canónica, $\alpha_1(X), \beta_1(Y)$ como las primeras variables canónicas y definir las sucesivas correlaciones canónicas. Sin embargo el cálculo de ρ_1 puede ser complicado (Cuadras, 2002a). Lancaster (1969) estudia estas correlaciones y demuestra que $f(x, y)$ se puede desarrollar en serie a partir de las correlaciones y funciones canónicas. Diversos autores han estudiado la estimación de las primeras funciones canónicas, como una forma de predecir una variable en función de la otra (Hastie y Tibshirani, 1990).

Capítulo 5

ANALISIS DE COMPONENTES PRINCIPALES

5.1 Definición y obtención de las componentes principales

Sea $\mathbf{X} = [X_1, \ldots, X_p]$ una matriz de datos multivariantes. Lo que sigue también vale si $\mathbf{X}$ es un vector formado por p variables observables.

Las componentes principales son unas variables compuestas incorrelacionadas tales que unas pocas explican la mayor parte de la variabilidad de $\mathbf{X}$.

Definició 5.1.1 *Las componentes principales son las variables compuestas*

$$Y_1 = \mathbf{X}\mathbf{t}_1, Y_2 = \mathbf{X}\mathbf{t}_2, \ldots, Y_p = \mathbf{X}\mathbf{t}_p$$

tales que:

1. *var*(Y_1) *es máxima condicionado a* $\mathbf{t}_1'\mathbf{t}_1 = 1$.
2. *Entre todas las variables compuestas* Y *tales que* $cov(Y_1, Y) = 0$, *la variable* Y_2 *es tal que var*(Y_2) *es máxima condicionado a* $\mathbf{t}_2'\mathbf{t}_2 = 1$.
3. Y_3 *es una variable incorrelacionada con* Y_1, Y_2 *con varianza máxima. Análogamente definimos las demás componentes principales.*

Si $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p]$ es la matriz $p \times p$ cuyas columnas son los vectores que definen las componentes principales, entonces la transformación lineal $\mathbf{X} \to \mathbf{Y}$

$$\mathbf{Y} = \mathbf{XT} \tag{5.1}$$

se llama transformación por componentes principales.

Teorema 5.1.1 *Sean* $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p$ *los* p *vectores propios normalizados de la matriz de covarianzas* $\mathbf{S}$, *es decir,*

$$\mathbf{St}_i = \lambda_i \mathbf{t}_i, \quad \mathbf{t}_i'\mathbf{t}_i = 1, \quad i = 1, \ldots, p.$$

Entonces:

1. *Las variables compuestas* $Y_i = \mathbf{Xt}_i$, $i = 1, \ldots, p$, *son las componentes principales.*

2. *Las varianzas son los valores propios de* $\mathbf{S}$

$$\mathrm{var}(Y_i) = \lambda_i, \quad i = 1, \ldots, p.$$

3. *Las componentes principales son variables incorrelacionadas:*

$$\mathrm{cov}(Y_i, Y_j) = 0, \quad i \neq j = 1, \ldots, p.$$

Demost.: Supongamos $\lambda_1 > \cdots > \lambda_p > 0$. Probemos que las variables $Y_i = \mathbf{Xt}_i$, $i = 1, \ldots, p$, son incorrelacionadas:

$$\begin{array}{lll} \mathrm{cov}(Y_i, Y_j) & = \mathbf{t}_i'\mathbf{St}_j & = \mathbf{t}_i'\lambda_j\mathbf{t}_j \quad = \lambda_j\mathbf{t}_i'\mathbf{t}_j, \\ \mathrm{cov}(Y_j, Y_i) & = \mathbf{t}_j'\mathbf{St}_i & = \mathbf{t}_j'\lambda_j\mathbf{t}_i \quad = \lambda_i\mathbf{t}_j'\mathbf{t}_i, \end{array}$$

$\Rightarrow (\lambda_j - \lambda_i)\mathbf{t}_i'\mathbf{t}_j = 0, \Rightarrow \mathbf{t}_i'\mathbf{t}_j = 0, \Rightarrow \mathrm{cov}(Y_i, Y_j) = \lambda_j\mathbf{t}_i'\mathbf{t}_j = 0$, si $i \neq j$.
Además:

$$\mathrm{var}(Y_i) = \lambda_i\mathbf{t}_i'\mathbf{t}_j = \lambda_i.$$

Sea ahora $Y = \sum_{i=1}^p a_i X_i = \sum_{i=1}^p \alpha_i Y_i$ una variable compuesta tal que $\sum_{i=1}^p \alpha_i^2 = 1$. Entonces

$$\mathrm{var}(Y) = \mathrm{var}(\sum_{i=1}^p \alpha_i Y_i) = \sum_{i=1}^p \alpha_i^2 \mathrm{var}(Y_i) = \sum_{i=1}^p \alpha_i^2\lambda_i \leq (\sum_{i=1}^p \alpha_i^2)\lambda_1 = \mathrm{var}(Y_1),$$

que prueba que Y_1 tiene varianza máxima.

Consideremos ahora las variables Y incorrelacionadas con Y_1. Las podemos expresar como:

$$Y = \sum_{i=1}^{p} b_i X_i = \sum_{i=2}^{p} \beta_i Y_i \text{ condicionado a } \sum_{i=2}^{p} \beta_i^2 = 1.$$

Entonces:

$$\text{var}(Y) = \text{var}(\sum_{i=2}^{p} \beta_i Y_i) = \sum_{i=2}^{p} \beta_i^2 \text{var}(Y_i) = \sum_{i=2}^{p} \beta_i^2 \lambda_i \leq (\sum_{i=2}^{p} \beta_i^2)\lambda_2 = \text{var}(Y_2),$$

y por lo tanto Y_2 está incorrelacionada con Y_1 y tiene varianza máxima. Si $p \geq 3$, la demostración de que $Y_3, \ldots, Y_p$ son también componentes principales es análoga.□

5.2 Variabilidad explicada por las componentes principales

La varianza de la componente principal Y_i es $\text{var}(Y_i) = \lambda_i$ y la variación total es $\text{tr}(\mathbf{S}) = \sum_{i=1}^{p} \lambda_i$. Por lo tanto:

1. Y_i contribuye con la cantidad λ_i a la variación total $\text{tr}(\mathbf{S})$.

2. Si $q < p$, $Y_1, \ldots, Y_q$ contribuyen con la cantidad $\sum_{i=1}^{q} \lambda_i$ a la variación total $\text{tr}(\mathbf{S})$.

3. El porcentaje de variabilidad explicada por las m primeras componentes principales es
$$P_m = 100 \frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}. \tag{5.2}$$

En las aplicaciones cabe esperar que las primeras componentes expliquen un elevado porcentaje de la variabilidad total. Por ejemplo, si $m = 2 < p$, y $P_2 = 90\%$, las dos primeras componentes explican una gran parte de la variabilidad de las variables. Entonces podremos sustituir $X_1, X_2, \ldots, X_p$ por las componentes principales Y_1, Y_2. En muchas aplicaciones, tales componentes tienen interpretación experimental.

5.3 Representación de una matriz de datos

Sea $\mathbf{X} = [X_1, \ldots, X_p]$ una matriz $n \times p$ de datos multivariantes. Queremos representar, en un espacio de dimensión reducida m (por ejemplo, $m = 2$), las filas $\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'$ de $\mathbf{X}$. Necesitamos introducir una distancia (ver Sección 1.9).

Definició 5.3.1 *La distancia euclídea (al cuadrado) entre dos filas de* $\mathbf{X}$

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}), \quad \mathbf{x}_j = (x_{j1}, \ldots, x_{jp}),$$

es

$$\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2.$$

La matriz $\Delta = (\delta_{ij})$ *es la matriz* $n \times n$ *de distancias entre las filas.*

Podemos representar las n filas de $\mathbf{X}$ como n puntos en el espacio R^p distanciados de acuerdo con la métrica δ_{ij}. Pero si p es grande, esta representación no se puede visualizar. Necesitamos reducir la dimensión.

Definició 5.3.2 *La variabilidad geométrica de la matriz de distancias* Δ *es la media de sus elementos al cuadrado*

$$V_\delta(\mathbf{X}) = \frac{1}{2n^2} \sum_{i,j=1}^{n} \delta_{ij}^2.$$

Si $\mathbf{Y} = \mathbf{XT}$ *es una transformación lineal de* $\mathbf{X}$*, donde* $\mathbf{T}$ *es una matriz* $p \times q$ *de constantes,*

$$\delta_{ij}^2(q) = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = \sum_{h=1}^{q} (y_{ih} - y_{jh})^2$$

es la distancia euclídea entre dos filas de $\mathbf{Y}$*. La variabilidad geométrica en dimensión* $q \leq p$ *es*

$$V_\delta(\mathbf{Y})_q = \frac{1}{2n^2} \sum_{i,j=1}^{n} \delta_{ij}^2(q).$$

Teorema 5.3.1 *La variabilidad geométrica de la distancia euclídea es la traza de la matriz de covarianzas*

$$V_\delta(\mathbf{X}) = tr(\mathbf{S}) = \sum_{h=1}^{p} \lambda_h.$$

Demost.: Si $x_1, \ldots, x_n$ es una muestra univariante con varianza s^2, entonces

$$\frac{1}{2n^2} \sum_{i,j=1}^{n} (x_i - x_j)^2 = s^2. \tag{5.3}$$

En efecto, si $\overline{x}$ es la media

$$\begin{aligned} \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}(x_i - x_j)^2 &= \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}(x_i - \overline{x} - (x_j - \overline{x}))^2 \\ &= \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}(x_i - \overline{x})^2 + \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}(x_j - \overline{x})^2 \\ &\quad + \tfrac{2}{n^2}\textstyle\sum_{i,j=1}^{n}(x_i - \overline{x})(x_j - \overline{x}))^2 \\ &= \tfrac{1}{n} n s^2 + \tfrac{1}{n} n s^2 + 0 = 2s^2. \end{aligned}$$

Aplicando (5.3) a cada columna de $\mathbf{X}$ y sumando obtenemos

$$V_\delta(\mathbf{X}) = \sum_{j=1}^{p} s_{jj} = \mathrm{tr}(\mathbf{S}).\square$$

Una buena representación en dimensión reducida q (por ejemplo, $q = 2$) será aquella que tenga máxima variabilidad geométrica, a fin de que los puntos estén lo más separados posible.

Teorema 5.3.2 *La transformación lineal* $\mathbf{T}$ *que maximiza la variabilidad geométrica en dimensión* q *es la transformación por componentes principales (5.1), es decir,* $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_q]$ *contiene los* q *primeros vectores propios normalizados de* $\mathbf{S}$.

Demost.: Aplicando (5.3), la variabilidad geométrica de $\mathbf{Y} = \mathbf{XT}$, donde $\mathbf{T}$ es cualquiera, es

$$V_\delta(\mathbf{Y})_q = \sum_{j=1}^{p} s^2(Y_j) = \sum_{j=1}^{p} \mathbf{t}_j' \mathbf{S} \mathbf{t}_j,$$

siendo $s^2(Y_j) = \mathbf{t}_j' \mathbf{S} \mathbf{t}_j$ la varianza de la variable compuesta Y_j. Alcanzamos la máxima varianza cuando Y_j es una componente principal: $s^2(Y_j) \leq \lambda_j$. Así:

$$\max V_\delta(\mathbf{Y})_q = \sum_{j=1}^{p} \lambda_j . \square$$

El porcentaje de variabilidad geométrica explicada por $\mathbf{Y}$ es

$$P_q = 100 \frac{V_\delta(\mathbf{Y})_q}{V_\delta(\mathbf{X})_p} = 100 \frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_p}.$$

Supongamos ahora $q = 2$. Si aplicamos la transformación (5.1), la matriz de datos $\mathbf{X}$ se reduce a

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{i1} & y_{i2} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix}.$$

Entonces, representando los puntos de coordenadas $(y_{i1}, y_{i2}), i = 1, \ldots, n$, obtenemos una representación óptima en dimensión 2 de las filas de $\mathbf{X}$.

5.4 Inferencia

Hemos planteado el ACP sobre la matriz $\mathbf{S}$, pero lo podemos también plantear sobre la matriz de covarianzas poblacionales Σ. Las componentes principales obtenidas sobre $\mathbf{S}$ son, en realidad, estimaciones de las componentes principales sobre Σ.

Sea $\mathbf{X}$ matriz de datos $n \times p$ donde las filas son independientes con distribución $N_p(\mu, \Sigma)$. Recordemos que:

1. $\overline{\mathbf{x}}$ es $N_p(\mu, \Sigma/n)$.

2. $\mathbf{U} = n\mathbf{S}$ es Wishart $W_p(\Sigma, n-1)$.

3. $\overline{\mathbf{x}}$ y $\mathbf{S}$ son estocásticamente independientes.

Sea $\Sigma = \Gamma\Lambda\Gamma'$ la diagonalización de Σ. Indiquemos

$$\Gamma = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p], \quad \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p], \quad \Lambda = diag(\lambda_1, \ldots, \lambda_p),$$

los vectores propios y valores propios de Σ. Por otra parte, sea $\mathbf{S} = \mathbf{GLG}'$ la diagonalización de $\mathbf{S}$. Indiquemos:

$$\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_p], \quad \mathbf{l} = [l_1, \ldots, l_p], \quad \mathbf{L} = \text{diag}(l_1, \ldots, l_p)$$

los vectores propios y valores propios de $\mathbf{S}$. A partir de ahora supondremos

$$\lambda_1 \geq \ldots \geq \lambda_p.$$

5.4.1 Estimación y distribución asintótica

Teorema 5.4.1 *Se verifica:*

1. *Si los valores propios son diferentes, los valores y vectores propios obtenidos a partir de* $\mathbf{S}$ *son estimadores máximo-verosímiles de los obtenidos a partir de* Σ

$$\widehat{\lambda}_i = l_i, \quad \widehat{\boldsymbol{\gamma}}_i = \mathbf{g}_i \quad , i = 1, \ldots, p.$$

2. *Cuando* $k > 1$ *valores propios son iguales a* λ

$$\lambda_1 > \ldots > \lambda_{p-k} = \lambda_{p-k+1} = \ldots = \lambda_p = \lambda,$$

el estimador máximo verosímil de λ *es la media de los correspondientes valores propios de* $\mathbf{S}$

$$\widehat{\lambda} = (l_{p-k+1} + \ldots + l_p)/k$$

Demost.: Los valores y vectores propios están biunívocamente relacionados con Σ y por lo tanto 1) es consecuencia de la propiedad de invariancia de la estimación máximo verosímil. La demostración de 2) se encuentra en Anderson (1959).□

Teorema 5.4.2 *Los vectores propios* $[\mathbf{g}_1, \ldots, \mathbf{g}_p]$ *y valores propios* $\mathbf{l} = [l_1, \ldots, l_p]$ *verifican asintóticamente:*

1. $\mathbf{l}$ *es* $N_p(\boldsymbol{\lambda}, 2\Lambda^2/n)$. *En particular:*

$$l_i \quad \text{es} \quad N(\lambda_i, 2\lambda_i^2/n), \qquad cov(l_i, l_j) = 0, \quad i \neq j,$$

es decir, l_i, l_j *son normales e independientes.*

2. $\mathbf{g}_i$ *es* $N_p(\boldsymbol{\gamma}_i, \mathbf{V}_i/n)$ *donde*

$$\mathbf{V}_i = \lambda_i \sum_{j \neq i} \frac{\lambda_i}{(\lambda_i - \lambda_j)^2} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i'$$

3. $\mathbf{l}$ *es independiente de* $\mathbf{G}$.

Demost.: Anderson (1959), Mardia, Kent y Bibby (1979).□

Como consecuencia de que l_i es $N(\lambda_i, 2\lambda_i^2/n)$, obtenemos el intervalo de confianza asintótico con coeficiente de confianza $1 - \alpha$

$$\frac{l_i}{(1 + az_{\alpha/2})^{1/2}} < \lambda_i < \frac{l_i}{(1 - az_{\alpha/2})^{1/2}}$$

siendo $a^2 = 2/(n-1)$ y $P(|Z| > z_{\alpha/2}) = \alpha/2$, donde Z es $N(0,1)$.

Se obtiene otro intervalo de confianza como consecuencia de que $\log l_i$ es $N(\log \lambda_i, 2/(n-1))$

$$l_i e^{-az_{\alpha/2}} < \lambda_i < l_i e^{+az_{\alpha/2}}.$$

5.4.2 Tests de hipótesis

Determinados tests de hipótesis relativos a las componentes principales son casos particulares de un test sobre la estructura de la matriz Σ.

A. Supongamos que queremos decidir si la matriz Σ es igual a una matriz determinada Σ_0. Sea $\mathbf{X}$ un matriz $n \times p$ con filas independientes $N_p(\mu, \Sigma)$. El test es:

$$H_0 : \Sigma = \Sigma_0 \quad (\mu \quad \text{desconocida})$$

Si L es la verosimilitud de la muestra, el máximo de $\log L$ bajo H_o es

$$\log L_0 = -\frac{n}{2} \log |2\pi\Sigma_0| - \frac{n}{2} tr(\Sigma_0^{-1}\mathbf{S}).$$

El máximo no restringido es

$$\log L = -\frac{n}{2} \log |2\pi\mathbf{S}| - \frac{n}{2} p.$$

El estadístico basado en la razón de verosimilitud λ_R es

$$\begin{aligned} -2\log\lambda_R &= 2(\log L - \log L_0) \\ &= n\text{tra}(\Sigma_0^{-1}\mathbf{S}) - n\log|\Sigma_0^{-1}\mathbf{S}| - np. \end{aligned} \tag{5.4}$$

Si $L_1, \ldots, L_p$ son los valores propios de $\Sigma_0^{-1}\mathbf{S}$ y a, g son las medias aritmética y geométrica

$$a = (L_1 + \ldots + L_p)/p, \quad g = (L_1 \times \ldots \times L_p)^{1/p}, \tag{5.5}$$

entonces, asintóticamente

$$-2\log\lambda_R = np(a - \log g - 1) \sim \chi^2_q, \tag{5.6}$$

siendo $q = p(p+1)/2 - \text{par}(\Sigma_0)$ el número de parámetros libres de Σ menos el número de parámetros libres de Σ_0.

B. Test de independencia completa.

Si la hipótesis nula afirma que las p variables son estocásticamente independientes, el test se formula como

$$H_0 : \Sigma = \Sigma_d = \text{diag}(\sigma_{11}, \cdots, \sigma_{pp}) \quad (\mu \quad \text{desconocida}).$$

Bajo H_0 la estimación de Σ_d es $\mathbf{S}_d = \text{diag}(s_{11}, \cdots, s_{pp})$ y $\mathbf{S}_d^{-1}\mathbf{S} = \mathbf{R}$ es la matriz de correlaciones. De (5.4) y de $\log|2\pi\mathbf{S}_d| - \log|2\pi\mathbf{S}| = \log|\mathbf{R}|$, $\text{tra}(\mathbf{R}) = p$, obtenemos

$$-2\log\lambda_R = -n\log|\mathbf{R}| \sim \chi^2_q$$

siendo $q = p(p+1)/2 - p = p(p-1)/2$. Si el estadístico $-n\log|\mathbf{R}|$ no es significativo, entonces podemos aceptar que las variables son incorrelacionadas y por lo tanto, como hay normalidad multivariante, independientes.

C. Test de igualdad de valores propios.

Este es un test importante en ACP. La hipótesis nula es

$$H_0 : \lambda_1 > \ldots > \lambda_{p-k} = \lambda_{p-k+1} = \ldots = \lambda_p = \lambda.$$

Indicamos los valores propios de $\mathbf{S}$ y de $\mathbf{S}_0$ (estimación de Σ si H_0 es cierta)

$$\mathbf{S} \sim (l_1, \ldots, l_k, l_{k+1}, \ldots, l_p), \quad \mathbf{S}_0 \sim (l_1, \ldots, l_k, a_0, \ldots, a_0),$$

donde $a_0 = (l_{k+1} + \ldots + l_p)/(p-k)$ (Teorema 5.4.1). Entonces

$$\mathbf{S}_0^{-1}\mathbf{S} \sim (1, \ldots, 1, l_{k+1}/a_0, \ldots, l_p/a_0),$$

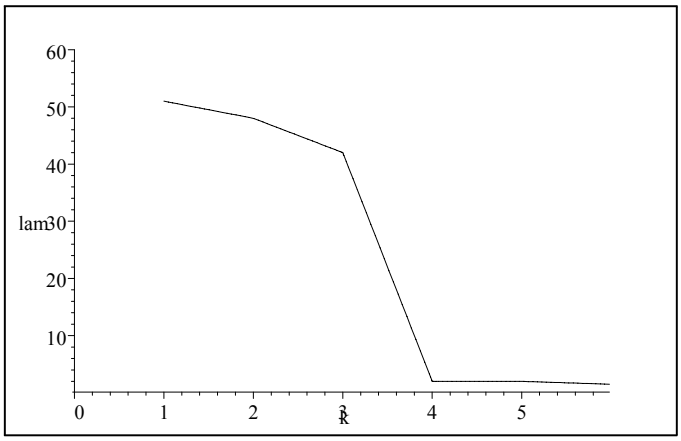

Figura 5.1: Ejemplo de representación de los valores propios, que indicaría 3 componentes principales.

las medias (5.5) son $a = 1$ y $g = (l_{k+1} \times \ldots \times l_p)^{1/p} a_0^{(k-p)/p}$ y aplicando (5.6)

$$-2 \log \lambda_R = n(p-k) \log(l_{k+1} + \ldots + l_p)/(p-k) - n(\sum_{i=k+1}^{p} \log l_i) \sim \chi_q^2, \quad (5.7)$$

donde $q = (p-k)(p-k+1)/2 - 1$.

5.5 Número de componentes principales

En esta sección presentamos algunos criterios para determinar el número $m < p$ de componentes principales.

5.5.1 Criterio del porcentaje

El número m de componentes principales se toma de modo que P_m sea próximo a un valor especificado por el usuario, por ejemplo el 80%. Por otra parte, si la representación de $P_1, P_2, \ldots, P_k, \ldots$ con respecto de k prácticamente se estabiliza a partir de un cierto m, entonces aumentar la dimensión apenas aporta más variabilidad explicada.

5.5.2 Criterio de Kaiser

Obtener las componentes principales a partir de la matriz de correlaciones $\mathbf{R}$ equivale a suponer que las variables observables tengan varianza 1. Por lo tanto una componente principal con varianza inferior a 1 explica menos variabilidad que una variable observable. El criterio, llamado de Kaiser, es entonces:

Retenemos las m primeras componentes tales que $\lambda_m \geq 1$,
donde $\lambda_1 \geq \ldots \geq \lambda_p$ son los valores propios de $\mathbf{R}$, que también son las varianzas de las componentes. Estudios de Montecarlo prueban que es más correcto el punto de corte $\lambda^* = 0.7$, que es más pequeño que 1.

Este criterio se puede extender a la matriz de covarianzas. Por ejemplo, m podría ser tal que $\lambda_m \geq v$, donde $v = \text{tra}(\mathbf{S})/p$ es la media de las varianzas. También es aconsejable considerar el punto de corte $0.7 \times v$.

5.5.3 Test de esfericidad

Supongamos que la matriz de datos proviene de una población normal multivariante $N_p(\mu, \Sigma)$. Si la hipótesis

$$H_0^{(m)} : \lambda_1 > \ldots > \lambda_m > \lambda_{m+1} = \ldots = \lambda_p$$

es cierta, no tiene sentido considerar más de m componentes principales. En efecto, no hay direcciones de máxima variabilidad a partir de m, es decir, la distribución de los datos es esférica. El test para decidir sobre $H_0^{(m)}$ está basado en el estadístico ji-cuadrado (5.7) y se aplica secuencialmente: Si aceptamos $H_0^{(0)}$ no hay direcciones principales, pero si rechazamos $H_0^{(0)}$, entonces repetimos el test con $H_0^{(1)}$. Si aceptamos $H_0^{(1)}$ entonces $m = 1$, pero si rechazamos $H_0^{(1)}$ repetimos el test con $H_0^{(2)}$, y así sucesivamente. Por ejemplo, si $p = 4$, tendríamos que $m = 2$ si rechazamos $H_0^{(0)}, H_0^{(1)}$ y aceptamos $H_0^{(2)} : \lambda_1 > \lambda_2 > \lambda_3 = \lambda_4$.

5.5.4 Criterio del bastón roto

Los valores propios suman $V_t = \text{tr}(\mathbf{S})$, que es la variabilidad total. Imaginemos un bastón de longitud V_t, que rompemos en p trozos al azar (asignando $p-1$ puntos uniformemente sobre el intervalo $(0, V_t)$) y que los trozos ordenados

son los valores propios $l_1 > l_2 > \ldots > l_p$. Si normalizamos a $V_t = 100$, entonces el valor esperado de l_j es

$$E(L_j) = 100 \times \frac{1}{p} \sum_{i=1}^{p-j} \frac{1}{j+i}.$$

Las m primeras componentes son significativas si el porcentaje de varianza explicada supera claramente el valor de $E(L_1) + \ldots + E(L_m)$. Por ejemplo, si $p = 4$, los valores son:

Porcentaje	$E(L_1)$	$E(L_2)$	$E(L_3)$	$E(L_4)$
Esperado	52.08	27.08	14.58	6.25
Acumulado	52.08	79.16	93.74	100

Si $V_2 = 93.92$ pero $V_3 = 97.15$, entonces tomaremos sólo dos componentes.

5.5.5 Un ejemplo

Exemple 5.5.1

Sobre una muestra de $n = 100$ estudiantes de Bioestadística, se midieron las variables

X_1 = peso (kg), X_2 =talla (cm.), X_3 =ancho hombros (cm.), X_4 = ancho caderas (cm.),

con los siguientes resultados:

1. medias: $\overline{x}_1 = 54.25, \overline{x}_2 = 161.73, \overline{x}_3 = 36.53, \overline{x}_4 = 30.1$.

2. matriz de covarianzas:

$$\mathbf{S} = \begin{pmatrix} 44.7 & 17.79 & 5.99 & 9.19 \\ 17.79 & 26.15 & 4.52 & 4.44 \\ 5.99 & 4.52 & 3.33 & 1.34 \\ 9.19 & 4.44 & 1.34 & 4.56 \end{pmatrix}$$

3. vectores y valores propios (columnas):

	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$
	.8328	.5095	.1882	.1063
	.5029	−.8552	.0202	.1232
	.1362	−.0588	.1114	−.9826
	.1867	.0738	−.9755	−.0892
Val. prop.	58.49	15.47	2.54	2.24
Porc. acum.	74.27	93.92	97.15	100

4. Número de componentes:

   a. Criterio de Kaiser: la media de las varianzas es $v = \text{tr}(\mathbf{S})/p = 19.68$. Los dos primeros valores propios son 58.49 y 15.47, que son mayores que $0.7 \times v$. Aceptamos $m = 2$.

   b. Test de esfericidad.

m	χ^2	g.l.
0	333.9	9
1	123.8	5
2	0.39	2

   Rechazamos $m = 0, m = 1$ y aceptamos $m = 2$.

   c. Test del bastón roto: Puesto que $P_2 = 93.92$ supera claramente el valor esperado 79.16 y que no ocurre lo mismo con P_3, aceptamos $m = 2$.

5. Componentes principales:

$$Y_1 = .8328X_1 + .5029X_2 + .1362X_3 + .1867X_4,$$
$$Y_2 = .5095X_1 - .8552X_2 - .0588X_3 + .0738X_4.$$

6. Interpretación: la primera componente es la variable con máxima varianza y tiene todos sus coeficientes positivos. La interpretamos como una componente de *tamaño*. La segunda componente tiene coeficientes positivos en la primera y cuarta variable y negativos en las otras dos. La interpretamos como una componente de *forma*. La primera componente ordena las estudiantes según su tamaño, de la más pequeña a la más grande, y la segunda según la forma, el tipo pícnico en contraste con el tipo atlético. Las dimensiones de tamaño y forma están incorrelacionadas.

5.6 Complementos

El Análisis de Componentes Principales (ACP) fué iniciado por K. Pearson en 1901 y desarrollado por H. Hotelling en 1933. Es un método referente a una población, pero W. Krzanowski y B. Flury han investigado las componentes principales comunes a varias poblaciones.

El ACP tiene muchas aplicaciones. Una aplicación clásica es el estudio de P. Jolicoeur y J. E. Mosimann sobre tamaño y forma de animales, en términos de la primera, segunda y siguientes componentes principales. La primera componente permite ordenar los animales de más pequeños a más grandes, y la segunda permite estudiar su variabilidad en cuanto a la forma. Nótese que tamaño y forma son conceptos "independientes".

El ACP puede servir para estudiar la capacidad. Supongamos que la caparazón de una tortuga tiene longitud L, ancho A, y alto H. La capacidad sería $C = L^\alpha A^\beta H^\gamma$, donde α, β, γ son parámetros. Aplicando logaritmos, obtenemos

$$\log C = \alpha \log L + \beta \log A + \gamma \log H = \log(L^\alpha A^\beta H^\gamma),$$

que podemos interpretar como la primera componente principal Y_1 de las variables $\log L, \log A, \log H$, y por tanto α, β, γ serían los coeficientes de Y_1.

Por medio del ACP es posible efectuar una regresión múltiple de Y sobre $X_1, \ldots, X_p$, considerando las primeras componentes principales $Y_1, Y_2, \ldots$ como variables explicativas, y realizar regresión de Y sobre $Y_1, Y_2, \ldots$, evitando así efectos de colinealidad, aunque las últimas componentes principales también pueden influir (Cuadras, 1993). La regresión ortogonal es una variante interesante. Supongamos que se quieren relacionar las variables $X_1, \ldots, X_p$ (todas con media 0), en el sentido de encontrar los coeficientes $\beta_1, \ldots, \beta_p$ tales que $\beta_1 X_1 + \ldots + \beta_p X_p \cong 0$. Se puede plantear el problema como $\mathrm{var}(\beta_1 X_1 + \ldots + \beta_p X_p) =$mínima, condicionado a $\beta_1^2 + \ldots + \beta_p^2 = 1$. Es fácil ver que la solución es la última componente principal Y_p.

Se pueden definir las componentes principales de un proceso estocástico y de una variable aleatoria. Cuadras y Fortiana (1995), Cuadras y Lahlou (2000) han estudiado las componentes principales de las variables uniforme, exponencial y logística.

Capítulo 6

ANÁLISIS FACTORIAL

6.1 Introducción

El Análisis Factorial (AF) es un método multivariante que pretende expresar p variables observables como una combinación lineal de m variables hipotéticas o latentes, denominadas *factores*. Tiene una formulación parecida al Análisis de Componentes Principales, pero el modelo que relaciona variables y factores es diferente en AF. Si la matriz de correlaciones existe, las componentes principales también existen, mientras que el modelo factorial podría ser aceptado o no mediante un test estadístico.

Ejemplos en los que la variabilidad de las variables observables se puede resumir mediante unas variables latentes, que el AF identifica como "factores", son:

1. La teoria clásica de la inteligencia suponía que los tests de inteligencia estaban relacionados por un factor general, llamado factor "g" de Spearman.

2. La estructura de la personalidad, también medida a partir de los tests, está dominada por dos dimensiones: el factor neuroticismo-estabilidad y el factor introversión-extroversión.

3. Las diferentes características políticas de ciertos países están influidas por dos dimensiones: izquierda-derecha y centralismo-nacionalismo.

El AF obtiene e interpreta los factores comunes a partir de la matriz de

correlaciones entre las variables:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ & \cdots & \cdots & \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}.$$

6.2 El modelo unifactorial

Consideremos $X_1, \ldots, X_p$ variables observables sobre una misma población. El modelo más simple de AF sólo contempla un factor común F, que recoge la covariabilidad de todas las variables, y p factores únicos $U_1, \ldots, U_p$, uno para cada variable. El modelo factorial es

$$X_i = a_i F + d_i U_i, \qquad i = 1, \ldots, p. \tag{6.1}$$

De acuerdo con este modelo, cada variable X_i depende del factor común F y de un factor único U_i. El modelo supone que:

a) las variables y los factores están estandarizados (media 0 y varianza 1).

b) Los $p+1$ factores están incorrelacionados.

De este modo F contiene la parte de la variabilidad común a todas las variables, y cada X_i está además influida por un factor único U_i, que aporta la parte de la variabilidad que no podemos explicar a partir del factor común. El coeficiente a_i es la *saturación* de la variable X_i en el factor F.

De (6.1) deducimos inmediatamente que

$$\begin{aligned} a_i^2 + d_i^2 &= 1, \\ \operatorname{cor}(X_i, F) &= a_i, \\ \operatorname{cor}(X_i, X_j) &= a_i a_j, \quad i \neq j. \end{aligned}$$

Por lo tanto la saturación a_i es el coeficiente de correlación entre X_i y el factor común. Por otra parte a_i^2, cantidad que recibe el nombre de *comunalidad,* indicada por h_i^2, es la proporción de variabilidad que se explica por F y la correlación entre X_i, X_j sólo depende de las saturaciones a_i, a_j.

Una caracterització del modelo unifactorial es

$$\frac{r_{ij}}{r_{i'j}} = \frac{r_{ij'}}{r_{i'j'}} = \frac{a_i}{a_{i'}}, \tag{6.2}$$

es decir, los cocientes entre elementos de la misma columna no diagonal de dos filas de la matriz de correlaciones $\mathbf{R}$ es constante. Esto es equivalente a decir que el determinante de todo menor de orden dos de $\mathbf{R}$, que no contenga elementos de la diagonal, es cero:

$$\begin{vmatrix} r_{ij} & r_{ij'} \\ r_{i'j} & r_{i'j'} \end{vmatrix} = r_{ij}r_{i'j'} - r_{ij'}r_{i'j'} = a_i a_j a_{i'} a_{j'} - a_i a_{j'} a_{i'} a_{j'} = 0. \tag{6.3}$$

Estas son las llamadas relaciones tetrádicas, que necesariamente se deben cumplir para que sea válido el modelo unifactorial.

La matriz de correlacions reducida $\mathbf{R}^*$ se obtiene substituyendo la diagonal de unos por las comunalidades (véase (6.7)). Es inmediato probar que $\mathbf{R}^*$ tiene rango 1, que todos los menores de orden dos se anulan y que las comunalidades se obtienen a partir de las correlaciones. Por ejemplo, la primera comunalidad es

$$h_1^2 = \frac{r_{12}r_{13}}{r_{23}} = \frac{r_{12}r_{14}}{r_{24}} = \cdots = \frac{r_{1p-1}r_{1p}}{r_{pp-1}}. \tag{6.4}$$

En las aplicaciones reales, tanto estas relaciones, com las tetrádicas, sólo se verifican aproximadamente. Así, la estimación de la primera comunalidad podría consistir en tomar la media de los cocientes (6.4).

Por ejemplo, la siguiente matriz de correlaciones

	C	F	I	M	D	Mu
C	1.00	0.83	0.78	0.70	0.66	0.63
F	0.83	1.00	0.67	0.67	0.65	0.57
I	0.78	0.67	1.00	0.64	0.54	0.51
M	0.70	0.67	0.64	1.00	0.45	0.51
D	0.66	0.65	0.54	0.45	1.00	0.40
Mu	0.63	0.57	0.51	0.51	0.40	1.00

relaciona las calificaciones en C (clásicas), F (francés), I (inglés), M (matemáticas), D (discriminación de tonos) y Mu (música) obtenidas por los alumnos de una escuela. Esta matriz verifica, aproximadamente, las relaciones (6.2). Si consideramos la primera y la tercera fila, tenemos que:

$$\frac{0.83}{0.67} \cong \frac{0.70}{0.64} \cong \frac{0.66}{0.54} \cong \frac{0.63}{0.51} \cong 1.2\ .$$

De acuerdo con el modelo unifactorial, estas calificaciones dependen esencialmente de un factor común.

6.3 El modelo multifactorial

6.3.1 El modelo

El modelo del análisis factorial de m factores comunes considera que las p variables observables $X_1, \ldots, X_p$ dependen de m variables latentes $F_1, \ldots, F_m$, lamadas factores comunes, y p factores únicos $U_1, \ldots, U_p$, de acuerdo con el modelo lineal:

$$\begin{array}{rcccccc} X_1 = & a_{11}F_1 & +\cdots+ & a_{1m}F_m & +d_1U_1 & & \\ X_2 = & a_{21}F_1 & +\cdots+ & a_{2m}F_m & & +d_2U_2 & \\ & & \cdots & & \cdots & & \\ X_p = & a_{p1}F_1 & +\cdots+ & a_{1p}F_m & & & +d_pU_p. \end{array} \tag{6.5}$$

Las hipótesis del modelo son:

1. Los factores comunes y los factores únicos están incorrelacionados dos a dos
$$\begin{aligned} \text{cor}(F_i, F_j) &= 0, \quad i \neq j = 1, \ldots, m, \\ \text{cor}(U_i, U_j) &= 0, \quad i \neq j = 1, \ldots, p. \end{aligned}$$

2. Los factores comunes están incorrelacionados con los factores únicos
$$\text{cor}(F_i, U_j) = 0, \quad i = 1, \ldots, m, \ j = 1, \ldots, p.$$

3. Tanto los factores comunes como los factores únicos són variables reducidas.

En el modelo factorial (6.5) se admite que las variables, en conjunto, dependen de los factores comunes, salvo una parte de su variabilidad, sólo explicada por el correspondiente factor específico. Los factores comunes representan dimensiones independentes en el sentido lineal, y dado que tanto los factores comunes como los únicos son variables convencionales, podemos suponer que tienen media 0 y varianza 1.

6.3.2 La matriz factorial

Los coeficientes a_{ij} son las *saturaciones* entre cada variable X_i y el factor F_j. La matriz $p \times m$ que contiene estos coeficientes es la matriz factorial

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ & \cdots & \\ a_{p1} & \cdots & a_{pm} \end{pmatrix}.$$

Si indicamos por $\mathbf{X} = (X_1, \ldots, X_p)'$ el vector columna de las variables, y análogamente $\mathbf{F} = (F_1, \ldots, F_m)', \mathbf{U} = (U_1, \ldots, U_p)'$, el modelo factorial en expresión matricial es

$$\mathbf{X} = \mathbf{AF} + \mathbf{DU}, \tag{6.6}$$

donde $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$ es la matriz diagonal con las saturaciones entre variables y factores únicos. El AF tiene como principal objetivo encontrar e interpretar la matriz factorial $\mathbf{A}$.

6.3.3 Las comunalidades

De las condiciones del modelo del AF se verifica

$$\text{var}(X_i) = a_{i1}^2 + \cdots + a_{im}^2 + d_i^2,$$

y por lo tanto a_{ij}^2 es la parte de la variabilidad de la variable X_i que es debida al factor común F_j, mientras que d_i^2 es la parte de la variabilidad explicada exclusivamente por el factor único U_i.

La cantidad

$$h_i^2 = a_{i1}^2 + \cdots + a_{im}^2 \tag{6.7}$$

se llama *comunalidad* de la variable X_i. La cantidad d_i^2 es la *unicidad.* Luego, para cada variable tenemos que:

$$\text{variabilidad} = \text{comunalidad} + \text{unicidad}.$$

La comunalidad es la parte de la variabilidad de las variables sólo explicada por los factores comunes.

Si supoemos que las variables observables son también reducidas, entonces tenemos que

$$1 = h_i^2 + d_i^2. \tag{6.8}$$

La matriz de correlaciones reducida se obtiene a partir de $\mathbf{R}$ substituyendo los unos de la diagonal por las comunalidades

$$\mathbf{R}^*= \begin{pmatrix} h_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & h_2^2 & \cdots & r_{2p} \\ & \cdots & \cdots & \\ r_{p1} & r_{p2} & \cdots & h_p^2 \end{pmatrix}.$$

Evidentmente se verifica

$$\mathbf{R} = \mathbf{R}^* + \mathbf{D}^2. \tag{6.9}$$

6.3.4 Número máximo de factores comunes

El número m de factores comunes está limitado por un valor máximo m_a, que podemos determinar teniendo en cuenta que hay $p(p-1)/2$ correlaciones diferentes y $p \cdot m$ saturaciones. Pero si $\mathbf{A}$ es matriz factorial también lo es $\mathbf{AT}$, donde $\mathbf{T}$ es matriz ortogonal, por tanto introduciremos $m(m-1)/2$ restricciones y el número de parámetros libres de $\mathbf{A}$ será $p \cdot m - m(m-1)/2$. El número de correlaciones menos el número de parámetros libres es

$$d = p(p-1)/2 - (p \cdot m - m(m-1)/2) = \frac{1}{2}[(p-m)^2 - p - m]. \tag{6.10}$$

Si igualamos d a 0 obtenemos una ecuación de segundo grado que un vez resuelta nos prueba que

$$m \leq m_a = \frac{1}{2}(2p + 1 - \sqrt{8p+1}).$$

Un modelo factorial es sobredeterminado si $m > m_a$, pues hay más saturaciones libres que correlaciones. Si $m = m_a$ el modelo es determinado y podemos encontrar $\mathbf{A}$ algebraicamente a partir de $\mathbf{R}$.

Desde un punto de vista estadístico, el caso más interesante es $m < m_a$, ya que entonces podemos plantear la estimación estadística de $\mathbf{A}$, donde $d > 0$ juega el papel de número de grados de libertad del modelo. El número máximo m^* de factores comunes en función de p es:

p	2	3	4	5	6	7	8	9	10	20	30	40
m^*	0	1	1	2	3	3	4	5	6	14	22	31

Asignamos a m^* el valor entero por defecto cuando m_a tiene parte fraccionaria.

6.3.5 El caso de Heywood

Una limitación del model factorial es que alguna comunalidad puede alcanzar (algebraicamente) un valor superior a 1, contradiciendo (6.8). Cuando esto ocurre, la solució se ha de interpretar con precaución. En algunos métodos, como el de la máxima verosimilitud, se resuelve este inconveniente (primeramente observado por H.B. Heywood) imponiendo la condición $h_i^2 \leq 1$ en la estimación de las comunalidades.

6.3.6 Un ejemplo

Las asignaturas clásicas de la enseñanza media, se dividen, en líneas generales, en asignaturas de Ciencias o de Letras, las primeras con contenido más racional y empírico, las segundas con contenido más humanístico y artístico. Consideremos las siguientes 5 asignaturas:

Ciencias Naturales (CNa), Matemáticas (Mat), Francés (Fra), Latín (Lat), Literatura (Lit). Supongamos que están influidas por dos factores comunes o variables latentes: Ciencias (C) y Letras (L). En otras palabras, suponemos que C y L son dos variables no observables, que de manera latente influyen sobre las cinco asignaturas. Las calificaciones de $n = 20$ alumnos en las asignaturas y en los factores se encuentran en la Tabla 6.1.

Vamos a suponer que la matriz factorial es

	C	L
CNa	.8	.2
Mat	.9	.1
Fra	.1	.9
Lla	.3	.8
Lit	.2	.8

Las dos primeras asignaturas están más influidas por el factor C, y las tres últimas por el factor L. Por ejemplo, Matemáticas tiene una correlación de 0.9 con Ciencias y sólo 0.1 con Letras.

La calificación del primer alumno en CNa es 7, debida a 7 puntos en Ciencias y 5 puntos en Letras. Según el modelo factorial:

$$7 = 0.8 \times 7 + 0.2 \times 5 + 0.4$$

	Asignaturas					Factors	
Alumno	CNa	Mat	Fra	Lat	Lit	Ciències	Lletres
1	7	7	5	5	6	7	5
2	5	5	6	6	5	5	6
3	5	6	5	7	5	6	5
4	6	8	5	6	6	7	5
5	7	6	6	7	6	6	6
6	4	4	6	7	6	4	6
7	5	5	5	5	6	5	6
8	5	6	5	5	5	6	5
9	6	5	7	6	6	5	6
10	6	5	6	6	6	5	6
11	6	7	5	6	5	7	5
12	5	5	4	5	4	6	4
13	6	6	6	6	5	6	6
14	8	7	8	8	8	7	8
15	6	7	5	6	6	6	5
16	4	3	4	4	4	3	4
17	6	4	7	8	7	5	7
18	6	6	7	7	7	6	7
19	6	5	4	4	4	5	4
20	7	7	6	7	6	7	6

Tabla 6.1: Calificaciones en 5 asignaturas y puntuaciones en 2 factores comunes de 20 alumnos.

De los 7 puntos, 5.6 se explican por el factor común C, 1 punto por el factor común L y 0.4 punts por el factor único. Este factor único representa la variabilidad propia de las CNa, independente de los conceptos C y L.

Las comunalidades son:

$$h_1^2 = 0.68,\ h_2^2 = 0.82,\ h_3^2 = 0.82,\ h_4^2 = 0.73,\ h_5^2 = 0.68.$$

Los porcentajes de la variabilidad explicada por los factores comunes y las comunalidades son:

	Factor C	Factor L	Comunalidades
C. Naturales	64	4	68
Matemáticas	81	1	82
Francés	1	81	82
Latín	9	64	73
Literatura	4	64	68

6.4 Teoremas fundamentales

El primer teorema, conocido como teorema de Thurstone, permite relacionar la matriz factorial con la matriz de correlaciones, o más exactamente, con la matriz de correlaciones reducida. El segundo teorema permite determinar, teóricamente, el número de factores comunes y los valores de las comunalidades.

Teorema 6.4.1 *Bajo las hipòtesis del modelo factorial lineal se verifica*

$$\begin{aligned} r_{ij} &= \textstyle\sum_{k=1}^{m} a_{ik}a_{jk}, \qquad i \neq j = 1, \ldots, p, \\ 1 &= \textstyle\sum_{k=1}^{m} a_{ik}^2 + d_i^2, \qquad i = 1, \ldots, p. \end{aligned}$$

En notación matricial

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{D}^2. \tag{6.11}$$

Demost.: Al ser las variables reducidas, $\mathbf{R} = E(\mathbf{X}\mathbf{X}')$ y de (6.6)

$$\begin{aligned} \mathbf{R} &= E((\mathbf{A}\mathbf{F} + \mathbf{D}\mathbf{U})(\mathbf{A}\mathbf{F} + \mathbf{D}\mathbf{U})') \\ &= \mathbf{A}E(\mathbf{F}\mathbf{F}')\mathbf{A}' + \mathbf{D}E(\mathbf{U}\mathbf{U}')\mathbf{D}' + 2\mathbf{A}E(\mathbf{F}\mathbf{U}')\mathbf{D}. \end{aligned}$$

Por las condiciones de incorrelación entre factores tenemos que $E(\mathbf{F}\mathbf{F}') = \mathbf{I}_m$, $E(\mathbf{U}\mathbf{U}') = \mathbf{I}_p$, $E(\mathbf{F}\mathbf{U}') = \mathbf{0}$, lo que prueba (6.11).□

De (6.9) vemos inmediatamente que

$$\mathbf{R}^* = \mathbf{A}\mathbf{A}'. \tag{6.12}$$

Una solución factorial viene dada por cualquier matriz $\mathbf{A}$ que cumpla la relación (6.12). Así pues, si $m > 1$, existen infinitas soluciones, pues si $\mathbf{A}$ es solución, también lo es $\mathbf{AT}$, siendo $\mathbf{T}$ una matriz $m \times m$ ortogonal. Por otro lado, (6.11) o (6.12) tampoco resuelven completamente el problema, ya que desconocemos las comunalidades. La obtención de las comunalidades está muy ligada al número de factores comunes.

Teorema 6.4.2 *Se verifica:*

1. *El modelo factorial existe si* $\mathbf{R}$ *es la suma de una matriz semidefinida positiva y una matriz diagonal con elementos no negativos.*

2. *El número* m *de factores comunes es el rango de la matriz* $\mathbf{R}^*$*. Por lo tanto* m *es el orden del más grande menor de* $\mathbf{R}$ *que no contiene elementos de la diagonal.*

3. *Les comunalidades son aquellos valores* $0 \leq h_i^2 \leq 1$ *tales que* $\mathbf{R}^*$ *es matriz semi-definida positiva (tiene* m *valores propios positivos).*

Prueba: Es una consecuencia de la relación (6.12) entre $\mathbf{R}^*$ y $\mathbf{A}$. El mayor menor de $\mathbf{R}$ quiere decir la submatriz cuadrada con determinante no negativo, que no contenga elementos de la diagonal.□

Hemos visto que a partir de $\mathbf{R}$ podemos encontrar m, pero la solución no es única. El *principio de parsimonia* en AF dice que entre varias soluciones admisibles, escogeremos la que sea más simple. El modelo factorial será pues aquel que implique un número mínimo m de factores comunes. Fijado m, las comunalidads se pueden encontrar, algebraicamente, a partir de la matriz de correlaciones $\mathbf{R}$. En la práctica, las comunalidades se hallan aplicando métodos estadísticos.

Finalmente, podemos probar de manera análoga, que si el análisis factorial lo planteamos a partir de la matriz de covarianzas $\mathbf{\Sigma}$, sin suponer las variables reducidas, aunque sí los factores, entonces obtenemos la estructura

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{A}' + \mathbf{D}^2. \tag{6.13}$$

6.5 Método del factor principal

Es un método de obtención de la matriz factorial con la propiedad de que los factores expliquen máxima varianza y sean incorrelacionados.

La variabilidad total de las variables, que suponemos reducidas, es p. La variabilidad de la variable X_i explicada por el factor F_j es a_{ij}^2. La suma de variabilidades explicadas por F_j es

$$V_j = a_{1j}^2 + \ldots + a_{pj}^2.$$

El primer factor principal F_1 es tal que V_1 es máximo. Consideremos pues el problema de maximizar V_1 con la restricción $\mathbf{R}^* = \mathbf{A}\mathbf{A}'$. Utilizando el método de los multiplicadores de Lagrange debemos considerar la función

$$V_1 + \sum_{j,j'=1}^{p} q_{jj'}(r_{jj'} - \sum_{k=1}^{m} a_{jk}a_{j'k}),$$

donde $q_{jj'} = q_{j'j}$ són los multiplicadores. Igualando las derivadas a cero se obtiene que las saturaciones $\mathbf{a}_1 = (a_{11}, \ldots, a_{p1})'$ del primer factor principal verifican

$$\mathbf{R}^*\mathbf{a}_1 = \lambda_1\mathbf{a}_1,$$

es decir, $\mathbf{a}_1$ es el primer vector propio de $\mathbf{R}^*$ y λ_1 es el primer valor propio. El valor máximo de V_1 es precisamente λ_1.

Si ahora restamos del modelo factorial el primer factor

$$X_i' = X_i - a_{i1}F_1 = a_{i2}F_2 + \ldots + a_{im}F_m + d_iU_i,$$

el modelo resultante contiene $m-1$ factores. Aplicando de nuevo el criterio del factor principal al modelo vemos que las saturaciones $\mathbf{a}_2 = (a_{12}, \ldots, a_{p2})'$ tales que la variabilidad explicada por el segundo factor

$$V_2 = a_{12}^2 + \ldots + a_{p2}^2,$$

sea máxima, corresponende al segundo vector propio de $\mathbf{R}^*$ con valor propio λ_2, que es precisament el valor máximo de V_2.

En general, si $\mathbf{R}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$ es la descomposición espectral de $\mathbf{R}^*$, la solución del factor principal es

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}^{1/2}.$$

Fijado un valor compatible de m, un algoritmo iterativo de obtención de la matriz factorial y de las comunalidads es:

Paso 1	$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$	(p valores y vectores propios)
Paso 2	$\mathbf{R}_1 = \mathbf{U}_m^{(1)}\mathbf{\Lambda}_m^{(1)}\mathbf{U}_m^{(1)\prime}$	(m primeros valores y vectores propios)
Paso i	$\mathbf{R}_i = \mathbf{U}_m^{(i)}\mathbf{\Lambda}_m^{(i)}\mathbf{U}_m^{(i)\prime}$,	
	$\mathbf{A}_i = \mathbf{U}_m^{(i)}(\mathbf{\Lambda}_m^{(i)})^{1/2}$	
Paso i+1	$\mathbf{R}_{i+1} = \text{diag}(\mathbf{A}_i\mathbf{A}_i') + \mathbf{R} - \mathbf{I}$	(volver al paso i)

La matriz $\mathbf{A}_i$ converge a la matriz factorial $\mathbf{A}$. Como criterio de convergencia podemos considerar la estabilidad de las comunalidades. Pararemos si pasando de i a $i+1$ los valores de las comunalidads, es decir, los valores en $\text{diag}(\mathbf{A}_i\mathbf{A}_i')$, prácticamente no varían. Esta refactorización podria fallar si se presenta el caso de Heywood o $\mathbf{R}$ no satisface el model factorial (6.11).

Ejemplo: Volviendo al ejemplo de las asignaturas, la solución por el método del factor principal encuentra dos factores que explican el 74.6% de la varianza:

	F_1	F_2
C. Naturales	.621	-.543
Matemáticas	.596	-.682
Francés	.796	.432
Latín	.828	.210
Literatura	.771	.292
Valor propio	2.654	1.076
Porcentaje	53.08	21.52

6.6 Método de la máxima verosimilitud

6.6.1 Estimación de la matriz factorial

Podemos plantear la obtención de la matriz factorial como un problema de estimación de la matriz de covarianzas Σ, con la restricción que Σ se descompone en la forma

$$\Sigma = \mathbf{A}\mathbf{A}' + \mathbf{V},$$

donde $\mathbf{V} = \mathbf{D}^2$ es una matriz diagonal (véase (6.13)). Si suponemos que las n observacions de las p variables provienen de una distribución normal con

$\boldsymbol{\mu} = \mathbf{0}$, el logaritmo de la función de verosimilitud es

$$\log L(\mathbf{X},\mu,\Sigma) = -\frac{n}{2}(\log|2\pi\Sigma| - tr(\Sigma^{-1}\mathbf{S})).$$

Cambiando de signo y modificando algunas constantes, se trata de estimar $\mathbf{A}$ y $\mathbf{V}$ de manera que

$$F_p(\mathbf{A},\mathbf{V}) = \log|\Sigma| + tr(\Sigma^{-1}\mathbf{S}) - \log|\mathbf{S}| - p \tag{6.14}$$

sea mínimo, siendo $\mathbf{S}$ la matriz de covarianzas muestrales. Las derivadas respecto de $\mathbf{A}$ y $\mathbf{V}$ son

$$\begin{aligned}\frac{\partial F_p}{\partial A} &= 2\Sigma^{-1}(\Sigma - S)\Sigma^{-1}\mathbf{A},\\ \frac{\partial F_p}{\partial V} &= \mathrm{diag}(\Sigma^{-1}(\Sigma - S)\Sigma^{-1}).\end{aligned}$$

Por tanto, las ecuaciones a resolver para obtener estimaciones de $\mathbf{A}$ y $\mathbf{V}$ son

$$\begin{array}{ll}\Sigma^{-1}(\Sigma - S)\Sigma^{-1}\mathbf{A} = \mathbf{0}, & \mathrm{diag}(\Sigma^{-1}(\Sigma - S)\Sigma^{-1}) = \mathbf{0},\\ \Sigma = \mathbf{A}\mathbf{A}' + \mathbf{V}, & \mathbf{A}'\mathbf{V}^{-1}\mathbf{A} \text{ es diagonal.}\end{array} \tag{6.15}$$

La última condición es sólo una restricción para concretar una solución, puesto que si $\mathbf{A}$ es solución, també lo es $\mathbf{AT}$, siendo $\mathbf{T}$ matriz ortogonal. Debe tenerse en cuenta que se trata de encontrar el espacio de los factores comunes. La solución final será, en la práctica, una rotación de la solución que verifique ciertos criterios de simplicidad. Las ecuaciones (6.15) no proporcionan una solución explícita, pero es posible encontrar una solución utilizando un método numérico iterativo.

6.6.2 Hipótesis sobre el número de factores

Una ventaja del método de la máxima verosimilitud es que permite formular un test de hipótesis sobre la estructura factorial de Σ y el número m de factores comunes.

Planteemos el test

$$H_0 : \Sigma = \mathbf{A}\mathbf{A}' + \mathbf{V} \quad \text{vs} \quad H_1 : \Sigma \text{ es definida positiva,}$$

donde $\mathbf{A}$ es de rango m.

Si $\widehat{\Sigma} = \widehat{\mathbf{A}}\widehat{\mathbf{A}}' + \widehat{\mathbf{V}}$, siendo $\widehat{\mathbf{A}}$ y $\widehat{\mathbf{V}}$ las estimaciones, los máximos del logaritmo de la razón de verosimilitud son (Sección 5.4.2)

$$\begin{array}{ll} H_0: & -\frac{n}{2}(\log|\widehat{\Sigma}| + \mathrm{tr}(\widehat{\Sigma}^{-1}\mathbf{S})), \\ H_1: & -\frac{n}{2}(\log|\mathbf{S}| + p). \end{array}$$

Aplicando el Teorema 3.5.1 tenemos que el estadístico

$$C_k = n(\log|\widehat{\Sigma}| - \log|\mathbf{S}| + \mathrm{tr}(\widehat{\Sigma}^{-1}\mathbf{S}) - p) = nF_p(\widehat{\mathbf{A}}, \widehat{\mathbf{V}})$$

sigue asinptóticamente la distribución ji-cuadrado con

$$k = p(p-1)/2 - (p \cdot m + p - m(m-1)/2) = \frac{1}{2}((p-m)^2 - p - m)$$

grados de libertad. Podemos observar que C_k es n veces el valor mínimo de la función (6.14) y que k coincide con (6.10).

6.7 Rotaciones de factores

La obtención de la matriz factorial, por aplicación de los dos métodos que hemos expuesto, no es más que el primer paso del AF. Normalmente la matriz obtenida no define unos factores interpretables. En el ejemplo de las asignaturas, la solución por el método del factor principal es en principio válida, pero define dos factores comunes F_1, F_2 que no son fácilmente identificables. Se hace necesario "rotar" estos dos factores hacia unos factores más fáciles de interpretar.

Se han propuesto diferentes versiones sobre como transformar la matriz factorial a fin de obtener una estructura simple de los factores. Esencialmente se trata de conseguir que unas saturaciones sean altas a costa de otras, que serán bajas, para así destacar la influencia de los factores comunes sobre las variables observables.

6.7.1 Rotaciones ortogonales

Dada una matriz factorial $\mathbf{A}$, queremos encontrar una matriz ortogonal $\mathbf{T}$ tal que la nueva matriz factorial $\mathbf{B} = \mathbf{AT}$ defina unos factores que tengan una estructura más simple. Un criterio analítico considera la función

$$G = \sum_{k=1}^{m} \sum_{k \neq j=1}^{m} [\sum_{i=1}^{p} a_{ij}^2 a_{ik}^2 - \frac{\gamma}{p} \sum_{i=1}^{p} a_{ij}^2 \sum_{i=1}^{p} a_{ik}^2], \tag{6.16}$$

donde γ es un parámetro tal que $0 \leq \gamma \leq 1$. Hay dos criterios especialmente interesantes.

Quartimax: Si $\gamma = 0$ minimizar G equivale a maximizar la varianza de los cuadrados de los $p \cdot m$ coeficientes de saturación. Si cada saturación a_{ij}^2 se divide por la comunalidad, es decir, se considera a_{ij}^2/h_i^2, la rotación se llama quartimax normalizada.

Varimax: Si $\gamma = 1$ minimizar G equivale a maximizar la suma de las varianzas de los cuadrados de los coeficientes de saturación de cada columna de $\mathbf{A}$. Análogamente si consideramos a_{ij}^2/h_i^2, la rotación se llama varimax normalizada.

6.7.2 Factores oblicuos

Los factores comunes pueden estar también correlacionados, y entonces se habla del model factorial oblícuo. Este modelo postula que las variables observables dependen de unos factores correlacionados $F_1', \ldots, F_m'$ y de p factores únicos. Así para cada variable X_i

$$X_i = p_{i1}F_1' + \ldots + p_{im}F_m' + d_iU_i, \qquad i = 1, \ldots, p. \tag{6.17}$$

La solución factorial oblicua consistirá en hallar las siguientes matrices:

1. Matriz del modelo factorial oblícuo

   $$\mathbf{P} = (p_{ij})$$

   siendo p_{ij} la saturación de la variable X_i en el factor F_j'.

2. Matriz de correlaciones entre factores oblícuos

   $$\mathbf{\Phi} = (\varphi_{ij}) \quad \text{siendo } \varphi_{ij} = \text{cor}(F_i', F_j').$$

3. Estructura factorial oblicua (estructura de referencia)

   $$\mathbf{Q} = (q_{ij}) \quad \text{siendo } q_{ij} = \text{cor}(X_i, F_j').$$

Si indicamos $\mathbf{F}^0 = (F_1', \ldots, F_m')'$ y escribimos el modelo (6.17) en forma matricial

$$\mathbf{X} = \mathbf{PF}^0 + \mathbf{DU},$$

fácilmente probamos la relación entre las tres matrices $\mathbf{P}, \mathbf{\Phi}$ y $\mathbf{Q}$

$$\mathbf{Q} = \mathbf{P\Phi},$$

y la versión del teorema de Thurstone para factores correlacionados

$$\mathbf{R} = \mathbf{P\Phi P}' + \mathbf{D}^2.$$

Si los factores son ortogonales, el modelo factorial coincide con la estructura factorial y tenemos que

$$\mathbf{P} = \mathbf{Q}, \qquad \mathbf{\Phi} = \mathbf{I}_m.$$

6.7.3 Rotación oblicua

Ya se ha dicho que hallar una matriz factorial $\mathbf{A}$ constituye el primer paso de la factorización. Queremos encontrar una matriz $\mathbf{L}$ tal que la nueva matriz factorial $\mathbf{P} = \mathbf{AL}$ defina unos factores oblicuos que tengan una estructura más simple. Un criterio analítico sobre la matriz de estructura factorial $\mathbf{Q}$ considera la función

$$H = \sum_{k=1}^{m} \sum_{k \neq j=1} [\sum_{i=1}^{p} q_{ij}^2 q_{ik}^2 - \frac{\gamma}{p} \sum_{i=1}^{p} q_{ij}^2 \sum_{i=1}^{p} q_{ik}^2]$$

donde γ es un parámetro tal que $0 \leq \gamma \leq 1$. Hay tres criterios especialmente interesantes, que tienen una interpretación parecida al caso ortogonal y que también se pueden formular, más adecuadamente, dividiendo por las comunalidades.

Quartimin: Si $\gamma = 0$ hay máxima oblicuidad entre los factores comunes.

Bi-quartimin: Si $\gamma = 1/2$ el criterio es intermedio entre quartimin y covarimin.

Covarimin: Si $\gamma = 1$ hay mínima oblicuidad entre los factores comunes.

Conviene tener en cuenta que las rotaciones ortogonales y oblícuas intentan simplificar la estructura factorial $\mathbf{A}$ y la estructura de referencia $\mathbf{Q}$, respectivamente.

Un criterio directo de rotación oblicua es el *promax*. Sea $\mathbf{A}$ la matriz factorial obtenida por el método varimax. Queremos destacar unas saturaciones sobre otras, por tanto definimos $\mathbf{P}^* = (p_{ij}^*)$ tal que

$$p_{ij}^* = |a_{ij}^{k+1}|/a_{ij}, \quad k > 1,$$

siendo k un número entero.

Cada elemento de $\mathbf{A}$ queda elevado a una potencia k conservando el signo. Seguidamente ajustamos $\mathbf{P}^*$ a $\mathbf{AL}$ en el sentido de los mínimos cuadrados

$$\mathbf{L} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{P}^*.$$

Es necesario normalizar la matriz $\mathbf{L}$ de manera que los vectores columna de $\mathbf{T} = (\mathbf{L}')^{-1}$ tengan módulo unidad. Obtenemos entonces

$$\mathbf{P} = \mathbf{AL}, \quad \mathbf{\Phi} = \mathbf{T}'\mathbf{T}, \quad \mathbf{Q} = \mathbf{AT}.$$

El grado de oblicuidad de los factores comunes aumenta con k. Se suele tomar $k = 4$.

Ejemplo: Continuando con el ejemplo de las 5 asignaturas, la estimación máximo verosímil y la matriz factorial rotada son:

	Máxim veros.		Varimax		Comun.
	F_1	F_2	C	L	
CNa	.659	.432	.636	.464	.62
Mat	.999	.005	.999	.046	.99
Fra	.104	.974	.055	.978	.96
Lat	.234	.809	.193	.820	.71
Lit	.327	.831	.280	.847	.79

El test de hipótesis de que hay $m = 2$ factores comunes da $\chi_1^2 = 1.22$, no significativo. Podemos aceptar $m = 2$. La rotación varimax pone de manifiesto la existencia de dos factores C, L, que podemos interpretar como dimensiones latentes de Ciencias y Letras.

La rotación oblicua promax con $k = 4$ da las matrices $\mathbf{P}, \mathbf{Q}, \mathbf{\Phi}$:

	Modelo factorial		Estruct. factorial		Correlaciones factores
	C	L	C	L	
CNa	.570	.375	.706	.581	
Mat	1.04	-.135	.992	.242	$\begin{pmatrix} 1 & .362 \\ .362 & 1 \end{pmatrix}$
Fra	-.150	1.024	.221	.970	
Lla	.028	.831	.330	.842	
Lit	.114	.844	.420	.885	

La Figura 6.1 representa los factores ortogonales iniciales F_1 y F_2, dibujados como vectores unitarios, y los factores oblícuos C y L. Las variables tienen una longitud proporcional a la raíz cuadrada de sus comunalidades.

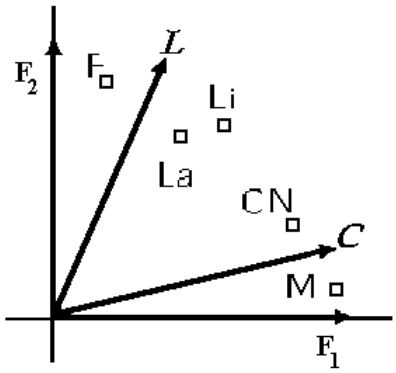

Figura 6.1: Proyección de las variables sobre los factors comunes ortogonals, y factores rotados (rotación promax), interpretados como factores de Ciencias y Letras.

6.7.4 Factores de segundo orden

Un vez hemos obtenido los factores oblícuos con matriz de correlaciones $\mathbf{\Phi}$, podemos suponer que estos m factores primarios dependen de m' factores secundarios de acuerdo con una matriz factorial $\mathbf{B}$ que verifica

$$\mathbf{\Phi} = \mathbf{B}\mathbf{B}' + \mathbf{E}^2,$$

siendo $\mathbf{E}$ la matriz $m \times m$ diagonal.

Si los factores secundarios son también oblicuos, el proceso de factorización puede continuar hasta llegar a un único factor común de orden superior.

Un ejemplo de aplicación nos lo proporciona la teoria clásica de la estructura factorial de la inteligencia. Los tests de aptitud dependen de un conjunto elevado de factores primarios, que dependen de un conjunto de 7 factores secundarios (verbal, numérico, espacial, razonamiento, memoria, percepción, psicomotores), que a su vez dependen de un factor general "g" (el factor "g" de Spearman), que sintetiza el hecho de que todas las aptitudes mentales están correlacionadas.

6.8 Medición de factores

Sea $\mathbf{x} = (x_1, \ldots, x_p)'$ los valores de las p variables observables obtenidas sobre un individuo ω. Nos planteamos ahora "medir los factores", es decir, encontrar los valores $\mathbf{f} = (f_1, \ldots, f_m)'$ de los factores comunes sobre ω. Se verifica

$$\mathbf{x} = \mathbf{Af} + \mathbf{Du}, \tag{6.18}$$

siendo $\mathbf{u} = (u_1 \ldots, u_p)'$ los valores de las unicidades.

Si interpretamos (6.18) como un modelo lineal, donde $\mathbf{x}$ es el vector de observaciones, $\mathbf{A}$ es la matriz de diseño, $\mathbf{f}$ es el vector de parámetros y $\mathbf{e} = \mathbf{Du}$ es el término de errror, el criterio de los mínimos cuadrado (véase (12.4)) nos da

$$\mathbf{f} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{x}.$$

Un método más elaborado (propuesto por M. S. Bartlett) considera que $\mathbf{f}$ es función lineal de $\mathbf{x}$ y que los valores de los factores únicos

$$\mathbf{u} = \mathbf{D}^{-1}(\mathbf{x} - \mathbf{Af})$$

son términos de error. Si queremos minimizar

$$\mathbf{u}'\mathbf{u} = u_1^2 + \ldots + u_p^2,$$

expresando (6.18) como $\mathbf{D^{-1}x} = \mathbf{D^{-1}Af} + \mathbf{u}$, es fácil ver que

$$\mathbf{f} = (\mathbf{A}'\mathbf{D}^{-2}\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}^{-2}\mathbf{x}.$$

Una modificación de este método (propuesta por T. W. Anderson y H. Rubin) consiste en añadir la condición de que los factores comunes estimados estén incorrelacionados. La solución que resulta es

$$\mathbf{f} = \mathbf{B}^{-1}\mathbf{A}'\mathbf{D}^{-2}\mathbf{x},$$

siendo $\mathbf{B}^2 = \mathbf{A}'\mathbf{D}^{-2}\mathbf{R}\mathbf{D}^{-2}\mathbf{A}$.

Ejemplo: Continuando con el ejemplo de las 5 asignaturas, las calificaciones en las asignatures de los 4 primeros alumnos (Tabla 6.1) y las puntuaciones (Anderson-Rubin) en los factores C y L, obtenidos con la rotación varimax, son:

Alumno	CNa	Mat	Fra	Lat	Lit	C	L
1	7	7	5	5	6	1.06	-.559
2	5	5	6	6	5	-.568	.242
3	5	6	5	7	5	.259	-.505
4	6	8	5	6	6	1.85	-.614

Teniendo en cuenta que los factores comunes son variables estandarizadas, el primer alumno tiene una nota relativamente alta en Ciencias y una nota algo por debajo de la media en Letras.

6.9 Análisis factorial confirmatorio

Los métodos del factor principal y de la máxima verosimilitud son exploratorios, en el sentido de que exploran las dimensiones latentes de las variables. El AF también se puede plantear en sentido confirmatorio, estableciendo una estructura factorial de acuerdo con el problema objeto de estudio, y seguidamente aceptando o rechazando esta estructura mediante un test de hipótesis. Por ejemplo, podemos considerar que la matriz factorial en el ejemplo de las 5 asignaturas es

	C	L
CNa	1	0
Mat	1	0
Fra	0	1
Lla	0	1
Lit	0	1

interpretando que las dos primeras sólo dependen del factor Ciencias y las otras tres del factor Letras. Entonces podemos realizar una transformación de la matriz factorial inicial para ajustarnos a la matriz anterior.

Si la solución inicial es $\mathbf{A}$, postulamos una estructura $\mathbf{B}$ y deseamos encontrar $\mathbf{T}$ ortogonal tal que $\mathbf{AT}$ se aproxime a $\mathbf{B}$ en el sentido de los mínimos cuadrados

$$\text{tr}(\mathbf{B} - \mathbf{AT})^2 = \text{mínimo},$$

entonces la solución es $\mathbf{T} = \mathbf{UV}'$, siendo $\mathbf{A}'\mathbf{B} = \mathbf{U\Lambda V}'$ la descomposición singular de $\mathbf{A}'\mathbf{B}$. Si $\mathbf{T}$ no es ortogonal y por lo tanto se admite una estructura oblicua, entonces $\mathbf{T}$ se obtiene siguiendo un procedimiento parecido a la rotación promax

$$\mathbf{T} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{B},$$

però normalizando a módulo 1 los vectores columna de $\mathbf{T}$.

Más generalmente, en AF confirmatorio se especifica el número de factores comunes, el tipo ortogonal u oblicuo de la solución, y los valores libres o fijos de las saturaciones.

Ejemplo: Un AF confirmatorio sobre 9 tests (estudiado por K. Joreskog) obtiene siete soluciones confirmatorias. De los 9 tests considerados, los tests 1,2,3 miden relaciones espaciales, los tests 4,5,6 inteligencia verbal y los tests 7,8,9 velocidad de percepción. La matriz de correlaciones es:

	1	2	3	4	5	6	7	8	9
1	1	.318	.468	.335	.304	.326	.116	.314	.489
2		1	.230	.234	.157	.195	.057	.145	.139
3			1	.327	.335	.325	.099	.160	.327
4				1	.722	.714	.203	.095	.309
5					1	.685	.246	.181	.345
6						1	.170	.113	.280
7							1	.585	.408
8								1	.512
9									1

Sólo comentaremos tres soluciones. La primera solución es oblicua no restringida, y se puede aceptar, puesto que la ji-cuadrado del ajuste no es significativa.

P			**Φ**			Comun.	
.71	.00	.00				.50	
.54	-.03	-.08				.26	
.67	.04	-.09				.46	
.00	.87	.00	1			.76	$\chi^2_{12} = 9.77$
-.03	.81	.13	.54	1		.70	$p = 0.64$
.01	.82	-.01	.24	.28	1	.68	
.00	.00	.78				.61	
.42	-.30	.73				.68	
.56	-.06	.41				.54	

La segunda solución es oblicua restringida. Se impone la condición de que los tres primeros tests correlacionen sólo con el primer factor, los tres siguientes sólo con el segundo y los tres últimos sólo con el tercero. No obstante, el valor ji-cuadrado es significativo y esta solución no debería aceptarse.

P			**Φ**			Comun.	
.68	.00	.00				.46	
.52	.00	.00				.27	
.69	.00	.00				.48	
.00	.87	.00	1			.77	$\chi^2_{24} = 51.19$
.00	.83	.00	.54	1		.69	$p = 0.001$
.00	.83	.00	.52	.34	1	.69	
.00	.00	.66				.43	
.00	.00	.80				.63	
.00	.00	.70				.49	

La tercera solución es ortogonal no restringida, con un factor general y tres factores específicos, en el sentido que el primero no correlaciona con la variable 4, el segundo no correlaciona con las variables 1 y 7 y el tercero no correlaciona con 1,2 y 4. El valor ji-cuadrado indica que esta solución es aceptable.

P				**Φ**				Comun.	
.38	.58	.00	.00					.48	
.24	.41	.35	.00					.37	
.38	.53	.30	-.03	1				.52	
.87	.00	.03	.00	.00	1			.75	$\chi^2_6 = 2.75$
.83	.01	-.13	.06	.00	.00	1		.72	$p = 0.84$
.83	.01	.04	-.02	.00	.00	.00	1	.68	
.24	.02	.00	.95					.95	
.15	.43	-.13	.57					.56	
.36	.59	-.22	.34					.64	

6.10 Complementos

Constituyen dos precedentes del Análisis Factorial el concepto de factor latente de F. Galton y de eje principal de K. Pearson. El primer trabajo, publicado en 1904, por Ch. Spearman (Spearman, 1904) desarrolla una teoría de la inteligencia alrededor de un factor común, el factor "g". Esta teoría,

que ordenaba la inteligencia de los individuos a lo largo de una sola dimensión, fue defendida por C. Burt, con consecuencias sociológicas importantes, pues proporcionó una base científica para financiar las escuelas privadas en detrimento de otras.

El Análisis Factorial moderno se inicia con la obra "Multiple Factor Analysis" de L.L. Thurstone, que postulaba más de un factor común, introducía la estructura simple y las rotaciones de factores. A partir de Thurstone la medida de la inteligencia era más "democrática", ya que poseía varias dimensiones latentes, quedando sin sentido una ordenación de los individuos, que si en una dimensión era posible hacerlo, en varias dimensiones no. Había una polémica similar sobre la personalidad. La teoria psicoanalítica defendía una continuidad entre la personalidad neurótica y la psicótica, mientras que el AF revela que neurosis y psicosis son dimensiones independientes.

Los modelos y métodos de Spearman, Burt, Thurstone y otros (Holzinger, Harman y Horst), son ya historia. Los métodos actuales para obtener la matriz factorial son: factor principal, análisis factorial canónico (C.R. Rao), método Alfa (H.F. Kaiser, J. Caffrey) y el método de la máxima verosimilitud (D.N. Lawley, K.G. Joreskog). Véase Joreskog (1967).

El método varimax de rotación ortogonal de Kaiser es uno de los más recomendados. J.B. Carroll introdujo la rotación oblicua quartimin y A.E. Hendrickson y P.O. White la promax. Anderson y Rubin (1956) publicaron un excelente trabajo sobre AF, tratando todo los aspectos algebraicos y estadísticos del tema. Véase Harman (1976), Torrens-Ibern (1972).

El estudio de las dimensiones latentes es un tema presente en la ciencia y siempre ha despertado interés. C. R. Rao demostró que si conocemos la distribución de k combinaciones lineales de p variables independientes, siendo $k(k-1)/2 < p \leq k(k+1)/2$, entonces la distribución de cada una de las p variables queda determinada (salvo la media o parámetro de localización). Si tenemos $p = 210$ variables independientes bastaría conocer la distribución de $k = 20$ combinaciones lineales adecuadas para determinar la distribución de las 210 variables. Este resultado proporciona una cierta justificación teórica acerca del hecho que la información multivariante posee una dimensionalidad latente mucha más pequeña.

La etapa inicial del AF (hasta 1966), era exploratoria, como una herramienta para explorar la dimensionalidad latente de las variables. Más tarde, el análisis factorial se ha entendido en sentido confirmatorio (Joreskog, Lawley, Maxwell, Mulaik), estableciendo una estructura factorial de acuerdo con el problema, y seguidamente aceptando o rechazando esta estructura

mediante un test de hipótesis (Joreskog, 1969, 1970). Consúltese Cuadras (1981).

Se han llevado a cabo muchas aplicaciones del AF. Citaremos tres, las dos primeras sobre AF exploratorio y la tercera sobre AF confirmatorio.

Rummel (1963) estudia 22 medidas de los conflictos de 77 naciones y encuentra tres dimensiones latentes, que identifica como: agitación, revolución y subversión, y ordena las naciones según las puntuaciones en los factors comunes.

Sánchez-Turet y Cuadras (1972) adaptan el cuestionario E.P.I. de personalidad (Eysenck Personality Inventory) y sobre un test de 69 ítems (algunos ítems detectan mentiras) encuentran tres factores: Introversión-Extroversión, Estabilidad-Inestabilidad, Escala de mentiras.

Joreskog (1969) explica un ejemplo de AF confirmatorio sobre 9 tests, previamente estudiado por Anderson y Rubin. Véase la Sección 6.9.

Finalmente, el Análisis de Estructuras Covariantes es una generalización del AF, que unifica este método con otras técnicas multivariantes (MANOVA, análisis de componentes de la varianza, análisis de caminos, modelos simplex y circumplexos, etc.). Se supone que la estructura general para la matriz de covarianzas es

$$\Sigma = \mathbf{B}(\mathbf{P}\mathbf{\Phi}\mathbf{P}' + \mathbf{D}^2)\mathbf{B}' + \Theta^2.$$

Otra generalización es el llamado modelo LISREL (Linear Structural Relationship), que permite relacionar un grupo de variables dependientes $\mathbf{Y}$ con un grupo de variables independientes $\mathbf{X}$, que dependen de unas variables latentes a través de un modelo de medida. Las variables latentes están relacionadas por un modelo de ecuaciones estructurales. LISREL (Joreskog y Sorbom, 1999) es muy flexible y tiene muchas aplicaciones (sociología, psicología, economía). Véase Satorra (1989), Batista y Coenders (2000).

Capítulo 7

ANÁLISIS CANÓNICO DE POBLACIONES

7.1 Introducción

Con el Análisis de Componentes Principales podemos representar los individuos de una población, es decir, representar una única matriz de datos. Pero si tenemos varias matrices de datos, como resultado de observar las variables sobre varias poblaciones, y lo que queremos es representar las poblaciones, entonces la técnica adecuada es el Análisis Canónico de Poblaciones (CANP).

Supongamos que de la observación de p variables cuantitativas $X_1, \ldots, X_p$ sobre g poblaciones obtenemos g matrices de datos

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_g \end{pmatrix} \begin{matrix} n_1 \times p \\ n_2 \times p \\ \vdots \\ n_g \times p \end{matrix}$$

donde $\mathbf{X}_i$ es la matriz $n_i \times p$ de la población i. Sean $\overline{\mathbf{x}}'_1, \overline{\mathbf{x}}'_2, \ldots, \overline{\mathbf{x}}'_g$ los vectores (fila) de las medias de cada población. $\mathbf{X}$ es de orden $n \times p$, siendo $n = \sum_{i=1}^{g} n_i$. Indiquemos

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{\mathbf{x}}'_1 - \overline{\mathbf{x}}' \\ \overline{\mathbf{x}}'_2 - \overline{\mathbf{x}}' \\ \vdots \\ \overline{\mathbf{x}}'_g - \overline{\mathbf{x}}' \end{pmatrix}$$

la matriz $g \times p$ con las medias de las g poblaciones. Tenemos dos maneras de cuantificar matricialmente la dispersión entre las poblaciones:

- La matriz de dispersión no ponderada entre grupos

$$\mathbf{A} = \overline{\mathbf{X}}'\overline{\mathbf{X}} = \sum_{i=1}^{g} (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'.$$

- La matriz de dispersión ponderada entre grupos

$$\mathbf{B} = \sum_{i=1}^{g} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'.$$

La matriz $\mathbf{A}$ es proporcional a una matriz de covarianzas tomando como datos sólo las medias de las poblaciones. La matriz $\mathbf{B}$ participa, juntamente con $\mathbf{W}$ (matriz de dispersión dentro de grupos) en el test de comparación de medias de g poblaciones. Aquí trabajaremos con la matriz $\mathbf{A}$, si bien los resultados serían parecidos si utilizáramos la matriz $\mathbf{B}$. También haremos uso de la matriz de covarianzas (véase (3.1)):

$$\mathbf{S} = \frac{1}{n-g} \sum_{i=1}^{g} n_i \mathbf{S}_i.$$

Entonces $\mathbf{A} = \overline{\mathbf{X}}'\overline{\mathbf{X}}$ juega el papel de matriz de covarianzas "entre" las poblaciones, $\mathbf{S}$ juega el papel de matriz de covarianzas "dentro" de las poblaciones.

7.2 Variables canónicas

Definició 7.2.1 *Sean* $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ *los vectores propios de* $\mathbf{A}$ *respecto de* $\mathbf{S}$ *con valores propios* $\lambda_1 > \ldots > \lambda_p$*, es decir,*

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{S}_i \mathbf{v}_i,$$

normalizados según

$$\mathbf{v}_i' \mathbf{S}_i \mathbf{v}_i = 1.$$

Los vectores $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *son los vectores canónicos y las variables canónicas son las variables compuestas*

$$Y_i = \mathbf{X}\mathbf{v}_i.$$

Si $\mathbf{v}_i = (v_{1i}, \ldots, v_{pi})'$ y $\mathbf{X} = [X_1, \ldots, X_p]$, la variable canónica Y_i es la variable compuesta

$$Y_i = \mathbf{X}\mathbf{v}_i = v_{1i}X_1 + \cdots + v_{pi}X_p$$

que tiene S-varianza 1 y $A-$varianza λ_i, es decir:

$$\mathrm{var}_A(Y_i) = \mathbf{v}_i'\mathbf{A}\mathbf{v}_i = \lambda_i, \quad \mathrm{var}_S(Y_i) = \mathbf{v}_i'\mathbf{S}_i\mathbf{v}_i = 1.$$

Trabajaremos con p variables canónicas, pero de hecho el número efectivo es $k = \min\{p, g-1\}$, ver Sección 7.5.3.

Teorema 7.2.1 *Las variables canónicas verifican:*

1. *Son incorrelacionadas dos a dos respecto a* $\mathbf{A}$ *y también respecto a* $\mathbf{S}$

$$cov_A(Y_i, Y_j) = cov_S(Y_i, Y_j) = 0 \quad si \quad i \neq j.$$

2. *Las A-varianzas son respectivamente máximas:*

$$var_A(Y_1) = \lambda_1 > \cdots > var_A(Y_p) = \lambda_p,$$

en el sentido de que Y_1 *es la variable con máxima varianza entre grupos, condicionada a varianza 1 dentro grupos,* Y_2 *es la variable con máxima varianza entre grupos, condicionada a estar incorrelacionada con* Y_1 *y tener varianza 1 dentro grupos, etc.*

Demost.: Supongamos $\lambda_1 > \cdots > \lambda_p > 0$. Probemos que las variables $Y_i = \mathbf{X}\mathbf{t}_i$, $i = 1, \ldots, p$, están incorrelacionadas:

$$\begin{array}{lll} \mathrm{cov}_A(Y_i, Y_j) & = \mathbf{t}_i'\mathbf{A}\mathbf{t}_j & = \mathbf{t}_i'\mathbf{S}\lambda_j\mathbf{t}_j \quad = \lambda_j\mathbf{t}_i'\mathbf{S}\mathbf{t}_j, \\ \mathrm{cov}_A(Y_j, Y_i) & = \mathbf{t}_j'\mathbf{A}\mathbf{t}_i & = \mathbf{t}_j'\mathbf{S}\lambda_j\mathbf{t}_i \quad = \lambda_i\mathbf{t}_j'\mathbf{S}\mathbf{t}_i, \end{array}$$

$\Rightarrow (\lambda_j - \lambda_i)\mathbf{t}_i'\mathbf{S}\mathbf{t}_j = 0 \Rightarrow \mathbf{t}_i'\mathbf{S}\mathbf{t}_j = 0 \Rightarrow \mathrm{cov}_A(Y_i, Y_j) = \lambda_j\mathbf{t}_i'\mathbf{S}\mathbf{t}_j = \mathrm{cov}_A(Y_i, Y_j) = 0$, si $i \neq j$. Además, de $\mathbf{t}_i'\mathbf{S}\mathbf{t}_j = 1$:

$$\mathrm{var}_A(Y_i) = \lambda_i\mathbf{t}_i'\mathbf{S}\mathbf{t}_j = \lambda_i.$$

Sea ahora $Y = \sum_{i=1}^p a_iX_i = \sum_{i=1}^p \alpha_iY_i$ una variable compuesta tal que $\mathrm{var}_S(Y) = \sum_{i=1}^p \alpha_i^2\mathrm{var}_S(Y_i) = \sum_{i=1}^p \alpha_i^2 = 1$. Entonces:

$$\mathrm{var}_A(Y) = \mathrm{var}_A(\sum_{i=1}^p \alpha_iY_i) = \sum_{i=1}^p \alpha_i^2\mathrm{var}_A(Y_i) = \sum_{i=1}^p \alpha_i^2\lambda_i \leq (\sum_{i=1}^p \alpha_i^2)\lambda_1 = \mathrm{var}_A(Y_1),$$

que prueba que Y_1 tiene máxima varianza entre grupos.

Consideremos a continuación las variables Y incorrelacionadas con Y_1, que podemos expresar como:

$$Y = \sum_{i=1}^{p} b_i X_i = \sum_{i=2}^{p} \beta_i Y_i \quad \text{condicionado a } \sum_{i=2}^{p} \beta_i^2 = 1.$$

Entonces:

$$\text{var}_A(Y) = \text{var}_A(\sum_{i=2}^{p} \beta_i Y_i) = \sum_{i=2}^{p} \beta_i^2 \text{var}_A(Y_i) = \sum_{i=2}^{p} \beta_i^2 \lambda_i \leq (\sum_{i=2}^{p} \beta_i^2)\lambda_2 = \text{var}_A(Y_2),$$

y por lo tanto Y_2 está incorrelacionada con Y_1 y tiene varianza máxima. La demostración de que $Y_3, \ldots, Y_p$ son también variables canónicas es análoga.

7.3 Distancia de Mahalanobis y transformación canónica

La distancia de Mahalanobis entre dos poblaciones es una medida natural de la diferencia entre las medias de las poblaciones, pero teniendo en cuenta las covarianzas. En la Sección 1.9 hemos introducido la distancia entre los individuos de una misma población. Ahora definimos la distancia entre dos poblaciones cuando hay más de dos poblaciones.

Definició 7.3.1 *Consideremos muestras multivariantes de g poblaciones con vectores de medias* $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \ldots, \overline{\mathbf{x}}_g$ *y matriz de covarianzas (común)* $\mathbf{S}$. *La distancia (al cuadrado) de Mahalanobis entre las poblaciones* i, j *es*

$$M^2(i,j) = (\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)' \mathbf{S}^{-1} (\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j).$$

Si $\overline{\mathbf{X}}$ es la matriz centrada con los vectores de medias y $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ es la matriz con los vectores canónicos (vectores propios de $\mathbf{A} = \overline{\mathbf{X}}' \overline{\mathbf{X}}$ respecto de $\mathbf{S}$), la transformación canónica es

$$\mathbf{Y} = \overline{\mathbf{X}} \mathbf{V}.$$

La matriz $\mathbf{Y}$ de orden $g \times p$ contiene las coordenadas canónicas de las g poblaciones.

Teorema 7.3.1 *La distancia de Mahalanobis entre cada par de poblaciones i, j coincide con la distancia Euclídea entre las filas i, j de la matriz de coordenadas canónicas* $\mathbf{Y}$. *Si* $\mathbf{y}_i = \overline{\mathbf{x}}_i\mathbf{V}$ *entonces*

$$d_E^2(i,j) = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = (\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j). \tag{7.1}$$

Demost.: Basta probar que los productos escalares coinciden

$$\mathbf{y}_i\mathbf{y}_j' = \overline{\mathbf{x}}_i\mathbf{S}^{-1}\overline{\mathbf{x}}_j' \Longleftrightarrow \overline{\mathbf{X}}\mathbf{S}^{-1}\overline{\mathbf{X}}' = \mathbf{Y}\mathbf{Y}'.$$

Sea Λ =diag$(\lambda_1, \ldots, \lambda_p)$ la matriz diagonal con los valores propios de $\mathbf{A} = \overline{\mathbf{X}}'\overline{\mathbf{X}}$ respecto de $\mathbf{S}$. Entonces

$$\mathbf{AV} = \mathbf{SV}\Lambda \quad \text{con} \quad \mathbf{V}'\mathbf{SV} = \mathbf{I}_p,$$

y la transformación canónica es $\mathbf{Y} = \overline{\mathbf{X}}\mathbf{V}$.

Sea $\mathbf{C}$ matriz ortogonal definida por $\mathbf{V} = \mathbf{S}^{-1/2}\mathbf{C}$, siendo $\mathbf{S} = \mathbf{UDU}'$, con $\mathbf{D}$ diagonal y $\mathbf{S}^{-1/2} = \mathbf{UD}^{-1/2}\mathbf{U}'$. Tenemos que $\mathbf{V}'\mathbf{AV} = \mathbf{V}'\mathbf{SV}\Lambda$ es

$$\mathbf{C}'\mathbf{S}^{-1/2}\mathbf{A}\mathbf{S}^{\mathbf{-1/2}}\mathbf{C} = \mathbf{C}'\mathbf{S}^{-1/2}\mathbf{S}\mathbf{S}^{-1/2}\mathbf{C}\Lambda = \Lambda,$$

es decir, $\mathbf{S}^{\mathbf{-1/2}}\mathbf{C}$ contiene los vectores propios de $\mathbf{A}$ con valores propios Λ. Entonces $\mathbf{AV} = \mathbf{SV}\Lambda$ implica

$$\mathbf{A}\mathbf{S}^{\mathbf{-1/2}}\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{C}\Lambda \quad \text{con} \quad \mathbf{C}'\mathbf{C} = \mathbf{C}\mathbf{C}' = \mathbf{I}_p.$$

La transformación canónica es pues $\mathbf{Y} = \overline{\mathbf{X}}\mathbf{S}^{-1/2}\mathbf{C}$, así que

$$\overline{\mathbf{X}}\mathbf{S}^{-1}\overline{\mathbf{X}}' = \overline{\mathbf{X}}\mathbf{S}^{-1/2}\mathbf{C}\mathbf{C}'\mathbf{S}^{-1/2}\overline{\mathbf{X}}' = \mathbf{Y}\mathbf{Y}'.\square$$

7.4 Representación canónica

La representación de las g poblaciones mediante las filas de $\overline{\mathbf{X}}$ con la métrica de Mahalanobis es bastante complicada: la dimensión puede ser grande y los ejes son oblícuos. En cambio, la representación mediante las coordenadas canónicas $\mathbf{Y}$ con la métrica Euclídea se realiza a lo largo de ejes ortogonales. Si además, tomamos las q primeras coordenadas canónicas ($q = 2$, por ejemplo), la representación es totalmente factible y es óptima en dimensión reducida, en el sentido de que maximiza la variabilidad geométrica.

Teorema 7.4.1 *La variabilidad geométrica de las distancias de Mahalanobis entre las poblaciones es proporcional a la suma de los valores propios:*

$$V_M(\overline{\mathbf{X}}) = \frac{1}{2g^2}\sum_{i,j=1}^{g} M(i,j)^2 = \frac{1}{g}\sum_{i=1}^{p}\lambda_i. \tag{7.2}$$

Si $\mathbf{Y} = \overline{\mathbf{X}}\mathbf{V}$, *donde* $\mathbf{V}$, *de orden* $p \times q$ *es la matriz de la transformación canónica en dimensión* q *y*

$$\delta_{ij}^2(q) = (\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)' = \sum_{h=1}^{q}(y_{ih} - y_{jh})^2$$

es la distancia Euclídea (al cuadrado) entre dos filas de $\mathbf{Y}$, *la variabilidad geométrica en dimensión* $q \leq p$ *es*

$$V_\delta(\mathbf{Y})_q = \frac{1}{2g^2}\sum_{i,j=1}^{g}\delta_{ij}^2(q) = \frac{1}{g}\sum_{i=1}^{q}\lambda_i,$$

y esta cantidad es máxima entre todas las transformaciones lineales en dimensión q.

Demost.: De (5.3) y (7.1)

$$V_M(\mathbf{X}) = \frac{1}{2g^2}\sum_{i,j=1}^{g} M(i,j)^2 = \frac{1}{2g^2}\sum_{i,j=1}^{g}\sum_{h=1}^{p}(y_{ih} - y_{jh})^2 = s_1^2 + \ldots + s_p^2$$

donde $s_j^2 = (\sum_{i=1}^{g} y_{ij}^2)/g$ representa la varianza ordinaria de la columna Y_j de $\mathbf{Y}$. Además

$$\frac{1}{g}\mathbf{Y}'\mathbf{Y} = \frac{1}{g}\mathbf{V}'\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{V} = \frac{1}{g}\mathbf{V}'\mathbf{A}\mathbf{V} = \frac{1}{g}\mathbf{\Lambda}$$

y por lo tanto $s_j^2 = \lambda_j/g$, lo que prueba (7.2).

Sea ahora $\widetilde{\mathbf{Y}} = \overline{\mathbf{X}}\mathbf{T}$ una transformación cualquiera tal que $\mathbf{T}'\mathbf{S}\mathbf{T} = \mathbf{I}$. Es decir, si

$$\overline{\mathbf{X}} = [\overline{X}_1, \ldots, \overline{X}_p] \rightarrow \widetilde{\mathbf{Y}} = \overline{\mathbf{X}}\mathbf{T} = [\widetilde{Y}_1, \ldots, \widetilde{Y}_p]$$

donde $\overline{X}_j, \widetilde{Y}_j$ son las columnas de $\overline{\mathbf{X}}, \widetilde{\mathbf{Y}}$, que son matrices centradas, y

$$\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_p] = \begin{pmatrix} t_{11} & \cdots & t_{1p} \\ \vdots & \ddots & \vdots \\ t_{p1} & \cdots & t_{pp} \end{pmatrix},$$

entonces $\widetilde{Y}_k = \overline{\mathbf{X}}\mathbf{t}_k = t_{1k}\overline{X}_1 + \ldots + t_{pk}\overline{X}_p$ tiene A-varianza

$$\mathrm{var}_A(\widetilde{Y}_k) = \mathbf{t}_k'\mathbf{A}\mathbf{t}_k = \mathbf{t}_k'\overline{\mathbf{X}}'\overline{\mathbf{X}}\mathbf{t}_k = \widetilde{Y}_k'\widetilde{Y}_k = g \cdot s^2(\widetilde{Y}_k)$$

donde $s^2(\widetilde{Y}_k)$ indica la varianza ordinaria. Puesto que la A-varianza máxima es λ_k, tenemos:

$$V_\delta(\widetilde{\mathbf{Y}})_q = \sum_{k=1}^{g} s^2(\widetilde{Y}_k) = \frac{1}{g}\sum_{k=1}^{g} \mathrm{var}_A(\widetilde{Y}_k) \leq \frac{1}{g}\sum_{k=1}^{g} \lambda_k.$$

El porcentaje de variabilidad geométrica explicada por las q primeras coordenadas canónicas es

$$P_q = 100\frac{V(\mathbf{Y})_q}{V_M(\overline{\mathbf{X}})} = 100\frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_p}.$$

7.5 Aspectos inferenciales

Supongamos que las matrices de datos $\mathbf{X}_1, \ldots, \mathbf{X}_g$ provienen de g poblaciones normales $N_p(\mu_1, \Sigma_1), \ldots, N_p(\mu_g, \Sigma_g)$. Para poder aplicar correctamente un análisis canónico de poblaciones conviene que los vectores de medias sean diferentes y que las matrices de covarianzas sean iguales.

7.5.1 Comparación de medias

El test

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g \tag{7.3}$$

ha sido estudiado en la Sección 3.3.3 y se decide calculando el estadístico $\Lambda = |\mathbf{W}|/|\mathbf{B} + \mathbf{W}|$ con distribución lambda de Wilks. Si aceptamos H_0 las medias de las poblaciones son teóricamente iguales y el análisis canónico, técnica destinada a representar las medias de las poblaciones a lo largo de ejes canónicos, no tiene razón de ser. Por lo tanto, conviene rechazar H_0.

7.5.2 Comparación de covarianzas

El test

$$H_0' : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_g$$

se resuelve mediante el test de razón de verosimilitud

$$\lambda_R = \frac{|\mathbf{S}_1|^{n_1/2} \times \cdots \times |\mathbf{S}_g|^{n_g/2}}{|\mathbf{S}|^{n/2}},$$

donde $\mathbf{S}_i$ es la matriz de covarianzas de las datos de la población i, estimación máximo verosímil de Σ_i y

$$\mathbf{S} = (n_1\mathbf{S}_1 + \cdots + n_g\mathbf{S}_g)/n = \mathbf{W}/n$$

es la estimación máximo verosímil de Σ, matriz de covarianzas común bajo H_0'. Rechazaremos H_0' si el estadístico

$$-2\log\lambda_R = n\log|\mathbf{S}| - (n_1\log|\mathbf{S}_1| + \cdots + n_g\log|\mathbf{S}_g|) \sim \chi^2_q$$

es significativo, donde $q = gp(p+1)/2 - p(p+1)/2 = (g-1)p(p+1)/2$ son los grados de libertad de la ji-cuadrado. Si rechazamos H_0', entonces resulta que no disponemos de unos ejes comunes para representar todas las poblaciones (la orientación de los ejes viene dada por la matriz de covarianzas), y el análisis canónico es teóricamente incorrecto. Conviene pues aceptar H_0'.

Debido a que el test anterior puede ser sesgado, conviene aplicar la corrección de Box,

$$c \cdot (n-g)\log|\mathbf{S}| - ((n_1-1)\log|\widehat{\mathbf{S}}_1| + \cdots + (n_g-1)\log|\widehat{\mathbf{S}}_g|)$$

donde $\widehat{\mathbf{S}}_i = (n_i/(n_i-1))\mathbf{S}_i$, y la constante c es

$$c = [1 - (\frac{2p^2+3p-1}{6(p+1)(g-1)})(\sum_{k=1}^{g}\frac{1}{n_g-1} - \frac{1}{n-g})].$$

7.5.3 Test de dimensionalidad

Como el rango de $\mathbf{A} = \overline{\mathbf{X}}'\overline{\mathbf{X}}$ no puede superar ni la dimensión p ni $g-1$, es obvio que el número efectivo de valores propios es

$$k = \min\{p, g-1\}.$$

Si los vectores de medias poblacionales están en un espacio R^m de dimensión $m < k$, entonces el espacio canónico tiene dimensión m y por lo tanto debemos aceptar la hipótesis

$$H_0^{(m)} : \lambda_1 > \ldots > \lambda_m > \lambda_{m+1} = \ldots = \lambda_k,$$

donde $\lambda_1 > \ldots > \lambda_m$ son los valores propios de $\overline{\mathbf{MM}}'$ (la versión poblacional de $\mathbf{A}$) respecto de Σ. Si

$$l_1 > \ldots > l_k$$

son los valores propios de $\mathbf{B}$ respecto de $\mathbf{W}$ (ver Sección 3.3.3), es decir, soluciones de

$$|\mathbf{B} - l\mathbf{W}| = 0,$$

entonces un test para decidir $H_0^{(m)}$ está basado en el estadístico

$$b_m = [n - 1 - \frac{1}{2}(p+g)] \sum_{i=m+1}^{k} \log(1 + l_i) \sim \chi_q^2,$$

donde $q = (p-m)(g-m-1)$. Este test asintótico, propuesto por Bartlett, se aplica secuencialmente: si b_0 es significativo, estudiaremos b_1; si b_1 es también significativo, estudiaremos b_2, etc. Si $b_0, \ldots, b_{m-1}$ son significativos pero b_m no, aceptaremos que la dimensión es m. Obsérvese que aceptar $H_0^{(0)}$ equivale a la hipótesis nula de igualdad de vectores de medias (que entonces coincidirían en un punto), es decir, equivale a aceptar (7.3).

Otros autores utilizan este test independienmente para cada dimensión. Así, el test $H_0 : \lambda_j = 0$ está basado en el estadístico

$$c_j = [n - 1 - \frac{1}{2}(p+g)] \log(1 + l_j) \sim \chi_r^2,$$

donde $r = p + g - 2j$ son los grados de liberdad. Rechazaremos H_0 si c_j es significativo.

7.5.4 Regiones confidenciales

Sean $\overline{\mathbf{y}}_i' = \overline{\mathbf{x}}_i'\mathbf{V}, i = 1, \ldots, g$ las proyecciones canónicas de los vectores de medias muestrales de las poblaciones. Podemos entender $\overline{\mathbf{y}}_i$ como una estimación de $\mu_i^* = \mu_i V$, la proyección canónica del vector de medias poblacional μ_i. Queremos encontrar regiones confidenciales para $\mu_i^*, i = 1, \ldots, g$.

Teorema 7.5.1 *Sea $1 - \alpha$ el coeficiente de confianza, F_α tal que $P(F > F_\alpha) = \alpha$, donde F sigue la distribución F con p y $(n - g - p + 1)$ g.l. y consideremos:*

$$R_\alpha^2 = F_\alpha \frac{(n-g)p}{(n-g-p+1)}.$$

Entonces las proyecciones canónicas μ_i^* de los vectores de medias poblacionales pertenecen a regiones confidenciales que son hiperesferas (esferas en dimensión 3, círculos en dimensión 2) de centros y radios

$$(\overline{\mathbf{y}}_i, R_\alpha/\sqrt{n_i}),$$

donde n_i es el tamaño muestral de la población i.

Demost.: $\overline{\mathbf{x}}_i - \mu_i$ es $N_p(\mathbf{0}, \Sigma/n_i)$ independiente de $\mathbf{W}$ que sigue la distribución $W_p(\Sigma, n-g)$. Por lo tanto

$$\begin{aligned}(n-g)n_i(\overline{\mathbf{x}}_i - \mu_i)'\mathbf{W}^{-1}(\overline{\mathbf{x}}_i - \mu_i) \\ = n_i(\overline{\mathbf{x}}_i - \mu_i)\mathbf{S}^{-1}(\overline{\mathbf{x}}_i - \mu_i)' \sim \quad T^2(p, n-g),\end{aligned}$$

y como la distribución de Hotelling equivale a una F, tenemos que

$$(\overline{\mathbf{x}}_i - \mu_i)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_i - \mu_i) \sim \frac{(n-g)p}{n_i(n-g-p+1)} F^p_{n-g-p+1}.$$

Así pues

$$P[(\overline{\mathbf{x}}_i - \mu_i)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_i - \mu_i) \leq \frac{R_\alpha^2}{n_i}] = 1-\alpha,$$

que define una región confidencial hiperelíptica para μ_i con coeficiente de confianza $1-\alpha$. Pero la transformación canónica $\overline{\mathbf{y}}_i' = \overline{\mathbf{x}}_i'\mathbf{V}$ convierte $(\overline{\mathbf{x}}_i - \mu_i)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_i - \mu_i)$ en $(\overline{\mathbf{y}}_i - \mu_i^*)'(\overline{\mathbf{y}}_i - \mu_i^*)$ y por lo tanto

$$P[(\overline{\mathbf{y}}_i - \mu_i^*)'(\overline{\mathbf{y}}_i - \mu_i^*) \leq \frac{R_\alpha^2}{n_i}] = 1-\alpha.$$

Esta transformación convierte además hiperelipses en hiperesferas (elipses en círculos si la dimensión es 2), ya que las variables canónicas son incorrelacionadas, lo que también es válido si reducimos la dimensión (tomamos las m primeras coordenadas canónicas).

Por ejemplo, si elegimos $1-\alpha = 0.95$ y una representación en dimensión reducida 2, cada población vendrá representada por un círculo de centro $\overline{\mathbf{y}}_i$ y radio $R_{0.05}/\sqrt{n_i}$, de manera que el vector de medias proyectado pertenece al círculo con coeficiente de confianza 0.95. La separación entre los centros indicará diferencias, mientras que si dos círculos se solapan, será indicio de que las dos poblaciones son posiblemente iguales.

Exemple 7.5.1

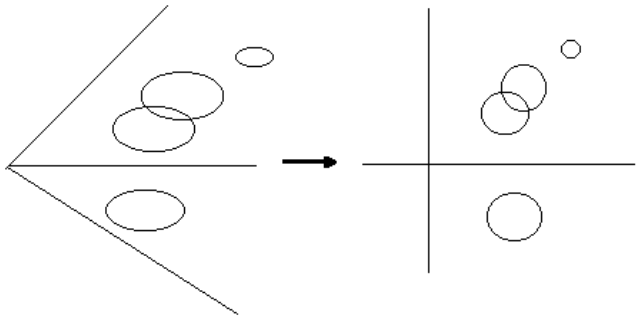

Figura 7.1: Proyeción canónica de cuatro poblaciones.

Se tienen medidas de 5 variables biométricas sobre coleópteros del género *Timarcha* de 5 especies encontradas en 8 localidades:

1. *T. sinustocollis* (Campellas, Pirineos) $n_1 = 40$.
2. *T. sinustocollis* (Planollas, Pirineos) $n_2 = 40$.
3. *T. indet* (vall de Llauset, Pirineos, Osca) $n_3 = 20$.
4. *T. monserratensis* (Collformic, Barcelona) $n_4 = 40$.
5. *T. monserratensis* (Collfsuspina, Barcelona) $n_5 = 40$.
6. *T. catalaunensis* (La Garriga, Barcelona) $n_6 = 40$.
7. *T. balearica* (Mahón, Baleares) $n_7 = 15$
8. *T. pimeliodes* (Palermo, Sicilia) $n_8 = 40$

Las medidas (en mm.) son:

X_1= long. prognoto, X_2=diam. máximo prognoto, X_3= base prognoto, X_4= long. élitros, X_5 = diam. máximo élitros.

Se quiere estudiar si existen diferencias entre las 8 especies y representarlas mediante la distancia de Mahalanobis. Los resultados del análisis canónico son:

- Matriz de covarianzas común:

$$\mathbf{S} = \begin{pmatrix} 3.277 & 3.249 & 2.867 & 5.551 & 4.281 \\ & 7.174 & 6.282 & 9.210 & 7.380 \\ & & 6.210 & 8.282 & 6.685 \\ & & & 20.30 & 13.34 \\ & & & & 13.27 \end{pmatrix}$$

- Test de Bartlett para homogeneidad de la matriz de covarianzas. Ji-cuadrado = 229.284, con 105 g.l. Significativo al 5%.

- Matriz de dispersión entre grupos:

$$\mathbf{B} = \begin{pmatrix} 6268 & 11386 & 8039 & 22924 & 17419 \\ & 21249 & 15370 & 42795 & 32502 \\ & & 11528 & 31009 & 23475 \\ & & & 86629 & 65626 \\ & & & & 49890 \end{pmatrix} \sim W_4(7, \Sigma)$$

- Matriz de dispersión dentro de grupos:

$$\mathbf{W} = \begin{pmatrix} 874.8 & 867.5 & 765.4 & 1482 & 1142 \\ & 1915 & 1677 & 2458.99 & 1970 \\ & & 1658 & 2211 & 1784 \\ & & & 5419 & 3562 \\ & & & & 3541 \end{pmatrix} \sim W_5(267, \Sigma)$$

- Matriz de dispersión total:

$$\mathbf{T} = \begin{pmatrix} 7143 & 12253 & 8804 & 24407 & 18562 \\ & 23164 & 17047 & 45254 & 34472 \\ & & 13186 & 33220 & 25260 \\ & & & 92049 & 69189 \\ & & & & 53432 \end{pmatrix}$$

- Test de comparación de medias:

  $\Lambda = |\mathbf{W}| \,/\, |\mathbf{B} + \mathbf{W}| = 0.0102 \sim \Lambda(5, 267, 7) \rightarrow F = 62.5$ (35 y 1108 g.l.)

  Existen diferencias muy significativas.

- Transformación canónica, valores propios y porcentaje acumulado:

	$\mathbf{v}_1$	$\mathbf{v}_2$
	-.0292	.2896
	.5553	.7040
	-.6428	-.9326
	.1259	-.1326
	.1125	.0059
λ	158.64	24.53
%	85.03	98.18

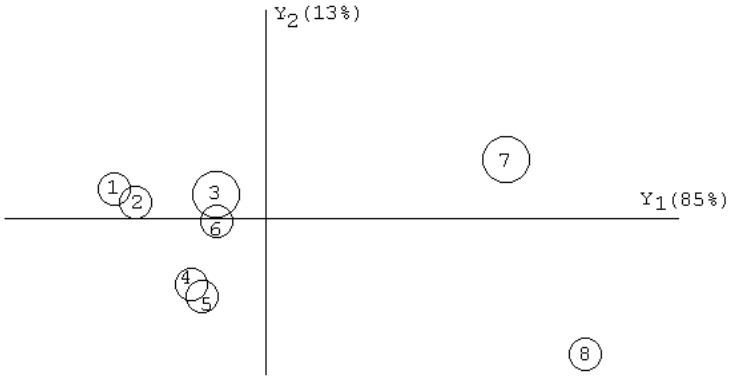

Figura 7.2: Representación canónica de 8 especies de coleópteros.

De acuerdo con la Fig. 7.2, las poblaciones 1 y 2 pertenecen claramente a la misma especie, así como la 4 y 5. Las poblaciones 3 y 6 son especies próximas, mientras que las 7 y 8 se diferencian mucho de las otras especies.

7.6 Complementos

El Análisis Canónico de Poblaciones (CANP) fué planteado por M.S. Bartlett en términos de correlación canónica entre las poblaciones y las variables observables. C. R. Rao lo relacionó con la distancia de Mahalanobis y lo estudió como una técnica para representar poblaciones. Su difusión es debido a Seal (1964).

Existen diferentes criterios para obtener la región confidencial para las medias de las poblaciones. Aquí hemos seguido un criterio propuesto por Cuadras (1974). Una formulación que no supone normalidad es debido a Krzanowski y Radley (1989). A menudo los datos no cumplen la condición de igualdad de las matrices de covarianzas, aunque el CANP es válido si las matrices muestrales son relativamente semejantes.

En el CANP, y más adelante en el Análisis Discriminante, interviene la descomposición $\mathbf{T} = \mathbf{B} + \mathbf{W}$, es decir:

$$\sum_{i=1}^{g}\sum_{h=1}^{n_i}(\mathbf{x}_{ih}-\overline{\mathbf{x}})(\mathbf{x}_{ih}-\overline{\mathbf{x}})'=\sum_{i=1}^{g} n_i(\overline{\mathbf{x}}_i-\overline{\mathbf{x}})(\overline{\mathbf{x}}_i-\overline{\mathbf{x}})'+\sum_{i=1}^{g}\sum_{h=1}^{n_i}(\mathbf{x}_{ih}-\overline{\mathbf{x}}_i)(\mathbf{x}_{ih}-\overline{\mathbf{x}}_i)'.$$

Si los datos provienen de g poblaciones con densidades $f_i(\mathbf{x})$, medias y matrices de covarianzas (μ_i, Σ_i) y probabilidades $p_i, i = 1, \ldots, g$, es decir, con densidad

$$f(\mathbf{x}) = p_1 f_1(\mathbf{x}) + \ldots + p_g f_g(\mathbf{x}),$$

entonces el vector de medias correspondiente a f es

$$\mu = p_1\mu_1 + \ldots + p_g\mu_g,$$

y la matriz de covarianzas es

$$\Sigma = \sum_{i=1}^{g} p_i(\mu_i - \mu)(\mu_i - \mu)' + \sum_{i=1}^{g} p_i\Sigma_i.$$

Esta descomposición de Σ es la versión poblacional de $\mathbf{T} = \mathbf{B} + \mathbf{W}$, y la versión multivariante de

$$\text{var}(Y) = E[\text{var}[Y|X]] + \text{var}[E[Y|X]],$$

donde $Y|X$ representa la distribución de una variable Y dada X. Ver Flury (1997).

Capítulo 8

ESCALADO MULTIDIMENSIONAL (MDS)

8.1 Introducción

Representar un conjunto finito cuando disponemos de una distancia entre los elementos del conjunto, consiste en encontrar unos puntos en un espacio de dimensión reducida, cuyas distancias euclídeas se aproximen lo mejor posible a las distancias originales.

Sea $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ un conjunto finito con n elementos diferentes, que abreviadamente indicaremos

$$\Omega = \{1, 2, ..., n\}.$$

Sea $\delta_{ij} = \delta(i,j) = \delta(j,i) \geq \delta(i,i) = 0$ una distancia o disimilaridad entre los elementos i, j de Ω. Consideremos entonces la matriz de distancias

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_n \end{pmatrix} \qquad \delta_{ij} = \delta_{ji} = \delta(i,j) \geq \delta_{ii} = 0.$$

Definició 8.1.1 *Diremos que $\Delta = (\delta_{ij})$ es una matriz de distancias Euclídeas si existen n puntos $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$, siendo*

$$\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip}), \quad i = 1, \ldots, n,$$

tales que

$$\delta_{ij}^2 = \sum_{\alpha=1}^{p}(x_{i\alpha} - x_{j\alpha})^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) \tag{8.1}$$

Indicaremos las coordenadas de los puntos $\mathbf{x}_1, \ldots, \mathbf{x}_n$, que representan los elementos $1, \ldots, n$ de Ω, en forma de matriz

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

El objetivo del *escalamiento multidimensional* es encontrar la X adecuada a partir de la matriz de distancias.

8.2 Cuando una distancia es euclídea?

Sea $\Delta^{(2)} = (\delta_{ij}^2)$ la matriz de cuadrados de las distancias. Si la distancia es euclídea entonces de (8.1)

$$\delta_{ij}^2 = \mathbf{x}_i'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_i'\mathbf{x}_j$$

La matriz de productos internos asociada a Δ es

$$\mathbf{G} = \mathbf{X}\mathbf{X}'.$$

Los elementos de $\mathbf{G} = (g_{ij})$ son $g_{ij} = \mathbf{x}_i'\mathbf{x}_j$. Relacionando $\Delta^{(2)} = (\delta_{ij}^2)$ con $\mathbf{G}$ vemos que

$$\Delta^{(2)} = \mathbf{1}\mathbf{g}' + \mathbf{g}\mathbf{1}' - 2\mathbf{G}, \tag{8.2}$$

donde $\mathbf{g} = (g_{11}, \ldots, g_{nn})'$ contiene los elementos de la diagonal de $\mathbf{G}$. Sea $\mathbf{H}$ la matriz de centrado (Cap. 1). Introducimos ahora las matrices $\mathbf{A} = -\frac{1}{2}\Delta^{(2)}$ y $\mathbf{B} = \mathbf{HAH}$.

Teorema 8.2.1 *La matriz de distancias Δ es euclídea si y sólo si $\mathbf{B} \geq 0$, es decir, los valores propios de $\mathbf{B}$ son no negativos.*

Demost.: La relación entre $\mathbf{B} = (b_{ij})$ y $\mathbf{A} = (a_{ij})$ es

$$b_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..},$$

donde $a_{i.}$ es la media de la columna i de $\mathbf{A}$, $a_{.j}$ es la media de la fila j y $a_{..}$ es la media de los n^2 elementos de $\mathbf{A}$. Entonces

$$b_{ii} = -a_{i.} - a_{.i} + a_{..}, \quad b_{jj} = -a_{j.} - a_{.j} + a_{..},$$

y por lo tanto

$$\delta_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij} = a_{ii} + a_{jj} - 2a_{ij}. \tag{8.3}$$

Supongamos que Δ es euclídea. Entonces $\mathbf{G} = \mathbf{X}\mathbf{X}'$. De (8.2) resulta que

$$\mathbf{A} = -(\mathbf{1}\mathbf{g}' + \mathbf{g}\mathbf{1}')/2 + \mathbf{G}.$$

Multiplicando ambos lados de $\mathbf{A}$ por $\mathbf{H}$, dado que $\mathbf{H1} = \mathbf{1}'\mathbf{H} = \mathbf{0}$, tenemos que

$$\mathbf{B} = \mathbf{HAH} = \mathbf{HGH} = \mathbf{HXX}'\mathbf{H} = \overline{\mathbf{X}\mathbf{X}}' \geq \mathbf{0},$$

lo que prueba que $\mathbf{B}$ es semidefinida positiva.

Supongamos ahora que $\mathbf{B} \geq 0$. Entonces $\mathbf{B} = \mathbf{Y}\mathbf{Y}'$ para alguna matriz $\mathbf{Y}$ de orden $n \times p$, es decir, $b_{ij} = \mathbf{y}_i'\mathbf{y}_j$, donde $\mathbf{y}_i'$ es la fila i- ésima de $\mathbf{Y}$. Aplicando (8.3) tenemos

$$\delta_{ij}^2 = \mathbf{y}_i'\mathbf{y}_i + \mathbf{y}_j'\mathbf{y}_j - 2\mathbf{y}_i'\mathbf{y}_j = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j),$$

que demuestra que Δ es matriz de distancias euclídeas.□

8.3 El análisis de coordenadas principales

Hemos visto que si $\mathbf{B} \geq 0$, cualquier matriz $\mathbf{Y}$ tal que $\mathbf{B} = \mathbf{Y}\mathbf{Y}'$ proporciona unas coordenadas cartesianas compatibles con la matriz de distancias Δ. Sea

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$$

la descomposición espectral de $\mathbf{B}$, donde $\mathbf{U}$ es una matriz $n \times p$ de vectores propios ortonormales de $\mathbf{B}$ y Λ es matriz diagonal que contiene los valores propios ordenados

$$\lambda_1 \geq \cdots \geq \lambda_p > \lambda_{p+1} = 0 \tag{8.4}$$

Obsérvese que $\mathbf{B1} = \mathbf{0}$, y por lo tanto $\lambda_{p+1} = 0$ es también valor propio de $\mathbf{B}$ de vector propio el vector $\mathbf{1}$ de unos. Entonces es evidente que la matriz $n \times p$

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2} \tag{8.5}$$

también verifica $\mathbf{B} = \mathbf{X}\mathbf{X}'$.

Definició 8.3.1 *La solución por coordenadas principales es la matriz de coordenadas (8.5), tal que sus columnas $X_1, \ldots, X_p$, que interpretaremos como variables, son vectores propios de $\mathbf{B}$ de valores propios (8.4). Las coordenadas del elemento $i \in \Omega$ son*

$$\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip}),$$

donde $\mathbf{x}_i$ es la fila i-ésima de $\mathbf{X}$. Reciben el nombre de coordenadas principales y cumplen (8.1).

La solución por coordenadas principales goza de importantes propiedades. En las aplicaciones prácticas, se toman las $q < p$ primeras coordenadas principales a fin de representar Ω. Por ejemplo, si $q = 2$, las dos primeras coordenadas de $\mathbf{X}$ proporcionan una representación a lo largo de los ejes X_1 y X_2 :

	X_1	X_2
1	x_{11}	x_{12}
2	x_{21}	x_{22}
$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}

Propiedades:

1. Las variables X_k (columnas de $\mathbf{X}$) tienen media 0.

   $$\overline{X}_1 = \cdots = \overline{X}_p = 0$$

   Prueba: $\mathbf{1}$ es vector propio de $\mathbf{B}$ ortogonal a cada X_k, por lo tanto $\overline{X}_k = \frac{1}{n}(\mathbf{1}'X_k) = 0$.

2. Las varianzas son proporcionales a los valores propios

   $$s_k^2 = \frac{1}{n}\lambda_k, \quad k = 1, \ldots, p$$

   Prueba: la varianza es $\frac{1}{n}X_k'X_k = \frac{1}{n}\lambda_k$.

3. Las variables son incorrelacionadas

$$\text{cor}(X_k, X_{k'}) = 0, \quad k \neq k' = 1, \ldots, p.$$

Prueba: como las medias son nulas, la covarianza es

$$\text{cov}(X_k, X_{k'}) = \frac{1}{n} X_k' X_{k'} = 0,$$

pues los vectores propios de $\mathbf{B}$ son ortogonales.

4. Las variables X_k son componentes principales de cualquier matriz de datos $\mathbf{Z}$ tal que las distancias euclídeas entre sus filas concuerden con Δ.

Prueba: Supongamos $\mathbf{Z}$ matriz de datos centrada. Tenemos que

$$\mathbf{B} = \mathbf{X}\mathbf{X}' = \mathbf{Z}\mathbf{Z}'$$

La matriz de covarianzas de $\mathbf{Z}$ es

$$\mathbf{S} = \frac{1}{n}\mathbf{Z}'\mathbf{Z} = \mathbf{T}\mathbf{D}\mathbf{T}',$$

donde $\mathbf{D}$ es diagonal y $\mathbf{T}$ es la matriz ortogonal de la transformación en componentes principales. Entonces:

$$\begin{aligned} \mathbf{Z}'\mathbf{Z} &= n\mathbf{T}\mathbf{D}\mathbf{T}', \\ \mathbf{Z}\mathbf{Z}'\mathbf{Z} &= n\mathbf{Z}\mathbf{T}\mathbf{D}\mathbf{T},' \\ \mathbf{B}\mathbf{Z}\mathbf{T} &= \mathbf{Z}\mathbf{T}n\mathbf{D}, \end{aligned}$$

y por lo tanto $\mathbf{Z}\mathbf{T}$ es matriz de vectores propios de $\mathbf{B}$ con valores propios los elementos diagonales de $n\mathbf{D}$, lo que implica $\mathbf{X} = \mathbf{Z}\mathbf{T}$. En consecuencia la matriz de coordenadas principales $\mathbf{X}$ coincide con la transformación por componentes principales de $\mathbf{Z}$.

5. La variabilidad geométrica de Δ es

$$V_\delta(\mathbf{X}) = \frac{1}{2n^2} \sum_{i,j=1}^{n} \delta_{ij}^2 = \frac{1}{n} \sum_{k=1}^{p} \lambda_k.$$

6. La variabilidad geométrica en dimensión q es máxima cuando tomamos las q primeras coordenadas principales. Es decir,

$$V_\delta(\mathbf{X})_q = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2(q) = \frac{1}{2n^2}\sum_{i,j=1}^{n}\sum_{k=1}^{q}(x_{ik}-x_{jk})^2 = \frac{1}{n}\sum_{k=1}^{q}\lambda_k$$

es máximo.

Prueba: Sea $x_1, ..., x_n$ una muestra con media $\overline{x}=0$ y varianza s^2. Se verifica

$$\begin{aligned}\tfrac{1}{2n^2}\textstyle\sum_{i,j=1}^{n}(x_i-x_j)^2 &= \tfrac{1}{2n^2}(\textstyle\sum_{i,j=1}^{n}x_i^2+\sum_{i,j=1}^{n}x_j^2-2\sum_{i,j=1}^{n}x_ix_j)\\ &= \tfrac{1}{2n^2}(n\textstyle\sum_{i=1}^{n}x_i^2+n\sum_{j=1}^{n}x_j^2-2\sum_{i=1}^{n}x_i\sum_{ij=1}^{n}x_j)\\ &= s^2,\end{aligned}$$

por lo tanto

$$V_\delta(\mathbf{X}) = \sum_{k=1}^{p}s_k^2.$$

Hemos demostrado que para cualquier matriz $\mathbf{X}$ tal que $\mathbf{B}=\mathbf{X}\mathbf{X}'$, la suma de las varianzas de las columnnas de $\mathbf{X}$ es igual a la variabilidad geométrica. Si en particular tenemos las coordenadas principales, esta suma de varianzas es la suma de los valores propios dividida por n, y como entonces las columnas son componentes principales, sus varianzas son respectivamente máximas.

El porcentaje de variabilidad explicada por los q primeros ejes principales es la proporción de variabilidad geométrica

$$P_q = 100\frac{V_\delta(\mathbf{X})_q}{V_\delta(\mathbf{X})} = 100\frac{\sum_{k=1}^{q}\lambda_k}{\sum_{k=1}^{p}\lambda_k}$$

Exemple 8.3.1

Consideremos $\Omega=\{1,2,3,4,5\}$ y la matriz de distancias (al cuadrado):

	1	2	3	4	5
1	0	226	104	34	101
2		0	26	104	29
3			0	26	9
4				0	41
5					0

Los valores propios de $\mathbf{B}$ son $\lambda_1 = 130$, $\lambda_2 = 10$, $\lambda_3 = \lambda_4 = \lambda_5 = 0$. Por lo tanto Δ es matriz de distancias euclídeas y Ω se puede representar en un espacio de dimensión 2. Las coordenadas principales son las columnas X_1, X_2 de:

	X_1	X_2	$\mathbf{1}$
1	-8	-1	1
2	7	0	1
3	2	1	1
4	-3	2	1
5	2	-2	1
λ	130	10	0
$\overline{x}$	0	0	1
s^2	26	2	0

8.4 Similaridades

En ciertas aplicaciones, especialmente en Biología y Psicología, en lugar de una distancia, lo que se mide es el grado de similaridad entre cada par de individuos.

Una similaridad s sobre un conjunto finito Ω es una aplicación de $\Omega \times \Omega$ en R tal que:

$$s(i,i) \geq s(i,j) = s(j,i) \geq 0.$$

La matriz de similaridades entre los elementos de Ω es

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

donde $s_{ij} = s(i,j)$.

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) consideremos la tabla

		j	
		1	0
i	1	a	b
	0	c	d

donde a, b, c, d las frecuencias de (1,1), (1,0), (0,1) y (0,0), respectivamente, con $p = a + b + c + d$. Un coeficiente de similaridad debería ser función de a, b, c, d. Son conocidos los coeficientes de similaridad:

$$\begin{aligned} s_{ij} &= \frac{a+d}{p} \qquad \text{(Sokal-Michener)} \\ s_{ij} &= \frac{a}{a+b+c} \qquad \text{(Jaccard)} \end{aligned} \tag{8.6}$$

que verifican : $s_{ii} = 1 \geq s_{ij} = s_{ji} \geq 0$.

Podemos transformar una similaridad en distancia aplicando la fórmula

$$d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}. \tag{8.7}$$

Entonces la matriz $\mathbf{A} = -(d_{ij}^2)/2$ es

$$\mathbf{A} = -\frac{1}{2}(\mathbf{S}_f + \mathbf{S}_f' - 2\mathbf{S}),$$

donde $\mathbf{S}_f$ tiene todas sus filas iguales, y como $\mathbf{HS}_f = \mathbf{S}_f'\mathbf{H} = \mathbf{0}$, resulta que

$$\mathbf{B} = \mathbf{HAH} = \mathbf{HSH}.$$

Por lo tanto:

1. Si $\mathbf{S}$ es matriz (semi)definida positiva, la distancia d_{ij} es euclídea.
2. $\text{rang}(\mathbf{HSH}) = \text{rang}(\mathbf{S}) - 1$.
3. Las coordenadas principales se obtienen diagonalizando $\mathbf{HSH}$.

8.5 Nociones de MDS no métrico

Supongamos que la matriz de distancias Δ es no euclídea. Entonces la matriz $\mathbf{B}$ (Teorema 8.2.1) tiene valores propios negativos:

$$\lambda_1 \geq \cdots \geq \lambda_p > 0 > \lambda_{p+1} \geq \cdots \geq \lambda_{p'}.$$

El fundamento del MDS no métrico es transformar las distancias δ_{ij} para convertirlas en euclídeas, pero conservando las relaciones de proximidad entre los elementos del conjunto Ω.

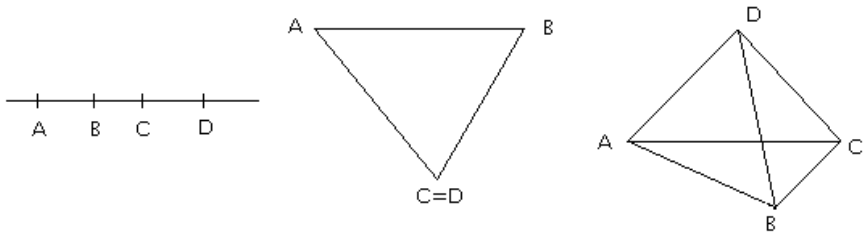

Figura 8.1: Representación de 4 objetos conservando las preordenaciones relacionadas a tres matrices de distancias.

Definició 8.5.1 *La preordenación asociada a la matriz de distancias* Δ *es la ordenación de las* $m = n(n-1)/2$ *distancias:*

$$\delta_{i_1 j_1} \leq \delta_{i_2 j_2} \leq \cdots \leq \delta_{i_m j_m}. \tag{8.8}$$

La preordenación es, de hecho, una propiedad asociada a Ω, es decir, podemos escribir

$$(i_1, j_1) \preceq (i_2, j_2) \preceq \cdots \preceq (i_m, j_m), \quad (i_k, j_k) \in \Omega \times \Omega,$$

donde

$$(i, j) \preceq (i', j') \quad \text{si} \quad \delta_{ij} \leq \delta_{i'j'}.$$

Se trata de representar Ω en un espacio que conserve la preordenación. Por ejemplo, si consideramos las tres matrices de distancias sobre {A,B,C,D}:

	A	B	C	D	A	B	C	D	A	B	C	D
A	0	1	2	3	0	1	1	1	0	1	1	1
B		0	1	2		0	1	1		0	1	1
C			0	1			0	0			0	1
D				0				0				0

las preordenaciones se pueden representar en 1, 2 ó 3 dimensiones (Fig. 8.1), respectivamente.

Si transformamos la distancia δ_{ij} en $\widehat{\delta}_{ij} = \varphi(\delta_{ij})$, donde φ es una función positiva creciente, es evidente que $\widehat{\delta}_{ij}$ tiene la misma preordenación (8.8), y por lo tanto, individuos próximos (alejados) según δ_{ij} estarán también próximos (alejados) con respecto a $\widehat{\delta}_{ij}$. Si además $\widehat{\delta}_{ij}$ es euclídea, tendremos la posibilidad de representar Ω, aplicando, por ejemplo, un análisis de coordenadas principales sobre la distancia transformada, pero conservando (aproximadamente) la preordenación. En general, la función φ no es lineal, y se obtiene por regresión monótona. Hay dos casos especialmente simples.

Definició 8.5.2 *La transformación q-aditiva de δ_{ij} se define como*

$$\widehat{\delta}_{ij}^2 = \begin{cases} \delta_{ij}^2 - 2a & si\ i \neq j \\ 0 & si\ i = j \end{cases}$$

donde $a < 0$ es una constante. La transformación aditiva se define como

$$\widehat{\delta}_{ij} = \begin{cases} \delta_{ij} + c & si\ i \neq j \\ 0 & si\ i = j \end{cases}$$

donde $c > 0$ es una constante.

Es evidente que las dos transformaciones aditiva y q-aditiva conservan la preordenación de la distancia. Probemos ahora que la primera puede dar lugar a una distancia euclídea.

Teorema 8.5.1 *Sea Δ una matriz de distancias no euclídeas y sea $\lambda_{p'} < 0$ el menor valor propio de $\mathbf{B}$. Entonces la transformación q-aditiva proporciona una distancia euclídea para todo a tal que $a \leq \lambda_{p'}$.*

Demost.: Sea $\widehat{\Delta} = (\widehat{\delta}_{ij})$ la matriz de distancias transformadas. Las matrices $\mathbf{A}, \mathbf{B}$ y $\widehat{\mathbf{A}}$,$\widehat{\mathbf{B}}$ (ver Teorema 8.2.1) verifican

$$\widehat{\mathbf{A}} = \mathbf{A} - a(\mathbf{I} - \mathbf{J}), \quad \widehat{\mathbf{B}} = \mathbf{B} - a\mathbf{H}.$$

Sea $\mathbf{v}$ vector propio de $\mathbf{B}$ de valor propio $\lambda \neq 0$. Entonces $\mathbf{Hv} = \mathbf{v}$ y por lo tanto

$$\widehat{\mathbf{B}}\mathbf{v} = (\mathbf{B} - a\mathbf{H})\mathbf{v} = (\lambda - a)\mathbf{v}.$$

Así $\widehat{\mathbf{B}}$ tiene los mismos vectores propios que $\mathbf{B}$, pero los valores propios son

$$\lambda_1 - a \geq \cdots \geq \lambda_p - a > 0 > \lambda_{p+1} - a \geq \cdots \geq \lambda_{p'} - a,$$

que son no negativos si $a \leq \lambda_{p'}$, en cuyo caso $\widehat{\mathbf{B}}$ es semidefinida positiva.□

La mejor transformación q-aditiva es la que menos distorsiona la distancia original. De acuerdo con este criterio, el mejor valor para la constante es $a = \lambda_{p'}$.

Las transformaciones aditiva y no lineal son más complicadas y las dejamos para otro dia. De hecho, los programas de MDS operan con transformaciones no lineales, siguiendo criterios de minimización de una función que mide la discrepancia entre la distancia original y la transformada. Por ejemplo, el método de Kruskal consiste en:

1. Fijar una dimensión Euclídea p.
2. Transformar la distancia δ_{ij} en la "disparidad" $\widehat{\delta}_{ij} = \varphi(\delta_{ij})$, donde φ es una función monótona creciente. Las disparidades conservan la preordenación de las distancias.
3. Ajustar una distancia euclídea d_{ij} a las disparidades $\widehat{\delta}_{ij}$ de manera que minimice
$$\sum_{i<j}(d_{ij} - \widehat{\delta}_{ij})^2.$$
4. Asociar a las distancias d_{ij} una configuración euclídea p-dimensional, y representar los n objetos a partir de las coordenades de la configuración.

Para saber si la representación obtenida refleja bien las distancias entre los objetos, se calcula la cantidad

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij} - \widehat{\delta}_{ij})^2}{\sum_{i<j} d_{ij}^2}},$$

denominada "stress", que verifica $0 \leq S \leq 1$, pero se expresa en forma de porcentaje. La representación es considerada buena si S no supera el 5%.

También es conveniente obtener el diagrama de Sheppard, que consiste en representar los $n(n-1)/2$ puntos (δ_{ij}, d_{ij}). Si los puntos dibujan una curva creciente, la representación es buena, porque entonces se puede decir que conserva bien la preordenación (Fig. 8.4).

8.6 Distancias estadísticas

En esta sección discutiremos algunos modelos de distancias estadísticas.

8.6.1 Variables cuantitativas

Siendo $\mathbf{x} = (x_1, x_2, \ldots, x_p), \mathbf{y} = (y_1, y_2, \ldots, y_p)$ dos puntos de R^p. La distancia de Minkowsky se define como

$$d_q(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^{p} |x_i - y_i|^q)^{1/q},$$

Casos particulares de la distancia d_q son:

1. Distancia "ciudad":

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

2. Distancia Euclídea:

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

3. Distancia "dominante":

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq p} \{|x_i - y_i|\}$$

Tienen también interés en las aplicaciones, la distancia normalizada por el rang R_i de la variable i

$$d_G(\mathbf{x}, \mathbf{y}) = \frac{1}{p} \sum_{i=1}^{p} \frac{|x_i - y_i|}{R_i},$$

y, cuando los valores de las variables son positivos, la métrica de Canberra

$$d_C(\mathbf{x}, \mathbf{y}) = \frac{1}{p} \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}.$$

d_G y d_C son invariantes por cambios de escala.

Supongamos ahora dos poblaciones Ω_1, Ω_2 con vectores de medias μ_1, μ_2 y matrices de covarianzas Σ_1, Σ_2. Cuando $\Sigma_1 = \Sigma_2 = \Sigma$, la distancia de Mahalanobis entre poblaciones es

$$M^2(\Omega_1, \Omega_2) = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$

Esta distancia, ya introducida previamente, es invariante por cambios de escala y tiene en cuenta la correlación entre las variables. Además, si M_p, M_q, M_{p+q} indican las distancias basada en $p, q, p+q$ variables, respectivamente, se verifica:

a) $M_p \leq M_{p+q}$.

b) $M^2_{p+q} = M^2_p + M^2_q$ si los dos grupos de p y q variables son independientes.

No es fácil dar una definición de distancia cuando $\Sigma_1 \neq \Sigma_2$. Una definición de compromiso es

$$(\mu_1 - \mu_2)'[\frac{1}{2}(\Sigma_1 + \Sigma_2)]^{-1}(\mu_1 - \mu_2).$$

8.6.2 Variables binarias

Cuando todas las variables son binarias (toman solamente los valores 0 y 1), entonces conviene definir un coeficiente de similaridad (Sección 8.4) y aplicar (8.7) para obtener una distancia. Existen muchas maneras de definir una similaridad s_{ij} en función del peso que se quiera dar a los a, b, c, d. Por ejemplo:

$$\begin{aligned} s_{ij} &= \frac{a}{a + 2(b + c)} \quad \text{(Sokal-Sneath)} \\ s_{ij} &= \frac{2a}{(a + b)(a + c)} \quad \text{(Dice)} \end{aligned} \tag{8.9}$$

Las similaridades definidas en (8.6) y (8.9) proporcionan distancias euclídeas.

8.6.3 Variables categóricas

Supongamos que las observaciones pueden ser clasificades en k categorías excluyentes $A_1, \ldots, A_k$, con probabilidades $\mathbf{p} = (p_1, \ldots, p_k)$, donde $\sum_{h=1}^{k} p_h = 1$. Podemos definir distancias entre individuos y entre poblaciones.

1. Entre individuos. Si dos individuos i, j tienen las categorías $A_h, A_{h'}$, respectivamente, una distancia (al cuadrado) entre i, j es:

$$d(i,j)^2 = \begin{cases} 0 & \text{si } h = h', \\ p_h^{-1} + p_{h'}^{-1} & \text{si } h \neq h'. \end{cases}$$

   Si hay varios conjuntos de variables categóricas, con un total de K categorías o estados, una similaridad es α/K ("matching coefficient"), donde α es el número de coincidencias.

2. Entre poblaciones. Si tenemos dos poblaciones representadas por $\mathbf{p} = (p_1, \ldots, p_k)$, $\mathbf{q} = (q_1, \ldots, q_k)$, dos distancias entre poblaciones son

$$\begin{aligned} d_a(\mathbf{p}, \mathbf{q}) &= 2 \textstyle\sum_{i=1}^k |p_i - q_i|/(p_i + q_i), \\ d_b(\mathbf{p}, \mathbf{q}) &= \arccos(\textstyle\sum_{i=1}^k \sqrt{p_i q_i}). \end{aligned}$$

8.6.4 Variables mixtas

En las aplicaciones a menudo los datos provienen de las observaciones de p_1 variables cuantitativas, p_2 variables dicotómicas (dos estados: presente, ausente) y p_3 variables categóricas o cualitativas (más de dos estados). Un coeficiente de similaridad (propuesto por J.C. Gower) es

$$s_{ij} = \frac{\sum_{h=1}^{p_1}(1 - |x_{ih} - x_{jh}|/R_h) + a + \alpha}{p_1 + (p_2 - d) + p_3},$$

donde R_h es el rango de la variable cuantitativa X_h, a y d son el número de dobles presencias y dobles ausencias de las variables dicotómicas, y α es el número de coincidencias entre las variables categóricas. Si solamente hay variables dicotómicas o variables categóricas, s_{ij} reduce la similaridad normalizada por el rango, al coeficiente de Jaccard o al "matching coefficient", respectivamente:

$$\begin{cases} 1 - \frac{1}{p_1} \sum_{h=1}^{p_1} |x_h - y_h|/R_h & \text{si } p_2 = p_3 = 0, \\ a/(a+b+c) & \text{si } p_1 = p_3 = 0, \\ \alpha/p_3 & \text{si } p_1 = p_2 = 0. \end{cases}$$

Este coeficiente verifica $0 \leq s_{ij} \leq 1$, y aplicando (8.7) se obtiene una distancia euclídea que además admite la posibilidad de datos faltantes.

8.6.5 Otras distancias

Existen muchos procedimientos para definir distancias, en función de los datos y el problema experimental. Veamos dos.

Modelo de Thurstone

Supongamos que queremos ordenar n estímulos $\omega_1, \ldots, \omega_n$ (por ejemplo, n productos comerciales)

$$\omega_{i_1} \preceq \ldots \preceq \omega_{i_n}$$

según una escala de preferencias $\theta_{i_1} \leq \cdots \leq \theta_{i_n}$, donde los θ_i son parámetros. Sea p_{ij} la proporción de individuos de la población que prefieren ω_j sobre ω_i. Un modelo es

$$p_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\theta_j - \theta_i} e^{-t^2/2} dt.$$

Si más de la mitad de los individuos prefieren ω_j sobre ω_i, entonces $\theta_i < \theta_j$. Así:

a) $p_{ij} < 0.5$ implica $\theta_i > \theta_j$,

b) $p_{ij} = 0.5$ implica $\theta_i = \theta_j$,

c) $p_{ij} > 0.5$ implica $\theta_i < \theta_j$.

La estimación de los parámetros a partir de las proporciones p_{ij} es complicada. Alternativamente, teniendo en cuenta que $p_{ij} + p_{ji} = 1$ podemos definir la distancia entre estímulos

$$d(\omega_i, \omega_j) = |p_{ij} - 0.5|$$

y aplicar un MDS sobre la matriz $(d(\omega_i, \omega_j))$. La representación de los estímulos a lo largo de la primera dimensión nos proporciona una solución a la ordenación de los estímulos.

Distancia de Rao

Sea $S_\theta = \{f(x, \theta), \theta \in \Theta\}$ un modelo estadístico y $z(\theta) = \frac{\partial}{\partial \theta} \log f(x, \theta)$ un vector columna. La matriz de información de Fisher $F(\theta)$ es la matriz de

covarianzas de los z's. Siendo θ_a, θ_b dos valores de los parámetros. Una distancia tipo Mahalanobis sería el valor esperado de

$$(z(\theta_a) - z(\theta_b))' F(\theta)^{-1} (z(\theta_a) - z(\theta_b)).$$

Pero z depende de x y θ varía entre θ_a, θ_b. Consideremos entonces a $F(\theta)$ como un tensor métrico sobre la variedad diferenciable S_θ. La distancia de Rao entre θ_a, θ_b es la distancia geodésica entre los puntos correspondientes de S_θ. La distancia de Rao es invariante por transformaciones de las variables y de los parámetros, generaliza la distancia de Mahalanobis y tiene aplicaciones en estadística matemática. Veamos tres ejemplos.

1. Distribución de Poisson: $f(x, \lambda) = e^{-x}\lambda^x/x!,\ x = 0, 1, 2, \ldots$. La distancia entre dos valores λ_a, λ_b es:

$$\Delta(\lambda_a, \lambda_b) = 2|\sqrt{\lambda_a} - \sqrt{\lambda_b}|.$$

2. Distribución multinomial. La distancia entre $\mathbf{p} = (p_1, \ldots, p_k)$ y $\mathbf{q} = (q_1, \ldots, q_k)$ es:

$$\Delta(\mathbf{p}, \mathbf{q}) = \arccos(\sum_{i=1}^{k} \sqrt{p_i q_i}).$$

3. Distribución normal. Si Σ es fija, la distancia (al cuadrado) entre dos vectores de medias es:

$$\Delta^2(\Omega_1, \Omega_2) = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2).$$

Finalmente, para un valor fijo de θ, podemos definir la distancia entre dos observaciones x_1, x_2 que dan $z_i(\theta) = \frac{\partial}{\partial \theta} \log f(x_i, \theta), i = 1, 2$, como

$$(z_1(\theta) - z_2(\theta))' F(\theta)^{-1} (z_1(\theta) - z_2(\theta)).$$

8.7 Dos ejemplos

Exemple 8.7.1

Un arqueólogo encontró 5 herramientas cortantes A,B,C,D,E y una vez examinadas, comprobó que estaban hechas de piedra, bronce y hierro, conforme a la siguiente 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

Utilizando la similaridad de Jaccard (8.6), obtenemos la matriz de similaridades:

	A	B	C	D	E
A	1	1/2	1/2	0	0
B		1	1/3	0	1/2
C			1	1/2	0
D				1	0
E					1

Los resultados del análisis de coordenadas principales son:

A	.0000	.6841	-.3446
B	.4822	.1787	.2968
C	-.4822	.1787	.2968
D	-.6691	-.5207	-.1245
E	.6691	-.5207	-.1245
valor propio	1.360	1.074	.3258
porc. acum.	44.36	79.39	90.01

La primera y segunda coordenadas explican el 80% de la variabilidad geométrica. La representación (Fig. 8.2) indica que las herramientas quedan ordenadas según su antigüedad: E es la más antigua (sólo contiene piedra) y D la más moderna (sólo contiene hierro).

Exemple 8.7.2

Una distancia genética es una medida que cuantifica las proximidades entre dos poblaciones a partir de las proporciones génicas. Por ejemplo, si existen k ordenaciones cromosómicas que se presentan en las proporciones

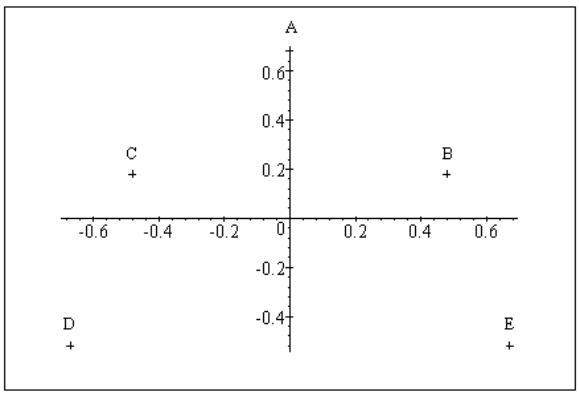

Figura 8.2: Representación por análisis de coordenadas principales de 5 herramientas prehistóricas.

$(p_1, \ldots, p_k), (q_1, \ldots, q_k)$, una distancia adecuada (propuesta por A. Prevosti) es

$$\frac{1}{2r} \sum_{i=1}^{k} |p_i - q_i|$$

donde r es el número de cromosomas diferentes.

Las distancias entre $n = 19$ poblaciones de *D. Suboscura* que provienen de Droback, Dalkeith, Groningen, Fontaineblau, Viena, Zurich, Huelva, Barcelona, Fornia, Foresta, Etna, Fruska-Gora, Thessaloniki, Silifke, Trabzon, Chalus, Orangerie, Agadir, Las Mercedes, se dan en la Tabla 8.1. Aplicando un MDS no métrico, se obtiene la representación de las 19 poblaciones (Fig. 8.3), con un "stress" de 2.84, que indica que la representación es buena. La Fig. 8.4 representa las distancias versus las disparidades.

8.8 Complementos

En un plano teórico, el MDS comienza con el teorema de I. J. Schoenberg acerca de la posibilidad de construir las coordenadas de un conjunto de puntos dadas sus distancias. A nivel aplicado, es de destacar a W. S. Torgerson, que en 1957 aplica el MDS a la psicología, y Gower (1966), que prueba su relación con el Análisis de Componentes Principales y el Canónico de Poblaciones, abriendo un fructífero campo de aplicación en la biología.

	Dro	Dal	Gro	Fon	Vie	Zur	Hue	Bar	For	For	Etn	Fru	The	Sil	Tra	Cha	Ora	Aga	Las
DROBA	0																		
DALKE	.307	0																	
GRONI	.152	.276	0																
FONTA	.271	.225	.150	0															
VIENA	.260	.370	.187	.195	0														
ZURIC	.235	.300	.112	.120	.128	0													
HUELV	.782	.657	.695	.580	.540	.623	0												
BARCE	.615	.465	.529	.412	.469	.445	.259	0											
FORNI	.780	.657	.693	.607	.606	.609	.373	.309	0										
FORES	.879	.790	.801	.764	.760	.761	.396	.490	.452	0									
ETNA	.941	.846	.873	.813	.818	.817	.414	.524	.451	.177	0								
FRUSK	.560	.505	.470	.442	.342	.391	.577	.460	.501	.681	.696	0							
THESS	.668	.545	.592	.514	.434	.500	.502	.392	.363	.590	.630	.315	0						
SILIF	.763	.643	.680	.584	.581	.610	.414	.357	.413	.646	.667	.544	.340	0					
TRABZ	.751	.619	.675	.582	.519	.587	.418	.342	.399	.587	.648	.439	.269	.286	0				
CHALU	.709	.489	.636	.548	.531	.549	.595	.489	.514	.635	.649	.444	.408	.574	.438	0			
ORANG	.947	.867	.864	.782	.837	.795	.573	.574	.568	.519	.535	.782	.733	.696	.698	.760	0		
AGADI	.927	.834	.844	.803	.789	.792	.428	.498	.485	.329	.303	.666	.661	.642	.631	.710	.321	0	
LASME	.931	.699	.846	.749	.802	.792	.404	.485	.429	.380	.253	.659	.566	.604	.551	.460	.615	.430	0

Tabla 8.1: Distancias genéticas respecto a las ordenaciones cromosómicas entre 19 poblaciones de D. Suboscura.

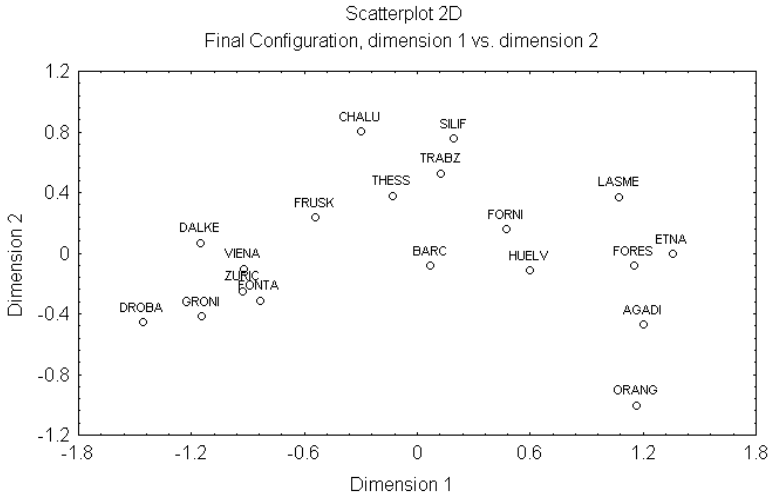

Figura 8.3: Representación MDS de 19 poblaciones de *D. Subobscura* respecto a las distancias genéticas entre ordenaciones cromosómicas.

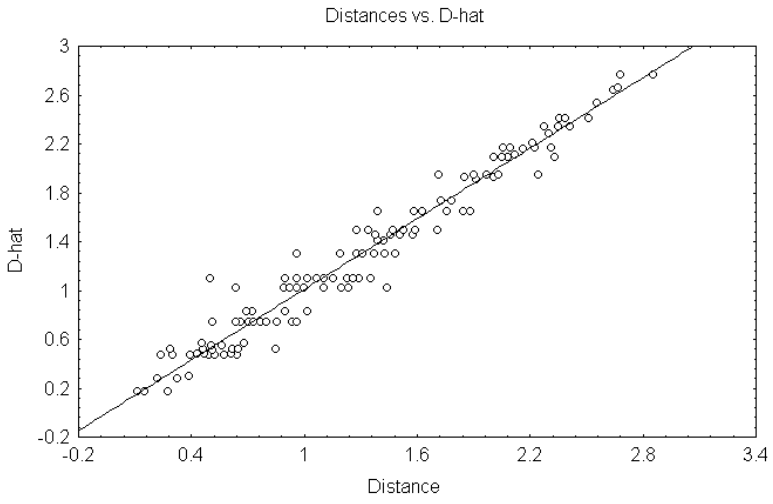

Figura 8.4: Representación de las distancias genéticas vs las disparidades.

El MDS no métrico es debido a R. N. Shepard, que en 1962 introdujo el concepto de preordenación, y J. B. Kruskal, que en 1964 propuso algoritmos efectivos que permitían encontrar soluciones. La transformación q-aditiva fue estudiada por J.C. Lingoes y K.V. Mardia. Diversos autores estudiaron la transformación aditiva, hasta que Cailliez (1983) encontró la solución definitiva. Consultar Cox y Cox (1994).

Existen diferentes modelos para tratar el problema de la representación cuando actúan diferentes matrices de distancias. Un modelo, propuesto por J.D. Carroll, es el INDSCAL. Un modelo reciente, propuesto por Cuadras y Fortiana (1998) y Cuadras (1998), es el "related metric scaling".

De la misma manera que se hace regresión sobre componentes principales, se puede hacer regresión de una variable dependiente Y sobre las dimensiones principales obtenidas aplicando MDS sobre una matriz de distancias entre las observaciones. Este modelo de regresión basado en distancias permite plantear la regresión con variables mixtas. Consultar Cuadras y Arenas (1990), Cuadras *et al.* (1996).

Una versión del MDS, denominada "continuous scaling", permite encontrar las coordenadas principales de una variable aleatoria. Consultar Cuadras y Fortiana (1993a,1995), Cuadras y Lahlou (2000).

P.C. Mahalanobis y C. R. Rao propusieron sus distancias en 1936 y 1945, respectivamente. Posteriormente Amari, Atkinson, Burbea, Mitchell, Oller y otros estudiaron la distancia de Rao. Consultar Oller (1987), Oller y Cuadras (1985), Cuadras (1988).

Capítulo 9

ANALISIS DE CORRESPONDENCIAS

9.1 Introducción

El Análisis de Correspondencias (AC) es una técnica multivariante que permite representar las categorías de las filas y columnas de una tabla de contingencia.

Supongamos que tenemos dos variables categóricas A y B con I y J categorías respectivamente, y que han sido observadas cruzando las I categorías A con las J categorías B, obteniendo $n = \sum_{ij} f_{ij}$ observaciones, donde f_{ij} es el número de veces en que aparece la intersección $A_i \cap B_j$, dando lugar a la tabla de contingencia $I \times J$:

$$\begin{array}{c|cccc|c} & B_1 & B_2 & \cdots & B_J & \\ A_1 & f_{11} & f_{12} & \cdots & f_{1J} & f_{1\cdot} \\ A_2 & f_{21} & f_{22} & \cdots & f_{2J} & f_{2\cdot} \\ \vdots & & & \ddots & & \vdots \\ A_I & f_{I1} & f_{I2} & \cdots & f_{IJ} & f_{I\cdot} \\ \hline & f_{\cdot 1} & f_{\cdot 2} & \cdots & f_{\cdot J} & n \end{array} \tag{9.1}$$

donde $f_{i\cdot} = \sum_j f_{ij}$ son las frecuencias de A_i, $f_{\cdot j} = \sum_i f_{ij}$ son las frecuencias de B_j. Hemos de tener en cuenta que la tabla (9.1) resume la matriz de datos

inicial, que típicamente es de la forma

	A_1	A_2	$\cdots$	A_I	B_1	B_2	$\cdots$	B_J
1	1	0	$\cdots$	0	1	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	0	0	$\cdots$	1	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	0	0	$\cdots$	1	0	0	$\cdots$	1

en la que damos el valor 1 cuando se presenta una característica y 0 cuando no se presenta. Así, el individuo "1" presentaría las características A_1 y B_1, el individuo "i" presentaria las características A_I y B_2, y el individuo "n" las características A_I y B_J. La matriz de datos $n \times (I + J)$ es pues

$$\mathbf{Z} = [\mathbf{X}, \mathbf{Y}].$$

A partir de ahora utilizaremos el nombre de variables filas y variables columnas a las variables A y B, respectivamente.

Indiquemos por $\mathbf{N} = (f_{ij})$ la matriz $I \times J$ con las frecuencias de la tabla de contingencia. La matriz

$$\mathbf{P} = \frac{1}{n}\mathbf{N},$$

es la matriz de correspondencias. Indiquemos por $\mathbf{r}$ el vector $I \times 1$ con los totales marginales de las filas de $\mathbf{P}$, y por $\mathbf{c}$ el vector $J \times 1$ con los totales marginales de las columnas de P :

$$\mathbf{r} = \mathbf{P1}, \quad \mathbf{c} = \mathbf{P'1}.$$

Tenemos entonces que

$$\mathbf{r} = \frac{1}{n}\mathbf{1'X}, \quad \mathbf{c} = \frac{1}{n}\mathbf{1'Y},$$

son los vectores de medias de las matrices de datos $\mathbf{X}, \mathbf{Y}$. Indiquemos además

$$\mathbf{D}_r = \text{diag}(\mathbf{r}), \quad \mathbf{D}_c = \text{diag}(\mathbf{c}),$$

las matrices diagonales que contienen los valores marginales de filas y columnas de $\mathbf{P}$. Se verifica

$$\mathbf{X'X} = n\mathbf{D}_r, \quad \mathbf{Y'Y} = n\mathbf{D}_c, \quad \mathbf{X'Y} = n\mathbf{P} = \mathbf{N}.$$

Por lo tanto, las matrice de covarianzas entre filas, entre columnas y entre filas y columnas, son

$$\mathbf{S}_{11} = \mathbf{D}_r - \mathbf{r}\mathbf{r}', \quad \mathbf{S}_{22} = \mathbf{D}_c - \mathbf{c}\mathbf{c}', \quad \mathbf{S}_{12} = \mathbf{P} - \mathbf{r}\mathbf{c}'.$$

Puesto que la suma de las variables es igual a 1, las matrices $\mathbf{S}_{11}$ y $\mathbf{S}_{22}$ son singulares.

9.2 Cuantificación de las variables categóricas

El problema de las variables categóricas, para que puedan ser manejadas en términos de AM clásico, es que no son cuantitativas. La cuantificación 0 ó 1 anterior es convencional. Asignemos pues a las categorías $\mathrm{A}_1, \ldots, \mathrm{A}_I$ de la variable fila, los valores numéricos $a_1, \ldots, a_I$, y a las categorías $\mathrm{B}_1, \ldots, \mathrm{B}_J$ de la variable columna, los valores numéricos $b_1, \ldots, b_J$. es decir, indiquemos los vectores

$$\mathbf{a} = (a_1, \ldots, a_I)', \quad \mathbf{b} = (b_1, \ldots, b_J)',$$

y consideremos las variables compuestas

$$U = \mathbf{X}\mathbf{a}, \quad V = \mathbf{Y}\mathbf{b}.$$

Si en un individuo k se observan las categorías A_i,B_j, entonces los valores de U, V sobre k son

$$U_k = a_i, \quad V_k = b_j.$$

Deseamos encontrar $\mathbf{a}, \mathbf{b}$ tales que las correlaciones entre U y V sean máximas. Claramente, estamos ante un problema de correlación canónica, salvo que ahora las matrices $\mathbf{S}_{11}$ y $\mathbf{S}_{22}$ son singulares. Una g-inversa de $\mathbf{S}_{11}$ es la matriz $\mathbf{S}_{11}^- = \mathbf{D}_r^{-1}$ que verifica

$$\mathbf{S}_{11}\mathbf{S}_{11}^-\mathbf{S}_{11} = \mathbf{S}_{11}.$$

En efecto,

$$\begin{aligned}(\mathbf{D}_r - \mathbf{r}\mathbf{r}')\mathbf{D}_r^{-1}(\mathbf{D}_r - \mathbf{r}\mathbf{r}') &= (\mathbf{D}_r - \mathbf{r}\mathbf{r}')(\mathbf{I} - \mathbf{1}\mathbf{r}') \\ &= \mathbf{D}_r - \mathbf{D}_r\mathbf{1}\mathbf{r}' - \mathbf{r}\mathbf{r}' + \mathbf{r}\mathbf{r}'\mathbf{1}\mathbf{r}' \\ &= \mathbf{D}_r - \mathbf{r}\mathbf{r}' - \mathbf{r}\mathbf{r}' + \mathbf{r}\mathbf{r}' \\ &= \mathbf{D}_r - \mathbf{r}\mathbf{r}'.\end{aligned}$$

Análogamente $\mathbf{S}_{22}^{-} = \mathbf{D}_c^{-1}$. Aplicando la teoria de la correlación canónica (Sección 4.3), podemos considerar la descomposición singular

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}', \tag{9.2}$$

donde $\mathbf{D}_\lambda$ es la matriz diagonal con los valores singulares en orden decreciente. Si $\mathbf{u}_1, \mathbf{v}_1$ son los primeros vectores canónicos, tendremos entonces

$$\mathbf{a} = \mathbf{S}_{11}^{-1/2}\mathbf{u}_1, \quad \mathbf{b} = \mathbf{S}_{22}^{-1/2}\mathbf{v}_1, \quad r = \lambda_1,$$

es decir, el primer valor singular es la máxima correlación entre las variables U y V. Pero pueden haber más vectores y correlaciones canonicas, y por lo tanto la solución general es

$$\mathbf{a}_i = \mathbf{D}_r^{-1/2}\mathbf{u}_i, \quad \mathbf{b}_i = \mathbf{D}_c^{-1/2}\mathbf{v}_i, \quad r_i = \lambda_i, \quad i = 1, \ldots, \min\{I, J\}.$$

En notación matricial, los vectores que cuantificn las categorías de las filas y de las columnas de $\mathbf{N}$, son las columnas de las matrices

$$\mathbf{A}_0 = \mathbf{D}_r^{-1/2}\mathbf{U}, \quad \mathbf{B}_0 = \mathbf{D}_c^{-1/2}\mathbf{V}.$$

También obtenemos correlaciones máximas considerando las matrices

$$\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\lambda, \quad \mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\lambda, \tag{9.3}$$

pues el producto por una constante (en este caso un valor singular), no altera las correlaciones.

9.3 Representación de filas y columnas

Los perfiles de las filas son

$$(\frac{p_{i1}}{r_i}, \frac{p_{i2}}{r_i}, \cdots, \frac{p_{iJ}}{r_i}),$$

es decir, las "probabilidades condicionadas" $P(\mathrm{B}_1/\mathrm{A}_i), \ldots, P(\mathrm{B}_J/\mathrm{A}_i)$. La matriz de perfiles de las filas es

$$\mathbf{Q} = \mathbf{D}_r^{-1}\mathbf{P}.$$

Definició 9.3.1 *La distancia ji-cuadrado entre las filas i, i' de* $\mathbf{N}$ *es*

$$\delta^2_{ii'} = \sum_{j=1}^{J} \frac{(p_{ij}/r_i - p_{i'j}/r_{i'})^2}{c_j}.$$

La matriz de productos escalares asociada a esta distancia es

$$\mathbf{G} = \mathbf{Q}\mathbf{D}_c^{-1}\mathbf{Q}',$$

y la relación entre $\Delta^{(2)} = (\delta^2_{ii'})$ y $\mathbf{G}$ es

$$\Delta^{(2)} = \mathbf{g}\mathbf{1}' + \mathbf{1}\mathbf{g}' - 2\mathbf{G},$$

siendo $\mathbf{g}$ el vector columna con los I elementos diagonales de $\mathbf{G}$. La solución MDS ponderada de las filas de $\mathbf{N}$ (Sección 9.8) se obtiene calculando la diagonalización

$$\mathbf{D}_r^{1/2}(\mathbf{I} - \mathbf{1}\mathbf{r}')\mathbf{G}(\mathbf{I} - \mathbf{r}\mathbf{1}')\mathbf{D}_r^{1/2} = \mathbf{U}\mathbf{D}_\lambda^2\mathbf{U}'$$

y seguidamente obteniendo las coordenadas principales

$$\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\lambda. \tag{9.4}$$

Las distancias euclídeas entre las filas de $\mathbf{A}$ coinciden con la distancia ji-cuadrado.

Relacionemos ahora estas coordenadas con las cuantificaciones anteriores. De (9.2) tenemos

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1}(\mathbf{P}' - \mathbf{c}\mathbf{r}')\mathbf{D}_r^{-1/2} = \mathbf{U}\mathbf{D}_\lambda^2\mathbf{U}',$$

y de

$$\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{D}_c^{-1}(\mathbf{P}'\mathbf{D}_r^{-1} - \mathbf{c}\mathbf{1}')\mathbf{D}_r^{1/2} = \mathbf{D}_r^{1/2}(\mathbf{Q} - \mathbf{1}\mathbf{r}'\mathbf{Q})\mathbf{D}_c^{-1}(\mathbf{Q}' - \mathbf{Q}'\mathbf{r}\mathbf{1}')\mathbf{D}_r^{1/2},$$

deducimos que

$$\mathbf{D}_r^{1/2}(\mathbf{I} - \mathbf{1}\mathbf{r}')\mathbf{Q}\mathbf{D}_c^{-1}\mathbf{Q}'(\mathbf{I} - \mathbf{r}\mathbf{1}')\mathbf{D}_r^{1/2} = \mathbf{U}\mathbf{D}_\lambda^2\mathbf{U}'.$$

Esta última expresión demuestra que las matrices $\mathbf{A}$ obtenidas en (9.3) y (9.4) son la misma.

Análogamente podemos definir la distancia ji-cuadrado entre columnas

$$\delta^2_{jj'} = \sum_{i=1}^{I} \frac{(p_{ij}/c_j - p_{ij'}/c_{j'})^2}{r_i},$$

y probar que las distancias euclídeas entre las filas de la matriz $\mathbf{B}$ obtenidas en (9.3), coinciden con esta distancia ji-cuadrado.

Así pues, si consideramos las dos primeras coordenadas principales:

	Filas		Columnas
A_1	(a_{11}, a_{12})	B_1	(b_{11}, b_{12})
A_2	(a_{21}, a_{22})	B_2	(b_{21}, b_{22})
$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	(a_{I1}, a_{I2})	B_J	(b_{J1}, b_{J2})

obtenemos una representación de filas y columnas de la matriz de frecuencias $\mathbf{N}$.

9.4 Relación entre filas y columnas y representación conjunta

Las coordenadas $\mathbf{A}$ y las coordenadas $\mathbf{B}$, que representan las filas y las columnas, están relacionadas. Premultiplicando (9.2) por $\mathbf{D}_r^{-1/2}$ y postmultiplicando por $\mathbf{V}$ obtenemos

$$\mathbf{D}_r^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}\mathbf{V} = \mathbf{D}_r^{-1/2}\mathbf{U},$$

luego

$$\mathbf{D}_r^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{B}\mathbf{D}_\lambda^{-1} = \mathbf{A}.$$

Análogamente se prueba que

$$\mathbf{D}_c^{-1}(\mathbf{P}' - \mathbf{c}\mathbf{r}')\mathbf{A}\mathbf{D}_\lambda^{-1} = \mathbf{B}.$$

Si ahora tenemos en cuenta que $\mathbf{r}'\mathbf{D}_r^{-1} = \mathbf{1}'$, premultiplicando por $\mathbf{r}'$

$$\mathbf{1}'(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{B}\mathbf{D}_\lambda^{-1} = \mathbf{r}'\mathbf{A}.$$

Como además $\mathbf{1}'\mathbf{P} = \mathbf{c}', \mathbf{1}'\mathbf{r} = 1$, vemos fácilmente que

$$(\mathbf{c}' - \mathbf{c}')\mathbf{B}\mathbf{D}_\lambda^{-1} = \mathbf{r}'\mathbf{A} = \mathbf{0}.$$

Análogamente, $\mathbf{c}'\mathbf{B} = \mathbf{0}$, es decir, las medias ponderadas de las coordenadas principales son cero. En consecuencia

$$\mathbf{A} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{B}\mathbf{D}_\lambda^{-1}, \quad \mathbf{B} = \mathbf{D}_c^{-1}\mathbf{P}'\mathbf{A}\mathbf{D}_\lambda^{-1}. \tag{9.5}$$

Conviene notar que $\mathbf{D}_r^{-1}\mathbf{P}$ son los perfiles de las filas, y $\mathbf{D}_c^{-1}\mathbf{P}'$ son los perfiles de las columnas. Así pues tenemos que, salvo el factor dilatador $\mathbf{D}_\lambda^{-1}$, (pues los elementos diagonales de $\mathbf{D}_\lambda$ son menores que 1), se verifica:

1. Las coordenadas de las filas son medias, ponderadas por los perfiles de las filas, de las coordenadas de las columnas.

2. Las coordenadas de las columnas son medias, ponderadas por los perfiles de las columnas, de las coordenadas de las filas.

Por ejemplo, la primera coordenada principal de las filas verifica:

$$a_{i1} = \frac{1}{\lambda_1}(b_{11}\frac{p_{i1}}{r_i} + b_{21}\frac{p_{i2}}{r_i} + \cdots + b_{J1}\frac{p_{iJ}}{r_i}), \quad i = 1, \ldots, I,$$

y la primera coordenada principal de las columnas verifica

$$b_{j1} = \frac{1}{\lambda_1}(a_{11}\frac{p_{1j}}{c_j} + a_{21}\frac{p_{2j}}{c_j} + \cdots + a_{I1}\frac{p_{Ij}}{c_j}), \quad j = 1, \ldots, J.$$

Ejemplo 1. La Tabla 9.1 contiene unos datos artificiales, que clasifican 400 clientes según la edad (joven, mediana, mayor) y los productos que compran en un supermercado.

Tenemos:

$$\mathbf{P} = \begin{pmatrix} .175 & 0 & 0 \\ .112\,5 & .1125 & 0 \\ .075 & .075 & .075 \\ 0 & .2 & .05 \\ .0875 & .0125 & .025 \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} .175 \\ .225 \\ .225 \\ .250 \\ .125 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} .45 \\ .40 \\ .15 \end{pmatrix}.$$

La matriz de perfiles de las filas es:

$$\begin{pmatrix} 1.00 & 0 & 0 \\ 0.50 & 0.50 & 0 \\ 0.33 & 0.33 & 0.33 \\ 0 & 0.80 & 0.20 \\ 0.70 & 0.10 & 0.20 \end{pmatrix}$$

	Edad			
Producto	Joven	Mediana	Mayor	Total
A	70	0	0	70
B	45	45	0	90
C	30	30	30	90
D	0	80	20	100
E	35	5	10	50
Total	180	160	60	400

Tabla 9.1: Clasificación de 400 clientes según edades y productos adquiridos en un supermercado.

Las coordenadas principales son:

$$\mathbf{A} = \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \text{Filas} \\ \left[\begin{array}{rr} 1.0990 & -0.1199 \\ 0.0551 & -0.4213 \\ -0.1834 & 0.4815 \\ -0.9231 & -0.1208 \\ 0.5384 & 0.3012 \end{array}\right] \end{matrix} \quad \mathbf{B} = \begin{matrix} \text{Columnas} \\ \left[\begin{array}{rr} 0.7525 & -0.0397 \\ -0.6770 & -0.2393 \\ -0.4522 & 0.7571 \end{array}\right] \end{matrix}$$

Los valores singulares son: $\lambda_1 = 0.6847, \lambda_2 = 0.3311$. La primera coordenada principal de las filas $A_1, \ldots, A_5$ verifica:

$$\begin{aligned} 1.0990 &= 0.6847^{-1}(.7525 \times 1 + 0 + 0) \\ 0.0551 &= 0.6847^{-1}(.7525 \times .5 - .677 \times .5 + 0) \\ -0.1834 &= 0.6847^{-1}(.7525 \times .33 - .677 \times .33 - .4522 \times .33) \\ -0.9231 &= 0.6847^{-1}(0 - .677 \times .8 - .4522 \times .2) \\ 0.5384 &= 0.6847^{-1}(.7525 \times .7 - .677 \times .1 - .4522 \times .2) \end{aligned}$$

Las coordenadas de las marcas A,B,C,D,E son medias de las coordenadas de las tres edades, ponderadas por la incidencia del producto en la edad.

9.5 Soluciones simétrica y asimétrica

La representación de filas y columnas utilizando las coordenadas principales $\mathbf{A}, \mathbf{B}$ es la solución simétrica. La representación conjunta es posible gracias a las fórmulas (9.5). La representación utilizando las matrices

$$\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\lambda, \quad \mathbf{B}_0 = \mathbf{D}_c^{-1/2}\mathbf{V},$$

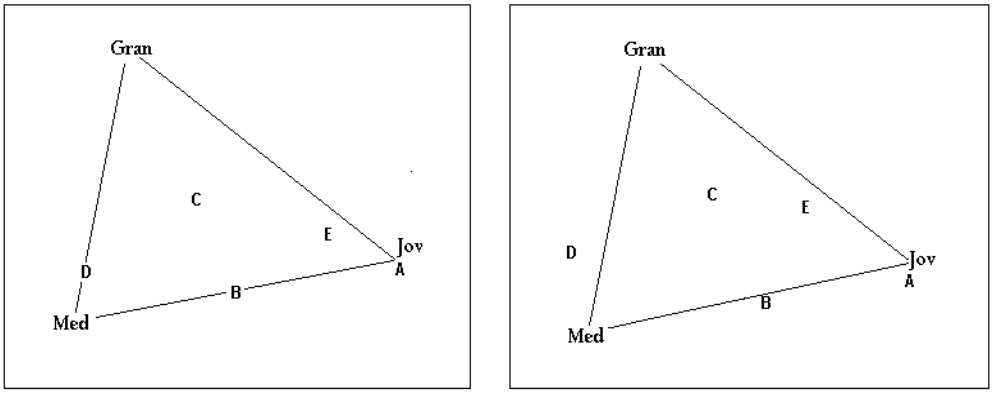

Figura 9.1: Representación asimétrica (izquierda) y simétrica (derecha) de las filas (productos) y columnas (edades) de la Tabla 9.1.

			Color	cabellos		
Color ojos	Rubio	Rojo	Castaño	Oscuro	Negro	Total
CLARO	688	116	584	188	4	1,580
AZUL	326	38	241	110	3	718
CASTAÑO	343	84	909	412	26	1,774
OSCURO	98	48	403	681	81	1,311
Total	1,455	286	2,137	1,391	114	5,383

Tabla 9.2: Classificación de 5383 individuos según el color de los ojos y del cabello.

es decir, coordenadas principales para las filas y coordenadas estándard para las columnas, es la llamada solución *asimétrica*. Esta solución verifica

$$\mathbf{P} - \mathbf{rc}' = \mathbf{D}_r \mathbf{A} \mathbf{B}_0' \mathbf{D}_c,$$

y por lo tanto reproduce mejor la dependencia entre filas y columnas.

Ejemplo 2. La Tabla 9.2 relaciona los colores de los cabellos y de los ojos de 5,383 individuos.

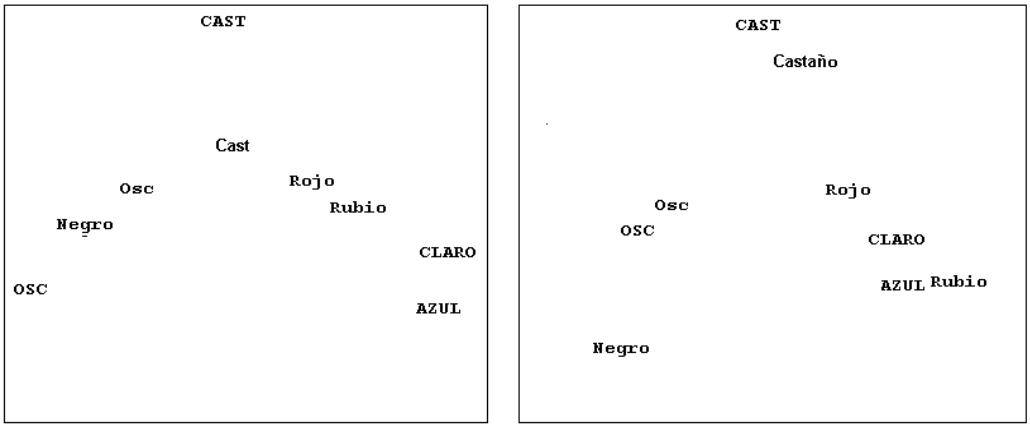

Figura 9.2: Representación asimétrica (izquierda) y simétrica (derecha) de los datos de los colores de ojos y cabellos.

Las coordenadas principales son:

$$\mathbf{A} = \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \text{Filas} \\ \left[\begin{array}{rr} 0.4400 & -0.0872 \\ 0.3996 & -0.1647 \\ -0.0361 & 0.2437 \\ -0.7002 & -0.1345 \end{array}\right] \end{matrix} \quad \mathbf{B} = \begin{matrix} \text{Columnas} \\ \left[\begin{array}{rr} 0.5437 & -0.1722 \\ 0.2324 & -0.0477 \\ 0.0402 & 0.2079 \\ -0.5891 & -0.1070 \\ -1.0784 & -0.2743 \end{array}\right] \end{matrix}$$

Los valores singulares son: $\lambda_1 = 0.449, \lambda_2 = 0.1727, \lambda_3 = 0.0292$. De acuerdo con (9.6), la variabilidad explicada por las dos primeras dimensiones principales es $P_2 = 86.8\%$. La Figura 9.2 proporciona las representaciones simétrica y asimétrica.

9.6 Variabilitadad geométrica (inercia)

Vamos a probar que

$$\chi^2 = n \sum_{k=1}^{m} \lambda_k^2,$$

siendo

$$\chi^2 = n \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j}/n)^2}{f_{i\cdot} f_{\cdot j}}$$

el estadístico ji-cuadrado con $(I-1)(J-1)$ g.l. que permite decidir si hay independencia entre filas y columnas de $\mathbf{N}$. Es decir, la ji-cuadrado es n veces la suma de los valores propios del AC.

El coeficiente ϕ^2 de Pearson se define como

$$\phi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = \frac{\chi^2}{n},$$

Es fácil probar que también podemos expresar

$$\phi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{p_{ij}^2}{r_i c_j} - 1.$$

La variabilidad geométrica ponderada de la distancia ji-cuadrado entre filas es

$$V_\delta = \frac{1}{2} \sum_{i=1}^{I} \sum_{i'=1}^{I} r_i \delta_{ii'}^2 r_{i'}.$$

Proposició 9.6.1 $V_\delta = \phi^2$.

Prueba:

$$\delta_{ii'}^2 = \sum_{j=1}^{J} \frac{(p_{ij}/r_i - p_{i'j}/r_{i'})^2}{c_j} = \sum_{j=1}^{J} (\frac{p_{ij}}{r_i c_j} - \frac{p_{i'j}}{r_{i'} c_j})^2 c_j$$

Por lo tanto

$$V_\delta = \frac{1}{2} \sum_{i=1}^{I} \sum_{i=1}^{I} \sum_{j=1}^{J} r_i (\frac{p_{ij}}{r_i c_j} - \frac{p_{i'j}}{r_{i'} c_j})^2 c_j r_{i'}$$

Si desarrollamos por un lado

$$\begin{aligned} \textstyle\sum_{i=1}^{I} \sum_{i'=1}^{I} \sum_{j=1}^{J} r_i \frac{p_{ij}^2}{r_i^2 c_j^2} c_j r_{i'} &= \textstyle\sum_{i=1}^{I} \sum_{i'=1}^{I} \sum_{j=1}^{J} \frac{p_{ij}^2}{r_i c_j} r_{i'} \\ &= \textstyle\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{p_{ij}^2}{r_i c_j}, \end{aligned}$$

y por otro lado, dado que $\sum_{i'=1}^{I} p_{ij} = c_j$,

$$\begin{aligned}\sum_{i=1}^{I}\sum_{i=1}^{I}\sum_{j=1}^{J} r_i \frac{p_{ij}p_{i'j}}{r_i c_j^2 r_{i'}} c_j r_{i'} &= \sum_{i=1}^{I}\sum_{i'=1}^{I}\sum_{j=1}^{J} \frac{p_{ij}p_{i'j}}{c_j} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{p_{ij}c_j}{c_j} = 1,\end{aligned}$$

es decir, vemos que $V_\delta = (\alpha + \alpha - 2)/2$, siendo $\alpha = \sum_{i,j} \frac{p_{ij}^2}{r_i c_j}$.

Proposició 9.6.2 $\phi^2 = \sum_{k=1}^{I} \lambda_k^2$.

Prueba: Sea

$$\mathbf{W} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}'.$$

Entonces

$$\phi^2 = \mathrm{tr}(\mathbf{W}\mathbf{W}') = \mathrm{tr}(\mathbf{U}\mathbf{D}_\lambda^2\mathbf{U}') = tr(\mathbf{D}_\lambda^2).$$

Proposició 9.6.3 *La variabilidad geométrica utilizando sólo las primeras m coordenadas principales es*

$$V_\delta(m) = \sum_{k=1}^{m} \lambda_k^2.$$

Prueba: Supongamos $m = J$. Podemos escribir la matriz de distancias entre filas como

$$\Delta^{(2)} = \mathbf{a}\mathbf{1}' + \mathbf{1}\mathbf{a}' - 2\mathbf{A}\mathbf{A}',$$

siendo $\mathbf{a}$ el vector columna que contiene los elementos de la diagonal de $\mathbf{A}\mathbf{A}'$. Entonces

$$V_\delta = \frac{1}{2}\mathbf{r}'\Delta^{(2)}\mathbf{r} = \mathbf{r}'\mathbf{a}\mathbf{1}'\mathbf{r} + \mathbf{r}'\mathbf{1}\mathbf{a}'\mathbf{r} - 2\mathbf{r}'\mathbf{A}\mathbf{A}'\mathbf{r} = \mathbf{r}'\mathbf{a}.$$

Pero

$$\mathbf{r}'\mathbf{a} = \mathrm{tr}(\mathbf{D}_r^{1/2}\mathbf{A}\mathbf{A}'\mathbf{D}_r^{1/2}) = \mathrm{tr}(\mathbf{U}\mathbf{D}_\lambda^2\mathbf{U}') = \mathrm{tr}(\mathbf{D}_\lambda^2).$$

Lo hemos probado para $m = J$, pero fácilmente vemos que la fórmula también vale para $m < J$.□

Así pues, en la representación por AC de las filas y columnas de $\mathbf{N}$ en dimensión m, el porcentaje de variabilidad geométrica o inercia viene dado por

$$P_m = 100 \times \frac{\sum_{k=1}^{m} \lambda_k^2}{\sum_{k=1}^{K} \lambda_k^2}. \tag{9.6}$$

9.7 Analisis de Correspondencias Múltiples

El AC combina y representa dos variables categóricas. Pero se puede adaptar para estudiar más de dos variables. Presentemos primero el procedimiento para dos, que después generalizaremos.

Escribimos la matriz $n \times (I + J)$ de datos binarios como una matriz $n \times (J_1 + J_2)$

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2].$$

Entonces tenemos que

$$\mathbf{B}_u = \mathbf{Z}'\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1'\mathbf{Z}_1 & \mathbf{Z}_1'\mathbf{Z}_2 \\ \mathbf{Z}_2'\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{Z}_2 \end{bmatrix} = n \begin{bmatrix} \mathbf{D}_r & \mathbf{P} \\ \mathbf{P}' & \mathbf{D}_c \end{bmatrix}.$$

La matriz de frecuencias, donde $\mathbf{F}$ y $\mathbf{C}$ contienen las marginales de filas y columnas,

$$\mathbf{B}_u = \begin{bmatrix} \mathbf{F} & \mathbf{N} \\ \mathbf{N}' & \mathbf{C} \end{bmatrix}$$

es la llamada matriz de Burt. A continuación podemos realizar tres análisis de correspondencias diferentes sobre las matrices:

a) $\mathbf{N}$. b) $[\mathbf{Z}_1, \mathbf{Z}_2]$. c) $\mathbf{B}_u$.

El análisis a) lo hemos vistos en las secciones anteriors. El resultado es una representación de filas y columnas de $\mathbf{N}$.

El análisis b) es sobre $[\mathbf{Z}_1, \mathbf{Z}_2]$, considerada una matriz binaria con n filas y $J_1 + J_2$ columnas. AC nos daría una representación de las $J_1 + J_2$ columnas, que es la interesante, y de los n individuos, pero esta segunda representación es innecessaria.

El análisis c) es sobre $\mathbf{B}_u$ que es la matriz simétrica de orden $(J_1 + J_2) \times (J_1 + J_2)$. Tendremos una representación idéntica por columnas y por filas.

En los tres casos vemos que podemos representar las filas y columnas de $\mathbf{N}$. Es posible demostrar que los tres análisis son equivalentes en el sentido de que proporcionan la misma representación, variando sólo los valores propios.

Todo esto se describe en el cuadro que sigue.

Tabla	Dimensión	Coordenadas	Valor propio
$\mathbf{N} = \mathbf{Z}_1'\mathbf{Z}_2$	$J_1 \times J_2$	**A** (filas) **B** (columnas)	λ
$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$	$n \times (J_1 + J_2)$	$\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}$	$\frac{1+\sqrt{\lambda}}{2}$
$\mathbf{B}_u = \mathbf{Z}'\mathbf{Z}$	$(J_1 + J_2) \times (J_1 + J_2)$	$\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}$	$(\frac{1+\sqrt{\lambda}}{2})^2$

Consideremos a continuación Q variables categòricas con $J_1, \ldots, J_Q$ estados, respectivamente, sobre n individuos. Sea $J = J_1 + \ldots + J_Q$. La tabla de datos, de orden $n \times J$ es la super-matriz de indicadores

$$\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_j, \ldots, \mathbf{Z}_q],$$

donde $\mathbf{Z}_j$ es $n \times J_j$ y contiene los datos binarios de la variable j. La tabla de contingencia que tabula la combinación de las variables i, j es $\mathbf{N}_{ij} = \mathbf{Z}_i'\mathbf{Z}_j$. La matriz de Burt, de orden $J \times J$ es

$$\mathbf{B}_u = \mathbf{Z}'\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1'\mathbf{Z}_1 & \mathbf{Z}_1'\mathbf{Z}_2 & \cdots & \mathbf{Z}_1'\mathbf{Z}_q \\ \mathbf{Z}_2'\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{Z}_2 & \cdots & \mathbf{Z}_2'\mathbf{Z}_q \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}_q'\mathbf{Z}_1 & \mathbf{Z}_q'\mathbf{Z}_2 & \cdots & \mathbf{Z}_q'\mathbf{Z}_q \end{bmatrix},$$

donde las matrices $\mathbf{Z}_j'\mathbf{Z}_j$ són diagonales.

El Anàlisis de Correspondències Múltiples intenta representar los $J = J_1 + \ldots + J_q$ estados de las q variables categòricas. Como en el caso $Q = 2$, lo podemos hacer aplicando un AC simple sobre leas matrices:

$$\text{a) } \mathbf{Z}. \qquad \text{b) } \mathbf{B}_u.$$

E en caso a) representampos las J columnas y ignoramos las n filas (individuos). En el caso b) tenemos una tabla de frecuencias $J \times J$ simétrica y podemos representar las filas (=columnas) aplicando AC simple. Los dos procedimientos son equivalentes, salvo que se cumple la relación

$$\lambda_k^B = (\lambda_k^Z)^2$$

entre los valores propios λ_i^B obtenidos a partir de la matriz de Burt y los λ_i^Z que surgen del análisis sobre $\mathbf{Z}$. Las inercias correspondientes son:

$$\begin{aligned}\phi^2(\mathbf{B}_u) &= \textstyle\sum_k \lambda_k^B = \frac{1}{Q^2}[\sum_{i\neq j}\phi^2(N_{ij}) + (J-Q)],\\ \phi^2(\mathbf{Z}) &= \textstyle\sum_k \lambda_k^Z = \frac{J}{Q} - 1,\end{aligned}$$

siemdo $\phi^2(N_{ij})$ la inercia para la tabla N_{ij}, véase Secció **??**. Así pues podemos constatar que AC puede servir también para representar más de dos variables categòriques.

Exemple 9.7.1 La Tabla 9.3 contiene las frecuencias con la clasifcación cruzada de 1257 individuos segun Edad (E), Sexo (S), intención de Voto (V) y Clase social (C). Tenemos $Q = 4, J = 12, J_1 = 4, J_2 = 2, J_3 = 3, J_4 = 2$. Los datos (matriz $\mathbf{Z}$, solo mostramos 5 individuos) son de la forma:

Edad					Votación		Clase			Sexo	
>73	51-73	41-50	26-40	<26	Lib	Con	Alt	Mit	Obr	H	D
0	1	0	0	0	1	0	0	1	0	1	0
0	1	0	0	0	0	1	1	0	0	0	1
0	0	0	0	1	1	0	0	0	1	1	0
1	0	0	0	0	0	1	1	0	0	0	1
0	1	0	0	0	1	0	0	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

La Tabla 9.4 (abajo) es la tabla de Burt. Observemos que es simétrica. El AC sobre esta tabla nos permite representar las 4 variables categóricas sobre el mismo gráfico, véase la Figura 9.3.

Edad	Hombres		Mujeres	
	Derecha	Izquierda	Derecha	Izquierda
		Clase	alta	
>73	4	0	10	0
51-73	27	8	26	9
41-50	27	4	25	9
26-40	17	12	28	9
<26	7	6	7	3
		Clase	media	
>73	8	4	9	1
51-73	21	13	33	8
41-50	27	12	29	4
26-40	14	15	17	13
<26	9	9	13	7
		Clase	obrera	
>73	8	15	17	4
51-73	35	62	52	53
41-50	29	75	32	70
26-40	32	66	36	67
<26	14	34	18	33

Tabla 9.3: Tabla de frecuencias combinando 1257 individuos según edad, sexo, clase social y tendencia de voto.

81	0	0	0	0	56	25	14	23	44	39	42
0	347	0	0	0	194	153	70	75	202	166	181
0	0	343	0	0	169	174	65	72	206	174	169
0	0	0	326	0	144	182	66	59	201	156	170
0	0	0	0	160	68	92	23	38	99	79	81
56	194	169	144	68	631	0	178	180	273	279	352
25	153	174	182	92	0	626	60	87	479	335	291
14	70	65	66	23	178	60	238	0	0	112	126
23	75	72	59	38	180	87	0	267	0	132	135
44	202	206	201	99	273	479	0	0	752	370	382
39	166	174	156	79	279	335	112	132	370	614	0
42	181	169	170	81	352	291	126	135	382	0	643

Tabla 9.4: Tabla de Burt con la clasificación de 1257 individuos según edad, sexo, clase social y tendencia de voto.

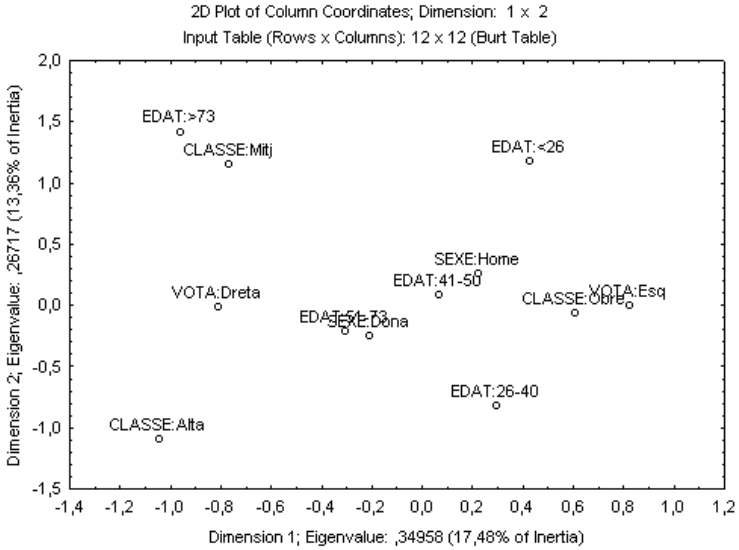

Figura 9.3: Representación por análisis de correspondencias múltiples de los datos de la Tabla 9.3.

9.8 MDS ponderado

En esta sección introducimos una variante del Análisis de Coordenadas Principales.

Definició 9.8.1 *Sea* $\Delta_g = (\delta_{ij})$ *una matriz de distancias* $g \times g$, $\mathbf{w} = (w_1, \ldots, w_g)'$ *un vector de pesos tal que*

$$\mathbf{w}'\mathbf{1} = \sum_{i=1}^{g} w_i = 1, \quad w_i \geq 0,$$

y consideremos la matriz diagonal $\mathbf{D}_w = diag(\mathbf{w})$. *La solución MDS ponderada de* Δ_g *es la matriz*

$$\mathbf{X} = \mathbf{D}_w^{-1/2}\mathbf{U}\mathbf{\Lambda},$$

siendo

$$\mathbf{D}_w^{1/2}(\mathbf{I}_g - \mathbf{1}\mathbf{w}')(-\frac{1}{2}\mathbf{\Delta}_g^{(2)})(\mathbf{I}_g - \mathbf{w}\mathbf{1}')\mathbf{D}_w^{1/2} = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}', \tag{9.7}$$

una descomposición espectral, donde $\Lambda = diag(\lambda_1^2, \ldots, \lambda_p^2)$ *contiene los valores propios y* $\Delta_g^{(2)} = (\delta_{ij}^2)$.

Definició 9.8.2 *La variabilidad geométrica ponderada de* Δ_g *es*

$$V_\delta = \frac{1}{2}\sum_{i,j=1}^{n} w_i \delta_{ij}^2 w_j = \frac{1}{2}\mathbf{w}'\Delta_g^{(2)}\mathbf{w}.$$

Las coordenadas principales son las filas de $\mathbf{X}$. Escribiendo

$$\mathbf{X} = [X_1, X_2, \ldots, X_p],$$

podemos interpretar las columnas de $\mathbf{X}$ como variables. Observemos que se verifica

$$(\mathbf{I}_g - \mathbf{1}\mathbf{w}')(-\frac{1}{2}\Delta_g^{(2)})(\mathbf{I}_g - \mathbf{w}\mathbf{1}') = \mathbf{X}\mathbf{X}'. \tag{9.8}$$

Propietades:

1. Las variables X_k (columnas de $\mathbf{X}$) tienen medias ponderadas iguales a cero:
$$\overline{X}_k = \mathbf{w}'X_k = 0.$$
Prueba:
$$\mathbf{w}'(\mathbf{I}_g - \mathbf{1}\mathbf{w}') = \mathbf{w}' - \mathbf{w}' = \mathbf{0} \Rightarrow \mathbf{w}'\mathbf{X}\mathbf{X}'\mathbf{w} = \mathbf{0} \Rightarrow \mathbf{w}'\mathbf{X} = \mathbf{0}.$$

2. Las varianzas ponderadas de las variables $X_k$son iguales a los valores propios:
$$s_k^2 = \lambda_k^2, \quad k = 1, \ldots, p.$$
Prueba: si la media de $x_1, \ldots, x_g$ es 0, la varianza ponderada es $\sum w_i x_i^2$, es decir,
$$s_k^2 = \mathbf{D}_w^{1/2}\mathbf{X}_k\mathbf{X}_k'\mathbf{D}_w^{1/2} = (\mathbf{U}_k'\boldsymbol{\lambda}_k)(\boldsymbol{\lambda}_k\mathbf{U}_k) = \lambda_k^2,$$
donde λ_k^2 es el valor propio de vector propio U_k.

3. Las variables (columnas de $\mathbf{X}$) están incorrelacionadas
$$\mathrm{cor}(X_k, X_{k'}) = 0, \quad k \neq k' = 1, \ldots, p.$$
Prueba: puesto que las medias son nulas la covarianza ponderada es
$$\mathrm{cov}(X_k, X_{k'}) = \mathbf{D}_w^{1/2}\mathbf{X}_k'\mathbf{X}_{k'}\mathbf{D}_w^{1/2} = \lambda_k^2 U_k' U_{k'} = 0,$$
ya que los vectores propios son ortogonales.

4. La variabilidad geométrica ponderada de Δ_g es

$$V_\delta = \sum_{k=1}^{p} \lambda_k^2.$$

Prueba: Expresemos la matriz de distancias al cuadrado como

$$\Delta_g^{(2)} = \mathbf{1d'} + \mathbf{d1'} - 2\mathbf{XX'},$$

siendo $\mathbf{d}$ un vector $g \times 1$ con los elementos diagonales de $\mathbf{XX'}$. Por una parte

$$\frac{1}{2}\mathbf{w'}\Delta_g^{(2)}\mathbf{w} = \mathbf{w'1d'w} - \mathbf{w'XX'w} = \mathbf{d'w}.$$

Por otra parte

$$\mathbf{d'w} = \mathrm{tr}(\mathbf{D}_w^{1/2}\mathbf{XX'D}_w^{1/2}) = \mathrm{tr}(\mathbf{U\Lambda^2U'}) = \mathrm{tr}(\Lambda^2).$$

5. Si tomamos las q primeras coordenadas principales de $\mathbf{X}$, la variabilidad geométrica ponderada es:

$$V_\delta(q) = \sum_{k=1}^{q} \lambda_k^2.$$

Estudiemos ahora la relación entre el Análisis de Coordenadas Principales ordinario (Cap. 8) y el ponderado. Supongamos que podemos expresar el vector de pesos como

$$\mathbf{w} = \frac{1}{n}(n_1, n_2, \ldots, n_k), \qquad n = \sum_{i=1}^{g} n_i,$$

donde n_i son enteros positivos y el peso w_i es igual (o muy próximo [1]) a n_i/n. Indiquemos por $\mathbf{M}$ la matriz $n \times g$ que contiene n_i filas $(0, \ldots, 1, \ldots, 0)$. Por ejemplo, si $g = 3$ y $n_1 = 2, n_2 = 3, n_3 = 1$, entonces

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

[1] Tomando n suficientmente grande, podemos aproximarlo tanto como queramos.

Si ahora suponemos que en vez de g objetos tenemos n objetos, pero el primer objeto está repetido n_1 veces, el segundo objeto n_2 veces, etc., entonces la matriz de distancias es

$$\Delta_n = \mathbf{M}\Delta_g\mathbf{M}', \tag{9.9}$$

y el análisis no ponderado sobre la matriz Δ_n es

$$(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}')(-\frac{1}{2}\Delta_n^{(2)})(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}') = \widetilde{\mathbf{U}}\mathbf{D}_\lambda^2\widetilde{\mathbf{U}}' = \mathbf{Y}\mathbf{Y}', \tag{9.10}$$

siendo $\widetilde{\mathbf{U}}$ la matriz $n \times p$ de los vectores propios. La solución no ponderada es

$$\mathbf{Y} = \widetilde{\mathbf{U}}\mathbf{D}_\lambda.$$

Teorema 9.8.1 *La solución no ponderada* $\mathbf{Y}$ *sobre* Δ_n *coincide con la solución ponderada* $\mathbf{X}$ *sobre* Δ_g*, en el sentido de que obtenemos* $\mathbf{Y}$ *repitiendo* $n_1, \ldots, n_g$ *veces las filas de* $\mathbf{X}$.

Prueba: De (9.9) podemos expresar la solución no ponderada (9.10) como

$$(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{M}(-\frac{1}{2}\Delta_g^{(2)})\mathbf{M}'(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}') = \mathbf{Y}\mathbf{Y}'.$$

Se verifica

$$(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{M} = \mathbf{M}(\mathbf{I}_g - \mathbf{1}_g\mathbf{w}').$$

Por lo tanto, de (9.8) tenemos

$$\mathbf{M}(\mathbf{I}_g - \mathbf{1}\mathbf{w}')(-\frac{1}{2}\Delta_g^{(2)})(\mathbf{I}_g - \mathbf{w}\mathbf{1}')\mathbf{M}' = \mathbf{M}\mathbf{X}\mathbf{X}'\mathbf{M}',$$

que demuestra que $\mathbf{Y} = \mathbf{M}\mathbf{X}$. En otras palabras, las coordenadas principales no ponderadas $\mathbf{Y}$ son el resultado de repetir $n_1, \ldots, n_g$ veces las coordenadas $\mathbf{X}$. La relación entre los valores singulares es

$$\widetilde{\lambda}_k = g\lambda_k, \quad k = 1. \ldots, p.$$

Por ejemplo, si $g = 3$ y $n_1 = 2, n_2 = 3, n_3 = 1$, obtenemos

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} x_{11} & x_{12} \\ x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{21} & x_{22} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix}.$$

9.9 Complementos

El Análisis de Correspondencias (AC) tiene una larga historia que se inicia en 1935 (H.O. Hirschfeld, R.A. Fisher, L. Guttman). Ha sido extensamente estudiado por Benzécri (1973) y Greenacre (1984).

Utilitzando coordenadas estándard $\mathbf{A}_0 = (a^0_{ik}), \mathbf{B}_0 = (b^0_{jk})$, podemos expresar la matriz de correspondencias $\mathbf{P} = (p_{ij})$ como

$$\mathbf{P} = \mathbf{r}\mathbf{c}' + \mathbf{D}_r\mathbf{A}_0\mathbf{D}_\lambda\mathbf{B}_0'\mathbf{D}_c.$$

Indicando $\mathbf{r} = (p_{1\cdot}, \ldots, p_{I\cdot})', \mathbf{c} = (p_{\cdot 1}, \ldots, p_{\cdot J})'$ los vectores marginales de filas y columnas de $\mathbf{P}$, la expresión escalar es

$$p_{ij} = p_{i\cdot} \times p_{\cdot j}(1 + \sum_{k=1}^{K} \lambda_k a^0_{ik} b^0_{jk}).$$

Si el término entre paréntesis $\alpha = \sum_{k=1}^{K} \lambda_k a^0_{ik} b^0_{jk}$, es suficientemente pequeño para que $\log(1+\alpha) \approx \alpha$, entonces

$$\log p_{ij} = \log p_{i\cdot} + \log p_{\cdot j} + \sum_{k=1}^{K} \lambda_k a^0_{ik} b^0_{jk},$$

que se adapta a un modelo log-lineal (Sección 11.5), donde α cuantificaría el término de interacción. El AC sería pues una manera de visualizar los términos de interacción (van der Heijden y de Leeuw, 1985).

CA verifica el "principio de equivalencia distribucional": si dos perfiles de columnas son idénticos, es decir,

$$p_{ij}/c_j = p_{ij'}/c_{j'}, \quad i = 1, \ldots, I,$$

entonces las columnas j, j' de $\mathbf{N}$ pueden juntarse y ser reemplazadas por su suma. En efecto, cuando se cumple este principio

$$\frac{p_{ij}}{c_j} = \frac{p_{ij'}}{c_{j'}} = \frac{p_{ij} + p_{ij'}}{c_j + c_{j'}}.$$

Luego

$$[(\frac{p_{ij}}{r_i c_j}) - (\frac{p_{i'j}}{r_{i'} c_j})]^2 c_j + [(\frac{p_{ij'}}{r_i c_{j'}}) - (\frac{p_{i'j'}}{r_{i'} c_{j'}})]^2 c_{j'} = [(\frac{p_{ij} + p_{ij'}}{r_i(c_j + c_{j'})}) - (\frac{p_{i'j} + p_{i'j'}}{r_{i'}(c_j + c_{j'})})]^2 (c_j + c_{j'}),$$

y la distancia ji-cuadrado queda inalterada si juntamos las columnas j y j'.

Una variante del AC propuesta por Rao (1995), se basa en la distancia de Hellinger

$$\widetilde{\delta}^2_{ii'} = \sum_{j=1}^{J} (\sqrt{p_{ij}/r_i} - \sqrt{p_{i'j}/r_{i'}})^2,$$

entre dos filas de $\mathbf{N}$, que tiene la ventaja de no depender de los perfiles de las columnas. Sin embargo los resultados pueden ser muy similares (Cuadras *et al*, 2004), y el método basado en esta distancia resulta más apropiado cuando las filas se ajustan a poblaciones multinomiales distintas.

Una forma alternativa de presentar el AC es el "reciprocal averaging" (RA). Supongamos que queremos encontrar las coordenadas de las filas $(a_1, \ldots, a_I)$, como medias ponderadas de las coordenadas de las columnas y recíprocamente, las coordenadas de las columnas $(b_1, \ldots, b_J)$ como medias ponderadas de las coordenadas de las filas

$$a_i = \sum_{j=1}^{J} b_j \frac{p_{ij}}{r_i}, \qquad b_j = \sum_{i=1}^{I} a_i \frac{p_{ij}}{c_j}.$$

Pero estas relaciones no se pueden verificar simultáneamente (por razones geométricas), así que hemos de introducir un factor multiplicativo $\beta > 1$ y escribir

$$a_i = \beta \sum_{j=1}^{J} b_j \frac{p_{ij}}{r_i}, \qquad b_j = \beta \sum_{i=1}^{I} a_i \frac{p_{ij}}{c_j}. \tag{9.11}$$

El objectivo del RA es encontrar las coordenadas verificando (9.11) tal que β sea mínimo. Entonces es posible probar que $\lambda = (1/\beta)^2$ es un valor propio. Esto mismo lo podemos plantear para la segunda y siguientes coordenadas y probar la equivalencia entre RA y AC. Los cálculos del RA se efectúan iterativamente, y es útil (especialmente en ecología), cuando la matriz de frecuencias $\mathbf{N}$ tiene dimensión grande y contiene muchos ceros (Hill, 1973). Por otra parte se conoce a (9.11) como la mejor representación $\beta-$baricéntrica sobre un eje (Lebart *et al.*, 1977).

Una extensión interesante del AC es el "Canonical Correspondence Analysis" (Ter Braak, 1986), que tiene en cuenta, para la representación, que los ejes sean combinación lineal de variables externas. Tiene aplicaciones en ecología, dado que permite relacionar las comunidades biológicas con las variables ambientales.

Una extensión continua del AC considera una densidad bivariante $h(x, y)$ con densidades marginales $f(x), g(y)$, y la descomposición singular

$$f(x)^{-1/2}h(x,y)g(y)^{-1/2} = \sum_{k=1}^{\infty} \rho_k u_k(x) v_k(y), \tag{9.12}$$

donde $\{\rho_k, k \geq 1\}$ son correlaciones canónicas y $\{u_k, k \geq 1\}, \{v_k, k \geq 1\}$ son sistemas de funciones ortonormales (Lancaster, 1969). Hay una interesante semejanza entre (9.12) y el AC, pues muchas propiedades se conservan. Véase una comparación sistemática en Cuadras *et al.* (2000) y Cuadras (2002b).

El AC ha sido también comparado con otros métodos de representación de tablas de contingencia (Cuadras *et al.*, 2006), propiciando una versión paramétrica que los engloba a todos (Cuadras y Cuadras, 2006).

Capítulo 10

CLASIFICACIÓN

10.1 Introducción

Clasificar los elementos de un conjunto finito consiste en realizar una partición del conjunto en subconjuntos homogéneos, siguiendo un determinado criterio de clasificación. Cada elemento pertenece a un único subconjunto, que a menudo tiene un nombre que lo caracteriza. Así clasificamos:

- Las personas en hombres y mujeres.
- Los trabajadores en actividades profesionales: servicios, industria, agricultura.
- Los animales en especies, géneros, familias y órdenes.
- Los libros de una biblioteca en arte, literatura, ciencia, informática y viajes.

Sea $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ un conjunto finito con n elementos diferentes, que abreviadamente indicaremos

$$\Omega = \{1, 2, ..., n\}.$$

Clasificar es también definir una relación de equivalencia $\mathcal{R}$ sobre Ω. Esta relación define una partición sobre Ω en m clases de equivalencia:

$$\Omega = c_1 + c_2 + \ldots + c_m,$$

donde + significa reunión disjunta. A la partición la llamaremos *clustering* y a las clases de equivalencia *clusters*.

10.2 Jerarquía indexada

Las clasificaciones pueden ser jerárquicas o no jerárquicas . Una clasificación jerárquica es una sucesión de clusterings tal que cada clustering se obtiene agrupando clusters. Por ejemplo, si $n = 5$,

$$\begin{aligned}\Omega &= \{1\} + \{2\} + \{3\} + \{4\} + \{5\}\\ \Omega &= \{1,2\} + \{3,4\} + \{5\}\\ \Omega &= \{1,2\} + \{3,4,5\}\\ \Omega &= \Omega\end{aligned}$$

Definició 10.2.1 *Una jerarquía indexada (C, α) sobre Ω está formada por una colección de clusters $C \subset \wp(\Omega)$ y un índice α tal que:*

- *Axioma de la intersección: Si $c, c' \in C$ entonces $c \cap c' \in \{c, c', \emptyset\}$.*
- *Axioma de la reunión: Si $c \in C$ entonces $c = \cup\{c' \mid c' \in C,\ c' \subset c\}$.*
- *La reunión de todos los clusters es el conjunto total: $\Omega = \cup\{c \mid c \in C\}$.*

El índice α es una aplicación de C sobre el conjunto de números reales positivos tal que:

$$\alpha(i) = 0, \forall i \in \Omega, \qquad \alpha(c) \leq \alpha(c') \text{ si } c \subset c'.$$

Diremos que una jerarquía es total si:

- $\forall i \in \Omega, \quad \{i\} \in C$.
- $\Omega \in C$.

Comentarios:

1. El primer axioma significa que si tenemos dos clusters, uno está incluido en el otro o ambos son disjuntos, es decir, $c \subset c'$, ó $c' \subset c$, ó $c \cap c' = \emptyset$. Se trata de evitar que un elemento de Ω pertenezca a dos clusters excluyentes a la vez, ya que entonces estaría mal clasificado.
2. El segundo axioma significa que cada cluster es reunión de los clusters que contiene. Es decir, reuniendo clusters obtenemos clusters más amplios. Por ejemplo, en el reino animal, un género es reunión de especies, una familia es reunión de géneros, etc.

3. El índice α mide el grado de heterogeneidad de cada cluster. Cuanto más grande es el cluster más heterogéneo es.

Teorema 10.2.1 *Para todo $x \geq 0$ la relación binaria $\mathcal{R}_x$ sobre los elementos de Ω*

$$i\mathcal{R}_x j \quad si \quad i,j \in c, \quad siendo \quad \alpha(c) \leq x, \tag{10.1}$$

es de equivalencia.

Demost.: La relación $\mathcal{R}_x$ es:

Reflexiva: $i\mathcal{R}_x i$ ya que $i \in \{i\}$, siendo $\alpha(\{i\}) = 0 \leq x$.

Simétrica: Evidente.

Transitiva: Sea c_{ij} el mínimo cluster que contiene i,j, y análogamente c_{jk}. Entonces :

$$i\mathcal{R}_x j \Rightarrow i,j \in c_{ij,} \quad \alpha(c_{ij}) \leq x, \quad j\mathcal{R}_x k \Rightarrow j,k \in c_{jk,} \quad \alpha(c_{jk}) \leq x,$$

$$\Rightarrow c_{ij} \cap c_{jk} \neq \emptyset \Rightarrow \left\{ \begin{array}{ll} a) & c_{ij} \subset c_{jk} \Rightarrow i,k \in c_{jk,} \\ b) & c_{jk} \subset c_{ij} \Rightarrow i,k \in c_{ij,} \end{array} \right. \Rightarrow i\mathcal{R}_x k. \square$$

La relación (10.1) define, para cada $x \geq 0$, una partición de Ω en clases de equivalencia. La partición se llama clustering al nivel x.

Ejemplo. Consideremos $n = 5$ partidos políticos: CU (Conveniencia y Unión), PP (Partido Pragmático), PSC (Partido Social Catalán), IC (Iniciativa Catalana) y ER (Entente Republicana). Un ejemplo (hipotético) de jerarquía indexada sobre Ω ={CU,PP,PSC,IC,ER} es:

C ={CU_0,PP_0,PSC_0,IC_0,ERC_0,$\{\mathrm{CU},\mathrm{PP}\}_1$,$\{\mathrm{PSC},\mathrm{IC}\}_{1.5}$,$\{\mathrm{PSC},\mathrm{IC},\mathrm{ERC}\}_2$,$\Omega_3$}, donde el índice α está indicado como un subíndice: α(CU)=0, α(CU,PP)=1, etc. tenemos entonces tenemos las siguientes particiones o clusterings:

		α	Nombre del clustering
$\Omega =$	$\{\mathrm{CU}\} + \{\mathrm{PP}\} + \{\mathrm{PSC}\} + \{\mathrm{IC}\} + \{\mathrm{ER}\}$	0	(partidos)
$\Omega =$	$\{\mathrm{CU},\mathrm{PP}\} + \{\mathrm{PSC},\mathrm{IC}\} + \{\mathrm{ER}\}$	1.5	(derecha, izquierda, centro)
$\Omega =$	$\{\mathrm{CU},\mathrm{PP}\} + \{\mathrm{PSC},\mathrm{IC},\mathrm{ER}\}$	2	(coaliciones)
$\Omega =$	Ω	3	(parlamento)

La representación de esta clasificación se encuentra en la Figura 10.1, que justificamos en la sección siguiente.

10.3 Geometría ultramétrica

Para presentar una clasificación utilizamos llaves. Por ejemplo, la clasificación divisiva de Nación, Comunidades Autónomas y Provincias (sólo vamos a considerar 8) es:

Nación Autonomías Provincias

$$\text{España} \begin{cases} \text{Aragón} \begin{cases} \text{Huesca} \\ \text{Teruel} \\ \text{Zaragoza} \end{cases} \\ \text{Catalunya} \begin{cases} \text{Barcelona} \\ \text{Gerona} \\ \text{Lérida} \\ \text{Tarragona} \end{cases} \\ \text{Madrid} \quad \text{Madrid} \end{cases}$$

Una generalización de las llaves es el árbol ultramétrico. Como veremos más adelante, una jerarquía indexada puede ser visualizada mediante un gráfico sencillo e intuitivo, llamado dendograma.

Definició 10.3.1 *Un espacio ultramétrico* (Ω, u) *es una estructura formada por un conjunto finito* Ω *y una función distancia* u *sobre* $\Omega \times \Omega$ *verificando, para todo* i, j, k *de* Ω*:*

- *No negatividad:* $u(i,j) \geq u(i,i) = 0$.
- *Simetría:* $u(i,j) = u(j,i)$.
- *Propiedad ultramétrica:*

$$u(i,j) \leq \sup\{u(i,k), u(j,k)\}.$$

La matriz $U = (u(i,j))$ de orden $n \times n$

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nn} \end{pmatrix} \qquad u_{ij} = u_{ji} = u(i,j), \quad u_{ii} = 0.$$

es la matriz de distancias ultramétricas .

Proposició 10.3.1 *Una distancia ultramétrica verifica la desigualdad triangular y por lo tanto es métrica.*

Demost.:

$$u(i,j) \leq \sup\{u(i,k), u(j,k)\} \leq u(i,k) + u(j,k). \quad \square$$

Definició 10.3.2 *Un triángulo $\{i,j,k\}$ formado por tres elementos de Ω es ultramétrico si es isósceles y su base es el lado más pequeño. Es decir, si $u(i,j)$ es la base, entonces*

$$u(i,j) \leq u(i,k) = u(j,k).$$

Teorema 10.3.2 *En un espacio ultramétrico todo triángulo es ultramétrico.*

Demost.: Sea $\{i,j,k\}$ un triángulo. Sea $u(i,j)$ es el lado más pequeño, entonces:

$$\begin{array}{lll} u(i,k) & \leq \sup\{u(i,j), u(j,k)\} & = u(j,k) \\ u(j,k) & \leq \sup\{u(i,j), u(i,k)\} & = u(i,k) \end{array} \Longrightarrow u(i,k) = u(j,k). \quad \square$$

Definició 10.3.3 *Un árbol ultramétrico (también llamado dendograma) es un grafo conexo, sin ciclos con un punto llamado raiz y n puntos extremos equidistantes de la raiz.*

Una propiedad importante es que todo espacio ultramétrico (Ω, u) se puede "dibujar" mediante un dendograma, como en la Figura 10.2.

Teorema 10.3.3 *Sea (Ω, u) un espacio ultramétrico. Entonces podemos representarlo mediante un árbol ultramétrico con extremos los elementos de Ω.*

Demost.: Supongamos el árbol en posición vertical. Sea $u(i,j)$ la distancia entre los extremos i,j medida como la mitad de la mínima longitud de las aristas verticales que unen i con j, es decir, la distancia vertical hasta el nudo γ que liga i con j. Consideremos un triángulo $\{i,j,k\}$ y supongamos que $\{i,j\}$ es el lado más pequeño. Entonces k se relaciona con i,j en un nudo γ' por encima de γ. Así $u(k,i) = u(k,j) = u(i,j) + \beta$, donde $\beta \geq 0$ es la distancia vertical entre γ y γ'. Esto demuestra que $\{i,j,k\}$ es un arbol ultramétrico.$\square$

Hay una versión del Teorema 10.2.1 para distancias ultramétricas.

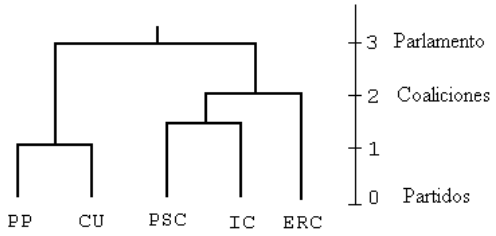

Figura 10.1: Representación en árbol ultramétrico (dendograma) de cinco partidos políticos.

Teorema 10.3.4 *Sea (Ω, u) un espacio métrico. Si u es distancia ultramétrica, entonces la relación binaria $\mathcal{R}_x$ sobre los elementos de Ω*

$$i\mathcal{R}_x j \quad si \quad u(i,j) \leq x, \tag{10.2}$$

es de equivalencia para todo $x \geq 0$. Recíprocamente, si la relación (10.2) es de equivalencia para todo $x \geq 0$, entonces u es distancia ultramétrica.

Demost.: Supongamos que u es ultramétrica. Entonces la relación $\mathcal{R}_x$ es:

Reflexiva: $u(i,i) = 0 \leq x$.

Simétrica: $u(i,j) = u(j,i) \leq x$.

Transitiva: Sea $\{i,j,k\}$ un triángulo ultramétrico con base $\{i,j\}$. entonces tenemos

$$u(i,j) \leq u(j,k) = u(i,k) \leq x,$$

que nos demuestra la transitividad.

Supongamos ahora que $\mathcal{R}_x$ es de equivalencia y que el triángulo $\{i,j,k\}$ verifica:

$$u(i,j) \leq u(j,k) \leq u(i,k).$$

Sea $x = u(j,k)$. Entonces $u(i,j) \leq x$, $u(j,k) \leq x \Rightarrow u(i,k) \leq x = u(j,k)$ por la transitividad de $\mathcal{R}_x$. Esto demuestra que $u(j,k) = u(i,k)$ y por lo tanto el triángulo $\{i,j,k\}$ es ultramétrico.□

Otra propiedad importante es que juntando elementos próximos de Ω seguimos manteniendo la propiedad ultramétrica, y esto vale para cualquier clustering.

Teorema 10.3.5 *Supongamos que sobre los m clusters del clustering*

$$\Omega = c_1 + c_2 + \ldots + c_m$$

hay definida una distancia ultramétrica u. Sean c_i, c_j los dos clusters más próximos: $u(c_i,c_j) =$ mínimo. Entonces uniendo c_i con c_j, se puede definir una distancia ultramétrica u' sobre los $m-1$ clusters del clustering

$$\Omega = c_1 + \ldots + c_i \cup c_j + \ldots + c_m.$$

Demost.: Si $k \neq i,j$, por la propiedad ultramétrica tenemos que $u(c_k,c_i) = u(c_k,c_j)$. Definimos:

$$\begin{array}{lll} u'(c_k, c_i \cup c_j) & = u(c_k,c_i) \quad = u(c_k,c_j), & k \neq i,j, \\ u'(c_a,c_b) & = u(c_a,c_b), & a,b \neq i,j. \end{array} \tag{10.3}$$

Consideremos el triángulo $\{c_a, c_b, c_i \cup c_j\}$. Entonces:

$$\begin{array}{ll} u'(c_a,c_b) & = u(c_a,c_b) \\ & \leq \sup\{u(c_a,c_i), u(c_b,c_i)\} = \sup\{u'(c_a, c_i \cup c_j), u'(c_b, c_i \cup c_j)\}, \\ u'(c_a, c_i \cup c_j) & = u(c_a,c_i) \\ & \leq \sup\{u(c_a,c_b), u(c_b,c_i)\} = \sup\{u'(c_a,c_b), u'(c_b, c_i \cup c_j)\}. \quad \square \end{array}$$

Finalmente, la propiedad ultramétrica es invariante por transformaciones monótonas.

Proposició 10.3.6 *Si u es distancia ultramétrica y $u' = \varphi(u)$ es una transformación de u donde φ es una función positiva monótona (creciente o decreciente), entonces u' es también distancia ultramétrica.*

Demost.: Si $\{i,j,k\}$ es un triángulo ultramétrico con base $\{i,j\}$ y φ es monótona, tendremos que

$$u(i,j) \leq u(i,k) = u(j,k) \Rightarrow u'(i,j) \leq u'(i,k) = u'(j,k). \quad \square$$

10.4 Algoritmo fundamental de clasificación

A partir de un espacio ultramétrico podemos construir una jerarquia indexada. Nos lo permite el siguiente

Algoritmo fundamental de clasificación

Sea (Ω, u) un espacio ultramétrico. El fundamento de este algoritmo consiste en el hecho de que, en virtud del Teorema 10.3.5, juntando elementos o clusters más próximos, conservamos la propiedad ultramétrica.

1. Comencemos con la partición:
$$\Omega = \{1\} + ... + \{n\}.$$

2. Sean i, j los dos elementos más próximos: $u(i, j) =$ mínimo. Los unimos
$$\{i\} \cup \{j\} = \{i, j\}$$
y definimos la nueva distancia ultramétrica u'
$$u'(k, \{i, j\}) = u(i, k) = u(j, k), \quad k \neq i, j,$$
(ver Teorema 10.3.5).

3. Consideremos la nueva partición:
$$\Omega = \{1\} + ... + \{i, j\} + \ldots + \{n\}$$
y repitamos el paso 2 hasta llegar a Ω. En este proceso, cada vez que unimos c_i con c_j tal que $u(c_i, c_j) =$ mínimo, definimos el índice
$$\alpha(c_i \cup c_j) = u(c_i, c_j). \tag{10.4}$$

El resultado de este proceso es una jerarquía indexada (C, α).

10.5 Equivalencia entre jerarquía indexada y ultramétrica

Una jerarquía indexada es una estructura conjuntista. Un espacio ultramétrico es una estructura geométrica. Ambas estructuras son equivalentes.

Teorema 10.5.1 *Sea (C, α) una jerarquía indexada total sobre un conjunto Ω. Entonces podemos definir una distancia ultramétrica u sobre Ω. Recíprocamente, todo espacio ultramétrico (Ω, u) define una jerarquía indexada (C, α).*

Demost.: A partir de (C, α) definimos la siguiente distancia

$$u(i, j) = \alpha(c_{ij}),$$

donde c_{ij} es el mínimo cluster (respecto a la relación de inclusión) que contiene i, j. Sea $\{i, j, k\}$ un triángulo y sean también c_{ik}, c_{jk} los mínimos clusters que contienen $\{i, k\}$, $\{j, k\}$ respectivamente. Tenemos que

$$c_{ik} \cap c_{jk} \neq \emptyset$$

y por tanto (axioma de la intersección) hay dos posibilidades:

$$\begin{array}{ll} a) & c_{ik} \subset c_{jk} \Rightarrow i, j, k \in c_{jk} \Rightarrow c_{ij} \subset c_{jk} \Rightarrow u(i, j) = \alpha(c_{ij}) \leq u(j, k) = \alpha(c_{jk}) \\ b) & c_{jk} \subset c_{ik} \Rightarrow i, j, k \in c_{ik} \Rightarrow c_{ij} \subset c_{ik} \Rightarrow u(i, j) = \alpha(c_{ij}) \leq u(i, k) = \alpha(c_{ik}) \end{array}$$

Así pues: $u(i, j) \leq \sup\{u(i, k), u(j, k)\}$.

La posibilidad de construir una jerarquía indexada a partir de una distancia ultramétrica es una consecuencia del algoritmo fundamental de clasificación. El índice de la jerarquía viene dado por (10.4).□

Comentarios:

1. Observa la analogía entre el Teorema 10.3.5 y el algoritmo fundamental de clasificación.

2. Observa además que (10.3) permite definir de manera inequívoca una distancia entre un cluster y la unión de los dos clusters más próximos. Esta propiedad es la que otorga importancia a la distancia ultramétrica.

10.6 Algoritmos de clasificación jerárquica

Supongamos que, en relación a unas variables observables, hemos obtenido una matriz de distancias $\Delta = (\delta(i, j))$ de orden $n \times n$ entre los elementos de un conjunto Ω :

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{pmatrix} \qquad \delta_{ij} = \delta_{ji} = \delta(i, j), \quad \delta_{ii} = 0.$$

Si la distancia δ es ultramétrica, entonces no hay ningún problema para llevar a cabo una clasificación construyendo una jerarquía indexada. Basta con aplicar el algoritmo fundamental de clasificación (Sección 10.4). Pero en general δ no cumple la propiedad ultramétrica y por lo tanto hemos de modificar adecuadamente este algoritmo.

Algoritmo de clasificación

Sea (Ω, δ) un espacio métrico. El algoritmo de clasificación se basa en el Teorema 10.3.5, en el sentido de que juntaremos los elementos o clusters más próximos, y procuraremos obtener triángulos ultramétricos.

1. Comencemos con la partición:

$$\Omega = \{1\} + ... + \{n\}.$$

2. Sean i, j los dos elementos más próximos: $\delta(i, j) =$ mínimo. Los unimos

$$\{i\} \cup \{j\} = \{i, j\}$$

y definimos la distancia de un elemento k al cluster $\{i, j\}$

$$\delta'(k, \{i, j\}) = f(\delta(i, k), \delta(j, k)), \quad k \neq i, j, \tag{10.5}$$

donde f es una función adecuada.

3. Consideremos la nueva partición:

$$\Omega = \{1\} + ... + \{i, j\} + \ldots + \{n\},$$

y repitamos el paso 2 hasta llegar a Ω. En este proceso, cada vez que unimos c_i con c_j tal que $\delta(c_i, c_j) =$ mínimo, definimos el índice

$$\alpha(c_i \cup c_j) = \delta'(c_i, c_j). \tag{10.6}$$

La función f en (10.5) se define adecuadamente a fin de que se cumpla la propiedad ultramétrica. El resultado de este proceso es una jerarquía indexada (C, α).

10.6.1 Método del mínimo

Los diferentes métodos de clasificación jerárquica dependen de la elección de f en (10.5). Una primera elección conveniente de f consiste simplemente en tomar el valor más *pequeño* de los dos lados $\{i,k\},\{j,k\}$ del triángulo $\{i,j,k\}$ con base $\{i,j\}$, es decir:

$$\delta'(k,\{i,j\}) = \min\{\delta(i,k),\delta(j,k)\}, \quad k \neq i,j. \tag{10.7}$$

En otras palabras, hacemos que el triángulo

$$\delta(i,j\} \leq \delta(i,k) = a \leq \delta(j,k),$$

se transforme en ultramétrico

$$\delta'(i,j\} \leq \delta'(i,k) = \delta'(j,k) = a.$$

Ejemplo. Sea Δ una matriz de distancias sobre $\Omega = \{1,2,3,4,5\}$. El método del mínimo proporciona una jerarquía indexada (C,α) asociada a una matriz ultramétrica $\underline{U}$:

$$\Delta = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & 1 & 3 & 4 & 7 \\ 2 & & 0 & 4 & 4 & 8 \\ 3 & & & 0 & 2 & 8 \\ 4 & & & & 0 & 7 \\ 5 & & & & & 0 \end{array} \rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & 3 & 4 & 7 \\ 3 & & 0 & 2 & 8 \\ 4 & & & 0 & 7 \\ 5 & & & & 0 \end{array} \rightarrow \begin{array}{c|ccc} & (1,2) & (3,4) & 5 \\ (1,2) & 0 & 3 & 7 \\ (3,4) & & 0 & 7 \\ 5 & & & 0 \end{array} \rightarrow$$

$$\begin{array}{c|cc} & (1,2,3,4) & 5 \\ (1,2,3,4) & 0 & 7 \\ 5 & & 0 \end{array} \rightarrow C = \{\{1\}_0,\ldots,\{5\}_0,\{1,2\}_1,\{3,4\}_2,\{1,2,3,4\}_3,\Omega_7\}$$

$$(C,\alpha) \longleftrightarrow \underline{U} = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & 1 & 3 & 3 & 7 \\ 2 & & 0 & 3 & 3 & 7 \\ 3 & & & 0 & 2 & 7 \\ 4 & & & & 0 & 7 \\ 5 & & & & & 0 \end{array}$$

El método del mínimo produce una distancia ultramétrica $\underline{u}$ que goza de la siguiente propiedad.

Teorema 10.6.1 *Sea*

$$\underline{\mathbf{U}} = \{u \mid u \text{ es ultramétrica, } u(i,j) \leq \delta(i,j)\}$$

el conjunto de distancias ultramétricas más pequeñas que δ. Entonces la distancia ultramétrica $\underline{u}$ resultante del método del mínimo es el elemento máximo de $\underline{\mathbf{U}}$

$$\underline{u}(i,j) \geq u(i,j), \quad u \in \underline{\mathbf{U}}, \quad \forall i,j \in \Omega.$$

Demost.: Sean $\{i,j\}$ los elementos más próximos. Entonces $\underline{u}(i,j) = \delta(i,j)$. La columna k $(\neq i,j)$ tendrá términos repetidos iguales a una distancia δ' construida tomando un mínimo. Si $u \leq \delta$ es otra distancia ultramétrica, entonces: a) si es estrictamente más pequeña es evidente que $\underline{u} > u$. b) si $u(k',k'')$ es más grande que $\underline{u}(k',k'')$ pero es igual a alguna δ, entonces la columna k tendrá elementos repetidos, y al menos uno será superior a δ'. Contradicción.

El razonamiento es parecido si consideramos un cluster c y un elemento $k \notin c$. Compárese Δ con $\underline{U}$ en el ejemplo anterior. Véase también el Teorema 10.7.3.

A la vista de este resultado, podemos decir que $\underline{u}$ es la mejor aproximación a δ por defecto.

10.6.2 Método del máximo

Una segunda elección razonable de f consiste en tomar el valor más *grande* de los dos lados $\{i,k\}, \{j,k\}$ del triángulo $\{i,j,k\}$ con base $\{i,j\}$, es decir:

$$\delta'(k,\{i,j\}) = \max\{\delta(i,k), \delta(j,k)\}, \quad k \neq i,j. \tag{10.8}$$

En otras palabras, hacemos que el triángulo

$$\delta(i,j\} \leq \delta(i,k) \leq \delta(j,k) = b,$$

se convierta en ultramétrico

$$\delta'(i,j\} \leq \delta'(i,k) = \delta'(j,k) = b.$$

El método del máximo produce una distancia ultramétrica $\overline{u}$ que goza de la siguiente propiedad.

Teorema 10.6.2 *Sea*

$$\overline{\mathbf{U}} = \{u \mid u \text{ es ultramétrica, } u(i,j) \geq \delta(i,j)\}$$

el conjunto de distancias ultramétricas más grandes que δ. Entonces la distancia ultramétrica $\overline{u}$ resultante del método del máximo es un elemento minimal de $\overline{\mathbf{U}}$

$$\overline{u}(i,j) \leq u(i,j), \quad u \in \overline{\mathbf{U}}, \quad \forall i,j \in \Omega.$$

Así $\overline{u}$ es la mejor aproximación a δ por exceso.

Comentarios:

1. Las distancias $\underline{u}, \overline{u}$, y δ verifican:

   $$\underline{u}(i,j) \leq \delta(i,j) \leq \overline{u}(i,j).$$

   Hay igualdad $\underline{u} = \delta = \overline{u}$ si y sólo si δ es ultramétrica.

2. $\underline{u}$ es elemento máximo y es único. El método del mínimo sólo tiene una solución.

3. $\overline{u}$ es elemento minimal y no es único. El método del máximo puede tener varias soluciones.

4. Si todos los elementos fuera de la diagonal de la matriz de distancias Δ son diferentes, entonces la solución aplicando el método del máximo es única y por tanto $\overline{u}$ es elemento mínimo .

Finalmente, una notable propiedad de los métodos del mínimo (también conocido como *single linkage*) y del máximo (*complete linkage*) es que conservan la ordenación de la distancia δ, en el sentido de la Proposición 10.3.6.

Teorema 10.6.3 *Los métodos del mínimo y del máximo son invariantes por transformaciones monótonas de la distancia δ :*

$$\delta' = \varphi(\delta) \Rightarrow u' = \varphi(u)$$

donde u, u' son las ultramétricas asociadas a δ, δ' y φ es una función monótona positiva.

Demost.: En el proceso de encontar la ultramétrica sólo intervienen los rangos de los valores de δ, que son los mismos que los rangos de los valores de δ'.

10.7 Otras propiedades del método del mínimo

Una propiedad de la distancia ultramétrica dice que todo elemento de una bola es también centro de la propia bola.

Proposició 10.7.1 *Sea $B(i_0, r)$ una bola cerrada de centro i_0 y radio r :*

$$B(i_0, r) = \{i \in \Omega | \quad u(i_0, i) \leq r\}.$$

Entonces

$$\forall i \in B(i_0, r) \quad verifica \quad B(i, r) = B(i_0, r).$$

La demostración es inmediata. También se verifica:

Proposició 10.7.2 *Sea $\{i_1, \ldots, i_m\}$. Se cumple la desigualdad*

$$u(i_1, i_m) \leq \sup\{u(i_\alpha, i_{\alpha+1}) | \alpha = 1, \ldots, m-1\}.$$

Demost.: Por recurrencia sobre m. Para $m = 2$ es la desigualdad ultramétrica. Supongamos cierto para $m - 1$. Tenemos:

$$\begin{aligned} u(i_1, i_m) &\leq \sup\{u(i_1, i_{m-1}), u(i_{m-1}, i_m)\} \\ &\leq \sup\{\sup\{u(i_\alpha, i_{\alpha+1}) | \alpha = 1, \ldots, m-2\}, u(i_{m-1}, i_m)\} \\ &\leq \sup\{u(i_\alpha, i_{\alpha+1}) | \alpha = 1, \ldots, m-1\}. \square \end{aligned}$$

Sea ahora $\Omega = \{1, 2, \ldots, n\}$ y δ una distancia sobre Ω.

Definició 10.7.1 *Una cadena $[i, j]_m$ es el conjunto $\{i = i_1, i_2, \ldots, j = i_m\}$.*

Definició 10.7.2 *Indiquemos*

$$\sup[i, j]_m = \sup_{1 \leq \alpha \leq m} \delta(i_\alpha, i_{\alpha+1})$$

el máximo salto de la cadena $[i, j]_m$. Definimos la distancia sobre Ω

$$\underline{u}(i, j) = \inf_m \sup[i, j]_m$$

Teorema 10.7.3 *Se verifica:*

1. *$\underline{u}$ es una ultramétrica tal que $\underline{u} \leq \delta$.*

2. *Si u es otra ultramétrica tal que $u \leq \delta$ entonces $u \leq \underline{u}$.*

3. *$\underline{u}$ es la ultramétrica que se obtiene por el método del mínimo.*

Demost.: $[i,j]_2 = \{i,j\}$ es una cadena que une i,j y por lo tanto

$$\underline{u}(i,j) \leq \sup[i,j]_2$$

Sea $[i,j,k]$ una cadena que une i,j pero que contiene k. El conjunto de las cadenas $[i,j,k]$ está contenido en el conjunto de las cadenas $[i,j]$. Por lo tanto:

$$\inf_m \sup[i,j]_m \leq \inf_{m'} \sup[i,k,j]_{m'} \tag{10.9}$$

Por otra parte, dadas las cadenas $[i,j],[j,k]$ podemos construir

$$[i,k,j] = [i,j] \cup [j,k]$$

de modo que

$$\sup[i,k,j] = \sup\{\sup[i,j], \sup[j,k]\}$$

Teniendo en cuenta (10.9) deducimos que

$$\underline{u}(i,j) \leq \sup\{\underline{u}(i,k), \underline{u}(j,k)\}$$

Sea ahora $u \leq \delta$. Aplicando la Proposición 10.7.2

$$u(i,j) \leq \sup_{1 \leq \alpha \leq m} u(i_\alpha, i_{\alpha+1}) \leq \sup[i,j]_m$$

Por lo tanto

$$u(i,j) \leq \inf_m \sup[i,j]_m = \underline{u}(i,j).$$

Conviene comparar este resultado con el Teorema 10.6.1.

10.8 Un ejemplo

Un grupo de $n = 11$ profesores de probabilidades y estadística de la Universidad de Barcelona han publicado, entre 1994 y 2000, unos 150 artículos

internacionales, algunos en colaboración. Con la finalidad de agrupar los profesores según los artículos que publicaron juntos, consideramos el coeficiente de similaridad

$$s(i,j) = \text{número de artículos que } i,j \text{ han publicado juntos.}$$

Definimos entonces la distancia

$$d(i,j) = 1 - s(i,j)/\min\{s(i,i), s(j,j)\}.$$

Obtenemos la matriz de distancias:

	Are	Cor	Cua	For	Mar	Nua	Oli	Oll	Rov	San	Sar
Arenas	0										
Corcuera	1	0									
Cuadras	0.50	1	0								
Fortiana	0.83	1	0.06	0							
Marquez	1	1	1	1	0						
Nualart	1	1	1	1	1	0					
Oliva	1	1	0.33	0.33	1	1	0				
Oller	1	0.75	1	1	1	1	1	0			
Rovira	1	1	1	1	1	1	1	1	0		
Sanz	1	1	1	1	0.33	0.93	1	1	0.11	0	
Sarra	1	1	1	1	0.75	1	1	1	1	0.25	0

Aplicando un análisis cluster, método del mínimo, a esta matriz, obtenemos el dendograma de la Figura 10.2. Este dendograma pone de manifiesto que hay tres grupos principales con 4, 2 y 5 profesores, que trabajan en análisis multivariante (AM), estadística matemática (EM) y análisis estocástico (AE), respectivamente.

10.9 Clasificación no jerárquica

Una clasificación no jerárquica de n objetos en relación a una matriz de datos cuantitativos $\mathbf{X}$, consiste en obtener g grupos homogéneos y excluyentes (clusters). Si tenemos g clusters, estamos en la misma situación contemplada en el Cap. 7, y podemos considerar la descomposición de la variabilidad total

$$\mathbf{T} = \mathbf{B} + \mathbf{W}$$

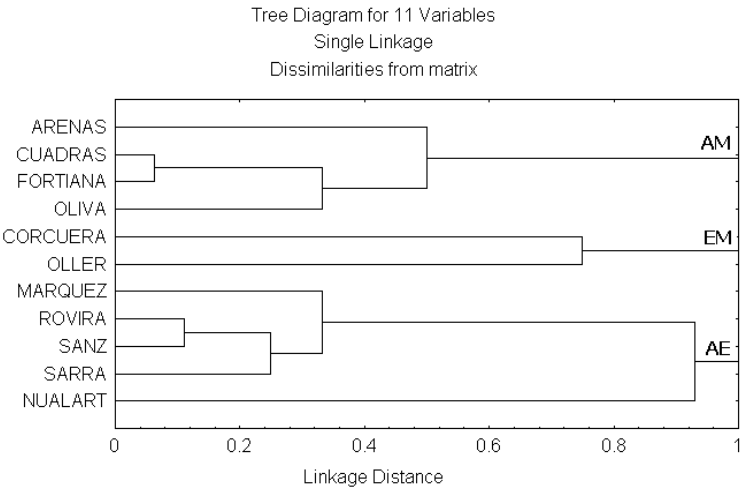

Figura 10.2: Representación que agrupa 11 profesores según los artículos publicados conjuntamente.

Una partición en g clusters que hace máxima $\mathbf{B}$ o mínima $\mathbf{W}$, en relación a algún criterio, dará una solución al problema, puesto que tendremos una máxima dispersión entre clusters. Algunos criterios, justificados por el análisis multivariante de la varianza, son:

a) Minimizar $\text{tr}(\mathbf{W})$

b) Minimizar $|\mathbf{W}|$.

c) Minimizar $\Lambda = |\mathbf{W}|/|\mathbf{T}|$.

d) Maximizar $\text{tr}(\mathbf{W}^{-1}\mathbf{B})$.

Pero la cantidad de maneras diferentes de agrupar n objetos en g clusters es del orden de $g^n/g!$, número muy grande incluso para valores moderados de n y g (necesitaríamos formar más de 10^{23} clusters si $n = 50, g = 3$). Por tanto, es necesario seguir algún algoritmo de agrupación.

El método de las *medias móviles* consiste en:

1. Comenzar con g puntos del espacio R^p y asignar los objetos a g clusters de acuerdo con la proximidad (distancia euclídea) a los g puntos iniciales.

2. Calcular los centroides de los g clusters obtenidos y reasignar los objetos según su proximidad al centroide de cada cluster.

3. Repetir el paso anterior, calculando cada vez la cantidad $|\mathbf{W}|$ (o el criterio de optimización escogido). Parar cuando $|\mathbf{W}|$ ya no disminuye.

Es posible probar que la suma de cuadrados de las distancias euclídeas de los puntos de cada cluster al centroide

$$\sum_{k=1}^{g}\sum_{i=1}^{n} d^2(\mathbf{x}_{ki}, \overline{\mathbf{x}}_k)$$

disminuye a cada paso.

10.10 Número de clusters

Diversos autores (Calinski, Harabasz, Hartigan, Krzanowski, Lai) han propuesto métodos para estimar el número de clusters de una clasificación. Es éste un tema abordado desde muchas perspectivas (véase Gordon, 1999).

Normalmente el usuario determina el número k de clusters. Un primer criterio consiste en tomar el valor k tal que maximice la cantidad

$$\mathrm{cl}_1(k) = \frac{\mathrm{tr}(\mathbf{B}(k))}{g-1} / \frac{\mathrm{tr}(\mathbf{W}(k))}{n-g},$$

donde $\mathbf{B}(k), \mathbf{W}(k)$ indican las matrices entre-grupos y dentro-grupos para k grupos. Otro criterio considera

$$\mathrm{dif}(k) = (k-1)^{2/p}\mathbf{W}(k-1) - k^{2/p}\mathbf{W}(k)$$

y elige k tal que maximiza

$$cl_2(k) = dif(k)/dif(k+1).$$

Pero cl_1 i cl_2 no estan definidos para $k = 1$. Un tercer criterio propone el estadístico

$$H(k) = (\frac{\mathbf{W}(k)}{\mathbf{W}(k+1)} - 1)/(n-k-1),$$

empieza con $k = 1$ y aumenta k si $H(k)$ crece significativamente de acuerdo con una aproximación a la distribución F.

Tibshirani *et al.* (2001) proponen un método que contempla también el caso $k = 1$. Partiendo del resultado de cualquier clasificación, jerárquica o no, comparan el cambio de $\mathbf{W}(k)$ respecto al cambio esperado para a una distribución apropiada de referencia

$$E(\log|\mathbf{W}(k)|) - \log|\mathbf{W}(k)|.$$

10.11 Complementos

La historia de la clasificación comienza con la sistemática de Carl von Linné, que permitía clasificar animales y plantas según género y especie. La clasificación moderna (denominada taxonomía numérica) se inicia en 1957 con la necesidad de proponer criterios objetivos de clasificación (Sokal, Sneath, Michener). Posteriormente, diversos autores relacionaron las clasificaciones jerárquicas con los espacios ultramétricos (Benzecri, Jardine, Sibson, Johnson), dado que la propiedad ultramétrica ya era conocida en otros campos de la matemática.

Una crítica que se ha hecho al análisis cluster es el excesivo repertorio de distancias y métodos de clasificación. Incluso se han realizado clasificaciones de las propias maneras de clasificar, y clasificaciones jerárquicas de las distancias. También se ha argumentado (Flury, 1997) que el planteamiento correcto del análisis cluster consiste en encontrar mixturas

$$f(\mathbf{x}) = p_1 f_1(\mathbf{x}) + \ldots + p_g f_g(\mathbf{x}),$$

donde cada densidad f_i representaría un cluster y f la densidad de los datos que hemos observado. Pero si una distancia mide razonablemente las diferencias entre los objetos, entonces se pueden obtener clasificaciones objetivas aplicando análisis cluster jerárquico. Por ejemplo, en el año 1999 se realizó la clasificación jerárquica del reino vegetal a partir de distancias entre secuencias de DNA, obteniendo una concordancia de un 60% con la clasificación tradicional basada en la similitud morfológica de las plantas.

J. C. Gower conjeturó y Holman (1972) probó, que toda distancia ultramétrica era euclídea con dimensión $n - 1$. Entonces interesó estudiar la relación entre representaciones en árbol y en coordenadas (Bock, Crithcley, Heiser, Kruskal). Critchley y Heiser (1988) probaron que, a pesar del resultado de Holman, es posible representar un espacio ultramétrico con una sola

dimensión utilizando una métrica adecuada. Un estudio de los vectores propios y las dimensiones principales de una matriz de distancias ultramétricas es debido a Cuadras y Oller (1987). Ver Cuadras *et al.* (1996).

N. Jardine y R. Simpson propusieron el método de clasificación denominado flexible, que consiste en definir la distancia de un cluster a la unión de dos clusters en función de unos parámetros, por ejemplo, inicialmente

$$\delta'(k,\{i,j\}) = \alpha_i\delta(i,k) + \alpha_j\delta(j,k) + \beta\delta(i,j) + \gamma|\delta(i,k) - \delta(j,k)|,$$

y análogamente en los siguientes pasos. Dando valores a los parámetros se obtienen los métodos siguientes (se incluye denominación estándar):

Criterio de agrupación	α_i	α_j	β	γ
Mínimo (single linkage)	$1/2$	$1/2$	0	$-1/2$
Máximo (complete linkage)	$1/2$	$1/2$	0	$+1/2$
Media (weighted average link)	$1/2$	$1/2$	0	0
UPGMA (group average link)	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	0	0

UPGMA (Unweighted pair group method using arithmetic averages) es un método recomendable porque proporciona una clasificación que se ajusta bien a la distancia inicial en el sentido de los mínimos cuadrados.

G.H. Ball, D.J. Hall, E. Diday y otros propusieron algoritmos eficientes de agrupación no jerárquica. Consúltese Everitt (1993).

Capítulo 11

ANALISIS DISCRIMINANTE

11.1 Introducción

Sean Ω_1, Ω_2 dos poblaciones, $X_1, ...,X_p$ variables observables, $\mathbf{x} = (x_1, ..., x_p)$ las observaciones de las variables sobre un individuo ω. El problema es asignar ω a una de las dos poblaciones. Este problema aparece en muchas situaciones: decidir si se puede conceder un crédito; determinar si un tumor es benigno o maligno; identificar la especie a que pertenece una planta.

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

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

Esta regla divide R^p en dos regiones

$$R_1 = \{\mathbf{x} | D(\mathbf{x}) > 0\}, \quad R_2 = \{\mathbf{x} | D(\mathbf{x}) < 0\}.$$

En la decisión de clasificar, nos equivocaremos si asignamos ω a una población a la que no pertenece. La probabilidad de clasificación errónea (pce) es

$$\text{pce} = P(R_2/\Omega_1)P(\Omega_1) + P(R_1/\Omega_2)P(\Omega_2). \tag{11.1}$$

11.2 Clasificación en dos poblaciones

11.2.1 Discriminador lineal

Sean μ_1, μ_2 los vectoros de medias de las variables en Ω_1, Ω_2, respectivamente, y supongamos que la matriz de covarianzas Σ es común. Las distancias de Mahalanobis de las observaciones $\mathbf{x} = (x_1, \ldots, x_p)'$ de un individuo ω a las poblaciones son

$$M^2(\mathbf{x},\mu_i) = (\mathbf{x}-\mu_i)'\Sigma^{-1}(\mathbf{x}-\mu_i), \quad i = 1, 2.$$

Un primer criterio de clasificación consiste en asignar ω a la población más próxima:

$$\begin{array}{rl} \text{Si } M^2(\mathbf{x},\mu_1) < M^2(\mathbf{x},\mu_2) & \text{asignamos } \omega \text{ a } \Omega_1, \\ \text{en caso contrario} & \text{asignamos } \omega \text{ a } \Omega_2. \end{array} \tag{11.2}$$

Expresando esta regla como una función discriminante, tenemos:

$$\begin{array}{rcl} M^2(\mathbf{x},\mu_2) - M^2(\mathbf{x},\mu_1) & = & \mathbf{x}'\Sigma^{-1}\mathbf{x} + \mu_2\Sigma^{-1}\mu_2 - 2\mathbf{x}'\Sigma^{-1}\mu_2 \\ & & -\mathbf{x}'\Sigma^{-1}\mathbf{x} - \mu_1\Sigma^{-1}\mu_1 + 2\mathbf{x}'\Sigma^{-1}\mu_1 \\ & = & (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 + \mu_1) + 2\mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_2) \end{array}$$

Definimos la función discriminante

$$L(\mathbf{x}) = \left[\mathbf{x} - \frac{1}{2}(\mu_1 + \mu_2)\right]' \Sigma^{-1}(\mu_1 - \mu_2). \tag{11.3}$$

Tenemos que

$$M^2(\mathbf{x},\mu_2) - M^2(\mathbf{x},\mu_1) = 2L(\mathbf{x}) - L((\mu_1 + \mu_2)/2)$$

y la regla (11.2) es

$$\begin{array}{rl} \text{Si } L(\mathbf{x}) > 0 & \text{asignamos } \omega \text{ a } \Omega_1, \\ \text{en caso contrario} & \text{asignamos } \omega \text{ a } \Omega_2. \end{array}$$

La función lineal (11.3) es el *discriminador lineal* de Fisher.

11.2.2 Regla de la máxima verosimilitud

Supongamos que $f_1(\mathbf{x}), f_2(\mathbf{x})$ son las densidades de $\mathbf{x}$ en Ω_1, Ω_2. Una regla de clasificación consiste en asignar ω a la población donde la verosimilitud de las observaciones $\mathbf{x}$ es más grande:

$$\begin{array}{rl} \text{Si } f_1(\mathbf{x}) > f_2(\mathbf{x}) & \text{asignamos } \omega \text{ a } \Omega_1, \\ \text{en caso contrario} & \text{asignamos } \omega \text{ a } \Omega_2. \end{array}$$

La función discriminante es

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

11.2.3 Regla de Bayes

En ciertas situaciones, se conocen las probabilidades a priori de que ω pertenezca a cada una de las poblaciones

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

Una vez que se dispone de las observaciones $\mathbf{x} = (x_1, \ldots, x_p)$, las probabilidades a posteriori de que ω pertenezca a las poblaciones (teorema de Bayes) son

$$P(\Omega_i/\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{q_1 f_1(\mathbf{x}) + q_2 f_2(\mathbf{x})}, \quad i = 1, 2.$$

La regla de clasificación de Bayes es

$$\begin{array}{rl} \text{Si } P(\Omega_1/\mathbf{x}) > P(\Omega_2/\mathbf{x}) & \text{asignamos } \omega \text{ a } \Omega_1, \\ \text{en caso contrario} & \text{asignamos } \omega \text{ a } \Omega_2. \end{array}$$

El discriminador de Bayes es

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

Cuando $q_1 = q_2 = 1/2$, entonces $B(\mathbf{x}) = V(\mathbf{x})$. Este discriminador es óptimo.

Teorema 11.2.1 *La regla de Bayes minimiza la probabilidad de clasificación errónea.*

Demost.: Supongamos que se dispone de otra regla que clasifica a Ω_1 si $\mathbf{x} \in R_1^*$, y a Ω_2 si $\mathbf{x} \in R_2^*$, donde R_1^*, R_2^* son regiones complementarias del espacio muestral. Indicando $d\mathbf{x} = dx_1 \cdots dx_p$. La probabilidad de clasificación errónea es

$$\begin{aligned} pce^* &= q_1 \int_{R_2^*} f_1(\mathbf{x}) d\mathbf{x} + q_2 \int_{R_1^*} f_2(\mathbf{x}) d\mathbf{x} \\ &= \int_{R_2^*} (q_1 f_1(\mathbf{x}) - q_2 f_{\mathbf{2}}(\mathbf{x})) d\mathbf{x} + q_2 (\int_{R_2} f_2(\mathbf{x}) d\mathbf{x} + \int_{R_1^*} f_2(\mathbf{x}) d\mathbf{x}) \\ &= \int_{R_2^*} (q_1 f_1(\mathbf{x}) - q_2 f_{\mathbf{2}}(\mathbf{x})) d\mathbf{x} + q_2. \end{aligned}$$

Esta última integral es mínima si R_2^* incluye todas las $\mathbf{x}$ tal que $q_1 f_1(\mathbf{x}) - q_2 f_{\mathbf{2}}(\mathbf{x}) < 0$ y excluye toda las $\mathbf{x}$ tal que $q_1 f_1(\mathbf{x}) - q_2 f_{\mathbf{2}}(\mathbf{x}) > 0$. Por tanto pce^* es mínima si $R_2^* = R_2$, donde $R_2 = \{\mathbf{x} | B(\mathbf{x}) < 0\}$.□

11.3 Clasificación en poblaciones normales

Supongamos ahora que la distribución de $X_1, ..., X_p$ en Ω_1 es $N_p(\mu_1, \Sigma_1)$ y en Ω_2 es $N_p(\mu_2, \Sigma_2)$, es decir,

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

11.3.1 Clasificador lineal

Si suponemos $\mu_1 \neq \mu_2, \Sigma_1 = \Sigma_2 = \Sigma$, entonces

$$\begin{aligned} V(\mathbf{x}) &= -\tfrac{1}{2}(\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1) + \tfrac{1}{2} (\mathbf{x} - \mu_2)' \Sigma^{-1} (\mathbf{x} - \mu_2) \\ &= L(\mathbf{x}) \end{aligned}$$

y por tanto los discriminadores máximo verosímil y lineal, el segundo basado en el criterio de la mínima distancia, coinciden.

Sea α la distancia de Mahalanobis entre las dos poblaciones

$$\alpha = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2).$$

Si suponemos que $\mathbf{x}$ proviene de $N_p(\mu_2, \Sigma)$, de $\mathbf{x} - \mu_1 = \mathbf{x} - \mu_2 + \mu_2 - \mu_1$, y de $E(\mathbf{x} - \mu_2)(\mathbf{x} - \mu_2)' = \Sigma$, $(\mathbf{x} - \mu_2)' \Sigma^{-1} (\mathbf{x} - \mu_2) \sim \chi_p^2$, tenemos que la esperanza de $U = (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1)$ es

$$E(U) = E[(\mathbf{x} - \mu_2)' \Sigma^{-1} (\mathbf{x} - \mu_2) + \alpha + 2(\mathbf{x} - \mu_2)' \Sigma^{-1} (\mu_2 - \mu_1)] = p + \alpha,$$

y la varianza de $V = (\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2)$ es la misma que la de $L(\mathbf{x})$ y es

$$\text{var}(V) = E((\mu_2 - \mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_2)(\mathbf{x}-\mu_2)'\Sigma^{-1}(\mu_2 - \mu_1)) = \alpha.$$

Entonces encontramos fácilmente la distribución de la función discriminante $L(\mathbf{x})$:

$$\begin{array}{l} L(\mathbf{x}) \text{ es } N(+\frac{1}{2}\alpha, \alpha) \text{ si } \mathbf{x} \text{ proviene de } N_p(\mu_1, \Sigma), \\ L(\mathbf{x}) \text{ es } N(-\frac{1}{2}\alpha, \alpha) \text{ si } \mathbf{x} \text{ proviene de } N_p(\mu_2, \Sigma). \end{array} \tag{11.4}$$

11.3.2 Regla de Bayes

Si suponemos $\mu_1 \neq \mu_2, \Sigma_1 = \Sigma_2 = \Sigma$, y conocemos las probabilidades a priori $q_1 = P(\Omega_1), q_2 = P(\Omega_2)$, entonces es fácil ver que

$$B(\mathbf{x}) = L(\mathbf{x}) + \log(q_1/q_2),$$

y la función discriminante de Bayes es el discriminador lineal más la constante $\log(q_1/q_2)$.

11.3.3 Probabilidad de clasificación errónea

La probabilidad de asignar $\mathbf{x}$ a Ω_2 cuando proviene de $N_p(\mu_1, \Sigma)$ es

$$P(L(\mathbf{x}) < 0|\Omega_1) = P((L(\mathbf{x}) - \frac{1}{2}\alpha)/\sqrt{\alpha}) = \Phi(-\frac{1}{2}\sqrt{\alpha}),$$

donde $\Phi(z)$ es la función de distribución $N(0,1)$. La probabilidad de clasificación errónea es

$$pce = q_1 P(L(\mathbf{x}) < 0|\Omega_1) + q_2 P(L(\mathbf{x}) > 0|\Omega_2) = \Phi(-\frac{1}{2}\sqrt{\alpha}).$$

Por tanto *pce* es una función decreciente de la distancia de Mahalanobis α entre las dos poblaciones.

11.3.4 Discriminador cuadrático

Supongamos $\mu_1 \neq \mu_2, \Sigma_1 \neq \Sigma_2$. Entonces el criterio de la máxima verosimilitud proporciona el discriminador

$$\begin{array}{rl} Q(\mathbf{x}) & = \frac{1}{2}\mathbf{x}'\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)\mathbf{x} + \mathbf{x}'\left(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2\right) \\ & + \frac{1}{2}\mu_2'\Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1'\Sigma_1^{-1}\mu_1 + \frac{1}{2}\log|\Sigma_2| - \frac{1}{2}\log|\Sigma_1| \end{array}$$

$Q(\mathbf{x})$ es el *discriminador cuadrático*. Análogamente podemos obtener el discriminador cuadrático de Bayes

$$B(\mathbf{x}) = Q(\mathbf{x}) + \log(q_1/q_2).$$

11.3.5 Clasificación cuando los parámetros son estimados

En las aplicaciones prácticas, $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ son desconocidos y se deberán estimar a partir de muestras de tamaños n_1, n_2 de las dos poblaciones sustituyendo μ_1, μ_2 por los vectores de medias $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2$, y Σ_1, Σ_2 por las matrices de covarianzas $\mathbf{S}_1, \mathbf{S}_2$. Si utilizamos el estimador lineal, entonces la estimación de Σ será

$$\mathbf{S} = (n_1\mathbf{S}_1 + n_2\mathbf{S}_2)/(n_1 + n_2)$$

y la versión muestral del discriminador lineal es

$$\widehat{L}(\mathbf{x}) = [\mathbf{x} - \frac{1}{2}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)]'\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2).$$

La distribución muestral de $\widehat{L}(\mathbf{x})$ es bastante complicada, pero la distribución asintótica es normal:

$$\widehat{L}(\mathbf{x}) \text{ es } N(+\tfrac{1}{2}\alpha, \alpha) \text{ si } \mathbf{x} \text{ proviene de } N_p(\mu_1, \Sigma),$$
$$\widehat{L}(\mathbf{x}) \text{ es } N(-\tfrac{1}{2}\alpha, \tfrac{1}{2}\alpha) \text{ si } \mathbf{x} \text{ proviene de } N_p(\mu_2, \Sigma),$$

donde $\alpha = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.

11.3.6 Un ejemplo

Exemple 11.3.1

Mytilicola intestinalis es un copépodo parásito del mejillón, que en estado larval presenta diferentes estadios de crecimiento. El primer estadio (Nauplis) y el segundo estadio (Metanauplius) son difíciles de distinguir.

Sobre una muestra de $n_1 = 76$ y $n_2 = 91$ copépodos que se pudieron identificar al microscopio como del primero y segundo estadio respectivamente, se midieron las variables

$$l = \text{longitud}, \quad a = \text{anchura},$$

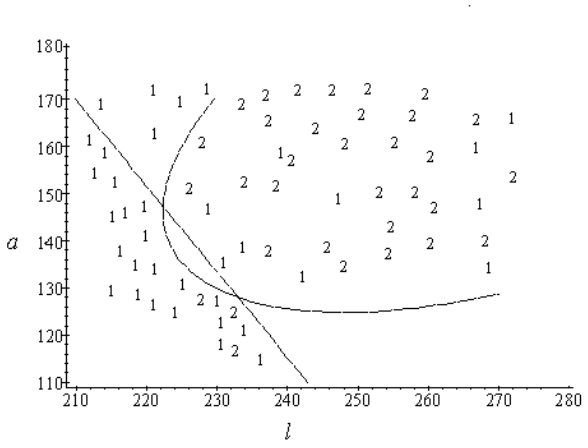

Figura 11.1: Discriminadores lineal y cuadrático en la clasificación de copépodos. La línea recta es el conjunto de puntos tales que $L = 0$. La parábola es el conjunto de puntos tales que $Q = 0$.

y se obtuvieron las siguientes medias y matrices de covarianzas:

$$\begin{array}{cccc} & \text{Estadio-1} & & \text{Estadio-2} \\ \overline{\mathbf{x}}_1 = & \begin{pmatrix} 219.5 & 138.1 \end{pmatrix} & \overline{\mathbf{x}}_2 = & \begin{pmatrix} 241.6 & 147.8 \end{pmatrix} \\ \mathbf{S}_1 = & \begin{pmatrix} 409.9 & -1.316 \\ -1.316 & 306.2 \end{pmatrix} & \mathbf{S}_2 = & \begin{pmatrix} 210.9 & 57.97 \\ 57.97 & 152.8 \end{pmatrix} \end{array}$$

Discriminador lineal

La estimación de la matriz de covarianzas común es

$$\mathbf{S} = (n_1\mathbf{S}_1 + n_2\mathbf{S}_2)/(n_1 + n_2) = \begin{pmatrix} 301.4 & 31.02 \\ 31.02 & 22.6 \end{pmatrix}$$

El discriminador lineal es:

$$\begin{aligned} L(l,a) &= ((l,a) - (461.1, 285.9)/2) \begin{pmatrix} 301.4 & 31.02 \\ 31.02 & 222.6 \end{pmatrix}^{-1} \begin{pmatrix} -22.1 \\ -9.7 \end{pmatrix} \\ &= 0.069l - 0.038a + 20.94 \end{aligned}$$

La tabla de clasificaciones es:

		Estadio asignado 1	2
Estadio	1	61	15
original	2	21	70

Discriminador de Bayes

Una larva, desde que eclosiona está 4 horas en el estadio 1 y 8 horas en el estadio 2. Al cabo de 12 horas, la larva pasa a un estadio fácilmente identificable. Por tanto, una larva tiene, a priori, una probabilidad $4/12 = 1/3$ de pertenecer al estadio 1 y una probabilidad $8/12 = 2/3$ de pertenecer al estadio 2. Así $q_1 = 1/3, q_2 = 2/3$, y el discriminador de Bayes es

$$B(l,a) = V(l,a) + \log(1/2) = 0.069l - 0.038a + 20.24$$

Probabilidad de clasificación errónea

Una estimación de la distancia de Mahalanobis es

$$\begin{pmatrix} -22.1 & -9.7 \end{pmatrix} \begin{pmatrix} 301.4 & 31.02 \\ 31.02 & 22.6 \end{pmatrix}^{-1} \begin{pmatrix} -22.1 \\ -9.7 \end{pmatrix} = 4.461.$$

La probabilidad de asignar una larva al estadio 1 cuando corresponde al estadio 2 o al estadio 2 cuando corresponde al estadio 1 es

$$pce = \Phi(-\frac{1}{2}\sqrt{4.461}) = 0.145.$$

Discriminador cuadrático

El test de homogeneidad de covarianzas nos da:

$$\chi^2 = [1 - \frac{13}{18}(\frac{1}{75} + \frac{1}{90} - \frac{1}{165})](1835.4 - 882.5 - 926.32) = 26.22$$

con 3 g.l. Las diferencias entre las matrices de covarianzas son significativas. Por tanto, el discriminador cuadrático puede resultar más apropiado. Efectuando cálculos se obtiene:

$$Q(l,a) = 0.0014l^2 + 0.002a^2 - 0.002al - 0.445l - 0.141a + 72.36$$

Con el clasificador cuadrático se han clasificado bien 2 individuos más (Fig. 11.1):

		Estadio asignado 1	2
Estadio original	1	59	17
	2	17	74

11.4 Discriminación en el caso de k poblaciones

Supongamos ahora que el individuo ω puede provenir de k poblaciones $\Omega_1, \Omega_2, \ldots, \Omega_k$, donde $k \geq 3$. Es necesario establecer una regla que permita asignar ω a una de las k poblaciones sobre la base de las observaciones $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ de p variables.

11.4.1 Discriminadores lineales

Supongamos que la media de las variables en Ω_i es μ_i, y que la matriz de covarianzas Σ es común. Si consideramos las distancias de Mahalanobis de ω a las poblaciones

$$M^2(\mathbf{x},\mu_i) = (\mathbf{x}-\mu_i)'\Sigma^{-1}(\mathbf{x}-\mu_i), \quad i = 1, \cdots, k,$$

un criterio de clasificación consiste en asignar ω a la población más próxima:

$$\text{Si } M^2(\mathbf{x},\mu_i) = \min\{M^2(\mathbf{x},\mu_1), \cdots, M^2(\mathbf{x},\mu_k)\}, \quad \text{asignamos } \omega \text{ a } \Omega_i. \quad (11.5)$$

Introduciendo las funciones discriminantes lineales

$$L_{ij}(\mathbf{x}) = \left(\mu_i - \mu_j\right)' \Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_i - \mu_j\right)' \Sigma^{-1}\left(\mu_i + \mu_j\right)$$

es fácil probar que (11.5) equivale a

$$\text{Si } L_{ij}(\mathbf{x}) > 0 \text{ para todo } j \neq i, \quad \text{asignamos } \omega \text{ a } \Omega_i.$$

Además las funciones $L_{ij}(\mathbf{x})$ verifican:

1. $L_{ij}(\mathbf{x}) = \frac{1}{2}[M^2(\mathbf{x},\mu_j) - M^2(\mathbf{x},\mu_i)]$.
2. $L_{ij}(\mathbf{x}) = -L_{ji}(\mathbf{x})$.
3. $L_{rs}(\mathbf{x}) = L_{is}(\mathbf{x}) - L_{ir}(\mathbf{x})$.

Es decir, sólo necesitamos conocer $k-1$ funciones discriminantes.

11.4.2 Regla de la máxima verosimilitud

Sea $f_i(\mathbf{x})$ la función de densidad de $\mathbf{x}$ en la población Ω_i. Podemos obtener una regla de clasificación asignando ω a la población donde la verosimilitud es más grande:

$$\text{Si } f_i(\mathbf{x}) = \max\{f_1(\mathbf{x}), \cdots, f_k(\mathbf{x})\}, \quad \text{asignamos } \omega \text{ a } \Omega_i.$$

Este criterio es más general que el geométrico y está asociado a las funciones discriminantes

$$V_{ij}(\mathbf{x}) = \log f_i(\mathbf{x}) - \log f_j(\mathbf{x}).$$

En el caso de normalidad multivariante y matriz de covarianzas común, se verifica $V_{ij}(\mathbf{x}) = L_{ij}(\mathbf{x})$, y los discriminadores máximo verosímiles coinciden con los lineales. Pero si las matrices de covarianzas son diferentes $\Sigma_1, \ldots, \Sigma_k$, entonces este criterio dará lugar a los discriminadores cuadráticos

$$\begin{aligned} Q_{ij}(\mathbf{x}) &= \tfrac{1}{2}\mathbf{x}'\left(\Sigma_j^{-1} - \Sigma_i^{-1}\right)\mathbf{x} + \mathbf{x}'\left(\Sigma_i^{-1}\mu_1 - \Sigma_j^{-1}\mu_2\right) \\ &\quad + \tfrac{1}{2}\mu_j'\Sigma_j^{-1}\mu_j - \tfrac{1}{2}\mu_i'\Sigma_i^{-1}\mu_i + \tfrac{1}{2}\log|\Sigma_j| - \tfrac{1}{2}\log|\Sigma_i|. \end{aligned}$$

11.4.3 Regla de Bayes

Si además de las funciones de densidad $f_i(\mathbf{x})$, se conocen las probabilidades a priori

$$q_1 = P(\Omega_1), \ldots, q_k = P(\Omega_k),$$

la regla de Bayes que asigna ω a la población tal que la probabilidad a posteriori es máxima

$$\text{Si } q_i f_i(\mathbf{x}) = \max\{q_1 f_1(\mathbf{x}), \cdots, q_k f_k(\mathbf{x})\}, \quad \text{asignamos } \omega \text{ a } \Omega_i,$$

está asociada a las funciones discriminantes

$$B_{ij}(\mathbf{x}) = \log f_i(\mathbf{x}) - \log f_j(\mathbf{x}) + \log(q_i/q_j).$$

Finalmente, si $P(j/i)$ es la probabilidad de asignar ω a Ω_j cuando en realidad es de Ω_i, la probabilidad de clasificación errónea es

$$pce = \sum_{i=1}^{k} q_i \Big(\sum_{j \neq i}^{k} P(j/i)\Big),$$

y se demuestra que la regla de Bayes minimiza esta *pce*.

11.4.4 Un ejemplo clásico

Continuando con el ejemplo 3.6.2, queremos clasificar a una de las 3 especies una flor de medidas

$$x_1 = 6.8 \quad x_2 = 2.8 \quad x_3 = 4.8 \quad x_4 = 1.4$$

La matriz de covarianzas común es

$$S = \begin{pmatrix} .2650 & .0927 & .1675 & .0384 \\ & .1154 & .05524 & .0327 \\ & & .18519 & .0426 \\ & & & .0418 \end{pmatrix}$$

Las distancies de Mahalanobis (al cuadrado) entre las 3 poblaciones son:

	Setosa	Versicolor	Virginica
Setosa	0	89.864	179.38
Versicolor		0	17.201
Virginica			0

Los discriminadores lineales son:

$$\begin{array}{l} L_{12}(x) = \frac{1}{2}\left[M^2(x, \overline{x}_2) - M^2(x, \overline{x}_1)\right], \\ L_{13}(x) = \frac{1}{2}\left[M^2(x, \overline{x}_3) - M^2(x, \overline{x}_1)\right], \\ L_{23}(x) = L_{13}(x) - L_{12}(x), L_{21}(x) = -L_{12}(x), \\ L_{31}(x) = -L_{13}(x),\ L_{32}(x) = -L_{23}(x). \end{array}$$

La regla de decisión consiste en asignar el individuo x a la población i si

$$L_{ij}(x) > 0 \quad \forall j \neq i.$$

Se obtiene:

Individuo	L_{12}	L_{13}	L_{21}	L_{23}	L_{31}	L_{32}	Población
x	-51.107	-44.759	51.107	6.3484	44.759	-6.3484	2

Por lo tanto clasificamos la flor a la especie I. Versicolor.

Para estimar la probabilidad de clasificación errónea *pce* podemos omitir una vez cada individuo, clasificarlo a partir de los demás y observar si sale bien clasificado (método *leaving-one-out*). El resultado de este proceso da:

		Población asignada		
		1	2	3
Población	1	50	0	0
original	2	0	48	2
	3	0	1	49

Sólo hay 3 individuos mal clasificados y la *pce* estimada es $3/150 = 0.02$.

11.5 Análisis discriminante basado en distancias

Los métodos que hemos descripto funcionan bien con variables cuantitativas o cuando se conoce la densidad. Pero a menudo las variables son binarias, categóricas o mixtas. Aplicando el principio de que siempre es posible definir una distancia entre observaciones, es posible dar una versión del análisis discriminante utilizando solamente distancias.

11.5.1 La función de proximidad

Sea Ω una población, $\mathbf{X}$ un vector aleatorio con valores en $E \subset R^p$ y densidad $f(x_1, ..., x_p)$. Sea δ una función de distancia entre las observaciones de $\mathbf{X}$. Definimos la variabilidad geométrica como la cantidad

$$V_\delta(\mathbf{X}) = \frac{1}{2}\int_E \delta^2(\mathbf{x}, \mathbf{y})\, f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y}$$

$V_\delta(\mathbf{X})$ es el valor esperado de las distancias (al cuadrado) entre observaciones independientes de $\mathbf{X}$.

Sea ω un individuo de Ω, y $\mathbf{x} = (x_1, ..., x_p)'$ las observaciones de $\mathbf{X}$ sobre ω. Definimos la función de proximidad de ω a Ω en relación con $\mathbf{X}$ como la función

$$\phi^2_\delta(\mathbf{x}) = E\left[\delta^2(\mathbf{x}, \mathbf{X})\right] - V_\delta(\mathbf{X}) = \int_E \delta^2(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) d\mathbf{t} - V_\delta(\mathbf{X}). \qquad (11.6)$$

$\phi^2_\delta(\mathbf{x})$ es la media de las distancias de $\mathbf{x}$, que es fija, a $\mathbf{t}$, que varía aleatoriamente, menos la variabilidad geométrica.

Teorema 11.5.1 *Supongamos que existe una representación de* (E, δ) *en un espacio* L *(Euclídeo o de Hilbert)*

$$(E, \delta) \to L$$

con un producto escalar $< ., . >$ *y una norma* $\|\mathbf{z}\|^2 =< \mathbf{z}, \mathbf{z} >$*, tal que*

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

donde $\psi(\mathbf{x}), \psi(\mathbf{y}) \in L$ *son las imágenes de* $\mathbf{x}, \mathbf{y}$*. Se verifica:*

- $V_\delta(\mathbf{X}) = E(\|\psi(\mathbf{X})\|^2) - \|E(\psi(\mathbf{X}))\|^2$.
- $\phi^2_\delta(\mathbf{x}) = \|\psi(\mathbf{x}) - E(\psi(\mathbf{X}))\|^2$.

En consecuencia, podemos afirmar que la variabilidad geométrica es una varianza generalizada, y que la función de proximidad mide la distancia de un individuo a la población.

11.5.2 La regla discriminante DB

Sean Ω_1, Ω_2 dos poblaciones, δ una función distancia. δ es formalmente la misma en cada población, pero puede tener diferentes versiones δ_1, δ_2 , cuando estemos en Ω_1, Ω_2, respectivamente. Por ejemplo, si las poblaciones son normales $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, 2$, y consideramos las distancias de Mahalanobis

$$\delta^2_i(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \mathbf{y}), \quad i = 1, 2,$$

lo único que cambia es la matriz Σ. Debe quedar claro que δ depende del vector aleatorio $\mathbf{X}$, que en general tendrá diferente distribución en Ω_1 y Ω_2.

Seguidamente, mediante (11.6), encontraremos las funciones de proximidad ϕ^2_1, ϕ^2_2, correspondientes a Ω_1, Ω_2. Sea ω un individuo que queremos clasificar, con valores $\mathbf{x} = \mathbf{X}(\omega)$.

La regla de clasificación DB (distance-based) es:

$$\begin{array}{ll} \text{Si } \phi^2_1(\mathbf{x}) \leq \phi^2_2(\mathbf{x}) & \text{asignamos } \omega \text{ a } \Omega_1, \\ \text{en caso contrario} & \text{asignamos } \omega \text{ a } \Omega_2. \end{array}$$

Teniendo en cuenta el Teorema 11.5.1, se cumple

$$\phi^2_i(\mathbf{x}) = \|\psi(\mathbf{x}) - E_{\Omega_i}(\psi(\mathbf{X}))\|^2, \quad i = 1, 2,$$

y por tanto la regla DB asigna ω a la población más próxima. La regla DB solamente depende de las distancias entre individuos.

11.5.3 La regla DB comparada con otras

Los discriminadores lineal y cuadrático son casos particulares de la regla DB.

1. Si las poblaciones son $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ y δ^2 es la distancia de Mahalanobis entre observaciones $\delta^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x}-\mathbf{y})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\mathbf{y})$, entonces las funciones de proximidad son

$$\phi_i^2(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})$$

y el discriminador lineal es

$$L(\mathbf{x}) = \frac{1}{2}\left[\phi_2^2(\mathbf{x}) - \phi_1^2(\mathbf{x})\right].$$

2. Si las poblaciones son $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ y δ_i^2 es la distancia de Mahalanobis más una constante

$$\begin{array}{ll} \delta_i^2(\mathbf{x}, \mathbf{y}) & = (\mathbf{x}-\mathbf{y})' \Sigma_i^{-1} (\mathbf{x}-\mathbf{y}) + \log|\Sigma_i|/2 \qquad \mathbf{x} \neq \mathbf{y}, \\ & = 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \mathbf{x} = \mathbf{y}, \end{array}$$

entonces el discriminador cuadrático es

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

3. Si δ es la distancia euclídea ordinaria entre observaciones, la regla DB equivale a utilizar el discriminador

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

conocido como *discriminador Euclídeo.* $E(\mathbf{x})$ es útil en determinadas circunstancias, por ejemplo, cuando la cantidad de variables es grande en relación al número de individuos, pues tiene la ventaja sobre $L(\mathbf{x})$ de que no necesita calcular la inversa de Σ.

11.5.4 La regla DB en el caso de muestras

En las aplicaciones prácticas, no se dispone de las densidades $f_1(\mathbf{x}), f_2(\mathbf{x})$, sino de dos muestras de tamaños n_1, n_2 de las variables $\mathbf{X} = (X_1, ..., X_p)$ en las poblaciones Ω_1, Ω_2. Sea $\Delta_1 = (\delta_{ij}(1))$ la matriz $n_1 \times n_1$ de distancias

entre las muestras de la primera población, y $\Delta_2 = (\delta_{ij}(2))$ la matriz $n_2 \times n_2$ de distancias entre las muestras de la segunda población. Indicamos (las representaciones Euclídeas de las muestras) por

$$\begin{array}{ll} \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{n_1} & \text{muestra de } \Omega_1, \\ \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{n_2} & \text{muestra de } \Omega_2, \end{array} \tag{11.7}$$

es decir, $\delta_{ij}(1) = \delta_E(\mathbf{x}_i, \mathbf{x}_j), \delta_{ij}(2) = \delta_E(\mathbf{y}_i, \mathbf{y}_j)$.

Las estimaciones de las variabilidades geométricas son:

$$\widehat{V}_1 = \frac{1}{2n_1^2} \sum_{i,j=1}^{n_1} \delta_{ij}^2(1), \quad \widehat{V}_2 = \frac{1}{2n_2^2} \sum_{i,j=1}^{n_2} \delta_{ij}^2(2).$$

Sea ω un individuo, $\delta_i(1), i = 1, \ldots, n_1$, las distancias a los n_1 individuos de Ω_1 y $\delta_i(2), i = 1, \ldots, n_2$, las distancias a los n_2 individuos de Ω_2. Si $\mathbf{x}$ son las coordenadas (convencionales) de ω cuando suponemos que es de Ω_1, y análogamente $\mathbf{y}$, las estimaciones de las funciones de proximidad son

$$\widehat{\phi}_1^2(\mathbf{x}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_i^2(1) - \widehat{V}_1, \quad \widehat{\phi}_2^2(\mathbf{y}) = \frac{1}{n_2} \sum_{i=1}^{n_2} \delta_i^2(2) - \widehat{V}_2.$$

La regla DB en el caso de muestras es

$$\begin{array}{ll} \text{Si } \widehat{\phi}_1^2(\mathbf{x}) \leq \widehat{\phi}_2^2(\mathbf{y}) & \text{asignamos } \omega \text{ a } \Omega_1, \\ \text{en caso contrario} & \text{asignamos } \omega \text{ a } \Omega_2. \end{array}$$

Esta regla solamente depende de distancias entre observaciones y es preciso insistir en que el conocimiento de $\mathbf{x}, \mathbf{y}$, no es necesario. La regla DB clasifica ω a la población más próxima:

Teorema 11.5.2 *Supongamos que podemos representar ω y las dos muestras en dos espacios euclídeos (posiblemente diferentes)*

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

respectivamente. Entonces se cumple

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

donde $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ son los centroides de las representaciones Euclídeas de las muestras.

Demost.: Consideremos $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, \overline{\mathbf{x}} = (\sum_{i=1}^n \mathbf{x}_i)/n$. Por un lado

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n d^2(\mathbf{x}_i, \mathbf{x}) &= \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \mathbf{x})'(\mathbf{x}_i - \mathbf{x}) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i + \mathbf{x}'\mathbf{x} - 2\overline{\mathbf{x}}'\overline{\mathbf{x}}. \end{aligned}$$

Por otro

$$\begin{aligned} \frac{1}{2n^2}\sum_{i,j=1}^n d^2(\mathbf{x}_i, \mathbf{x}_j) &= \frac{1}{2n^2}\sum_{i,j=1}^n (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i - \overline{\mathbf{x}}'\overline{\mathbf{x}}. \end{aligned}$$

Restando

$$\widehat{\phi}^2(\mathbf{x}) = \mathbf{x}'\mathbf{x} + \overline{\mathbf{x}}'\overline{\mathbf{x}} - 2\overline{\mathbf{x}}'\mathbf{x} = d_E^2(\mathbf{x}, \overline{\mathbf{x}}) . \square$$

11.6 Complementos

El Análisis Discriminante se inicia en 1936 con el trabajo de R.A. Fisher sobre clasificación de flores del género Iris. A. Wald y T.W. Anderson estudiaron las propiedades del discriminador lineal. L. Cavalli y C.A.B. Smith introdujeron el discriminador cuadrático.

J.A. Anderson estudió la discriminación logística. Si definimos

$$y(\omega, \mathbf{x}) = P(\Omega_1/\mathbf{x}) = q_1 f_1(\mathbf{x})/(q_1 f_1(\mathbf{x}) + q_2 f_2(\mathbf{x})),$$

la regla de classificación es

$$\omega \text{ es de } \Omega_1 \text{ si } y(\omega, \mathbf{x}) > 1/2, \text{ de } \Omega_2 \text{ en caso contrario.}$$

Entonces el modelo logístico supone

$$y(\omega, \mathbf{x}) = \frac{1}{1 + e^{-\alpha - \boldsymbol{\beta}'\mathbf{x}}}.$$

Existen otros métodos de análisis discriminante, algunos no-paramétricos, otros para variables mixtas, como el método del núcleo, del vecino mas próximo, el basado en el "location model" de W. Krzanowski, etc. Consultar McLachlan (1992).

Los métodos de análisis discriminante basados en distancias pueden abordar todo tipo de datos y han sido estudiados por Cuadras (1989, 1992b), Cuadras *et al.* (1997).

Capítulo 12

EL MODELO LINEAL

12.1 El modelo lineal

Supongamos que una variable observable Y depende de varias variables explicativas (caso de la regresión múltiple), o que ha sido observada en diferentes situaciones experimentales (caso del análisis de la varianza). Entonces tendremos n observaciones de Y, que en muchas situaciones aplicadas, se ajustan a un *modelo lineal*

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{im}\beta_m + e_i, \quad i = 1, \ldots, n, \tag{12.1}$$

que en notación matricial es

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Los elementos que intervienen en el modelo lineal son:

1. El vector de observaciones de Y

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)'.$$

2. El vector de parámetros

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)'.$$

3. La matriz de diseño

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ & & \ddots & \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}.$$

4. El vector de desviaciones aleatorias

$$\mathbf{e} = (e_1, e_2, \ldots, e_n)'$$

La notación matricial compacta del modelo es:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Solamente $\mathbf{y}$ y $\mathbf{X}$ son conocidas. En los modelos de regresión, $\mathbf{X}$ contiene las observaciones de m variables explicativas. En los modelos de análisis de la varianza, $\mathbf{X}$ contiene los valores $0, 1$ ó -1, según el tipo de diseño experimental.

12.2 Suposiciones básicas del modelo

Supongamos que las desviaciones aleatorias o errores e_i del modelo lineal se asimilan a n variables aleatorias con media 0, incorrelacionadas y con varianza común σ^2, es decir, satisfacen:

1. $E(e_i) = 0, \quad i = 1, \ldots, n.$
2. $E(e_i e_j) = 0, \quad i \neq j = 1, \ldots, n.$
3. $\text{var}(e_i) = \sigma^2, \quad i = 1, \ldots, n.$

Estas condiciones equivalen a decir que el vector de medias y la matriz de covarianzas del vector $\mathbf{e} = (e_1, e_2, \ldots, e_n)'$ son:

$$E(\mathbf{e}) = \mathbf{0}, \quad \Sigma_e = \sigma^2 \mathbf{I}_p.$$

Si podemos suponer que los errores son normales y estocásticamente independientes, entonces estamos ante un *modelo lineal normal*

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_p).$$

La cantidad $r = \text{rang}(\mathbf{X})$ es el rango del diseño. Tenemos $r \leq m$ y cuando $r = m$ se dice que es un modelo de rango máximo.

12.3 Estimación de parámetros

12.3.1 Parámetros de regresión

La estimación de los parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ en función de las observaciones $\mathbf{y} = (y_1, \ldots, y_n)'$, se plantea mediante el criterio de los mínimos cuadrados (LS, "least squares"). Se desea encontrar $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)'$ tal que

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - \ldots - x_{im}\beta_m)^2 \tag{12.2}$$

sea mínimo.

Teorema 12.3.1 *Toda estimación LS de $\boldsymbol{\beta}$ es solución de las ecuaciones*

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \tag{12.3}$$

denominades ecuaciones normales del modelo.

Demost.:

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + 2\boldsymbol{\beta}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

Derivando vectorialmente respecto de $\boldsymbol{\beta}$ e igualando a cero

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{e}'\mathbf{e} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

obtenemos (12.3).□

Distinguiremos dos casos según el rango del diseño.

a) $r = m$. Entonces la estimación de $\boldsymbol{\beta}$ es única:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{12.4}$$

b) $r < m$. Cuando el diseño no es de rango máximo una solución es

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y},$$

donde $(\mathbf{X}'\mathbf{X})^{-}$ es una inversa generalizada de $\mathbf{X}'\mathbf{X}$.

La suma de cuadrados residual de la estimación de $\boldsymbol{\beta}$ es

$$R_0^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2,$$

siendo

$$\widehat{y}_i = x_{i1}\widehat{\beta}_1 + \ldots + x_{im}\widehat{\beta}_m.$$

12.3.2 Varianza

La varianza común de los términos de error, σ^2 =var(e_i), es el otro parámetro que hemos de estimar en función de las observaciones $\mathbf{y} = (y_1, \ldots, y_n)'$ y de $\mathbf{X}$. En esta estimación interviene de manera destacada la suma de cuadrados residual.

Lema 12.3.2 *Sea $C_r(\mathbf{X})$ el subespacio de R^n de dimensión r generado por las columnas de $\mathbf{X}$. Entonces $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \in C_r(\mathbf{X})$ y $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ es ortogonal a $C_r(\mathbf{X})$.*

Demost.: Por las ecuaciones normales

$$\mathbf{X}'\widehat{\mathbf{e}} = \mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0}.$$

Teorema 12.3.3 *Sea $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ el modelo lineal donde $\mathbf{e}$ satisface las suposiciones básicas del modelo (Sección 12.2). Entonces el estadístico*

$$\widehat{\sigma}^2 = R_0^2/(n - r),$$

siendo R_0^2 la suma de cuadrados residual y $r = rang(\mathbf{X})$ el rango del modelo, es un estimador insesgado de σ^2.

Demost.: Sea $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_r, \mathbf{t}_{r+1}, \ldots, \mathbf{t}_n]$ una matriz ortogonal tal que sus columnas formen una base ortonormal de R^n, de manera que las r primeras generen el subespacio $C_r(\mathbf{X})$ y por tanto las otras $n - r$ sean ortogonales a $C_r(\mathbf{X})$. Definimos $\mathbf{z} = \mathbf{T}'\mathbf{y}$. Entonces $\mathbf{z}$ =$(z_1, \ldots, z_n)'$ verifica

$$\begin{aligned} E(z_i) \quad &= \mathbf{t}_i'\mathbf{X}\boldsymbol{\beta} \quad = \eta_i \quad \text{si } i \leq r, \\ &\qquad\qquad\quad = 0 \quad \text{si } i > r, \end{aligned}$$

pues $\mathbf{t}_i$ es ortogonal a $C_r(\mathbf{X})$ si $i > r$. Consideremos $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$. Entonces $\mathbf{T}'\widehat{\mathbf{e}} = \mathbf{z} - \mathbf{T}'\mathbf{X}\widehat{\boldsymbol{\beta}}$, donde las r primeras componentes de $\mathbf{T}'\widehat{\mathbf{e}}$ son cero (por el lema anterior) y las $n-r$ componentes de $\mathbf{T}'\mathbf{X}\widehat{\boldsymbol{\beta}}$ son también cero. Por tanto $\mathbf{T}'\widehat{\mathbf{e}}$ es

$$\mathbf{T}'\widehat{\mathbf{e}} = (0, \ldots, 0, z_{r+1}, \ldots, z_n)'$$

y en consecuencia

$$R_0^2 = \widehat{\mathbf{e}}'\widehat{\mathbf{e}} = \widehat{\mathbf{e}}'\mathbf{T}\mathbf{T}'\widehat{\mathbf{e}} = \sum_{i=r+1}^{n} z_i^2.$$

La matriz de covarianzas de $\mathbf{y}$ es $\sigma^2\mathbf{I}_n$, y por ser $\mathbf{T}$ ortogonal, la de $\mathbf{z}$ es también $\sigma^2\mathbf{I}_n$. Así

$$E(z_i) = 0, \quad E(z_i^2) = \text{var}(z_i) = \sigma^2, \quad i > r,$$

y por tanto

$$E(R_o^2) = \sum_{i=r+1}^{n} E(z_i^2) = (n-r)\sigma^2.\square$$

Bajo el modelo lineal normal, la estimación de $\boldsymbol{\beta}$ es estocásticamente independiente de la estimación de σ^2, que sigue la distribución ji-cuadrado.

Teorema 12.3.4 *Sea* $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_p)$ *el modelo lineal normal de rango máximo* $m = rang(\mathbf{X})$. *Se verifica:*

1. *La estimación LS de* $\boldsymbol{\beta}$ *es también la estimación máximo verosímil de* $\boldsymbol{\beta}$. *Esta estimación es además insesgada y de varianza mínima.*
2. $\widehat{\boldsymbol{\beta}} \sim N_m(\boldsymbol{\beta},\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.
3. $U = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_m$.
4. $\widehat{\boldsymbol{\beta}}$ *es estocásticamente independiente de* R_0^2.
5. $R_0^2/\sigma^2 \sim \chi^2_{n-r}$.

12.4 Algunos modelos lineales

12.4.1 Regresión múltiple

El modelo de regresión múltiple de una variable respuesta Y sobre m variables explicativas $X_1, \ldots, X_m$ es

$$y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{im}\beta_m + e_i, \quad i = 1, \ldots, n, \tag{12.5}$$

donde y_i es la i-ésima observación de Y, y $x_{i1}, \ldots, x_{im}$ son las i-ésimas observaciones de las variables explicativas. La matriz de diseño es

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix}.$$

12.4.2 Diseño de un factor

Supongamos que una variable observable Y ha sido observada en k condiciones experimentales diferentes, y que disponemos de n_i réplicas (observaciones independentes de Y) $y_{i1}, \ldots, y_{in_i}$ bajo la condición experimental i. El modelo es

$$y_{ih} = \mu + \alpha_i + e_{ih}, \quad i = 1, \ldots, k;\ h = 1, \ldots, n_i, \tag{12.6}$$

donde μ es la media general y α_i es el efecto aditivo de la condición i. Las desviaciones aleatorias e_{ih} se suponen normales independientes. En el modelo (12.6), se supone la restricción lineal

$$\alpha_1 + \ldots + \alpha_k = 0,$$

y por tanto cabe considerar solamente los parámetros $\mu, \alpha_1, \ldots, \alpha_{k-1}$. Por ejemplo, si $k = 3, n_1 = n_2 = 2, n_3 = 3$, la matriz de diseño es

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccc} \mu & \alpha_1 & \alpha_2 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \end{array}$$

12.4.3 Diseño de dos factores

Supongamos que las $n = a \times b$ observaciones de una variable observable Y se obtienen combinando dos factores con a y b niveles, respectivamente, denominados factor fila y columna (por ejemplo, producción de trigo obtenida en $9 = 3 \times 3$ parcelas, 3 fincas y 3 fertilitzantes en cada finca). El modelo es

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \tag{12.7}$$

donde μ es la media general, α_i es el efecto aditivo del nivel i del factor fila, β_j es el efecto aditivo del nivel j del factor columna. Las desviaciones aleatorias e_{ij} se suponen normales independientes. En el modelo (12.6) se suponen las restricciones lineales

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0. \tag{12.8}$$

Por ejemplo, si $a = b = 3$ la matriz de diseño es

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccccc} \mu & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 \end{array} \\ \left(\begin{array}{ccccc} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & -1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & -1 & -1 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 \\ 1 & 0 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 \end{array} \right) \end{array}$$

12.5 Hipótesis lineales

Consideremos el modelo lineal normal $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. Una hipótesis lineal es una restricción lineal sobre los parámetros $\boldsymbol{\beta}$ del modelo.

Definició 12.5.1 *Una hipótesis lineal de rango t sobre los parámetros $\boldsymbol{\beta}$ es una restricción lineal*

$$h_{i1}\beta_1 + \ldots + h_{im}\beta_m = 0, \qquad i = 1, \ldots, t.$$

Indicando la matriz $t \times m$, con $t < m$ filas linealmente independientes,

$$\mathbf{H} = \begin{pmatrix} h_{11} & \cdots & h_{1m} \\ & \cdots & \\ h_{t1} & \cdots & h_{tm} \end{pmatrix}$$

la notación matricial de una hipótesis lineal es

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{0}. \tag{12.9}$$

Definició 12.5.2 *Una hipótesis lineal es demostrable si las filas de $\mathbf{H}$ son combinación lineal de las filas de $\mathbf{X}$. Dicho de otra manera, si existe una matriz $\mathbf{A}$ de orden $t \times n$ tal que*

$$\mathbf{H} = \mathbf{A}\mathbf{X}.$$

Observaciones:

a) Suponemos que la matriz $\mathbf{H}$ es de rango t.

b) Solamente podremos construir un test (el test F) para decidir si podemos aceptar o no una hipótesis lineal si esta hipótesis es "demostrable".

c) Es evidente que si el modelo es de rango máximo, $r = \text{rang}(\mathbf{X}) = m$, cualquier hipótesis lineal es demostrable.

Cuando una hipótesis (12.9) es cierta, los parámetros $\boldsymbol{\beta}$ se convierten en $\boldsymbol{\theta}$ y la matriz de diseño $\mathbf{X}$ en $\widetilde{\mathbf{X}}$. Así el modelo lineal, bajo H_0, es

$$\mathbf{y} = \widetilde{\mathbf{X}}\boldsymbol{\theta} + \mathbf{e}. \tag{12.10}$$

Para obtener (12.10), consideramos los subespacios $F(\mathbf{H})$,$F(\mathbf{X})$ generados por las filas de $\mathbf{H}$ y $\mathbf{X}$. Entonces $F(\mathbf{H}) \subset F(\mathbf{X}) \subset R^m$. Sea $\mathbf{C}$ una matriz $m \times (r-t)$ tal que $F(\mathbf{C}') \subset F(\mathbf{X})$ y $\mathbf{HC} = \mathbf{0}$. En otras palabras, las columnas de $\mathbf{C}$ pertenecen a $F(\mathbf{X})$ y son ortogonales a $F(\mathbf{H})$. Si definimos los parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{r-t})'$ tales que

$$\boldsymbol{\beta} = \mathbf{C}\boldsymbol{\theta},$$

entonces $\mathbf{H}\boldsymbol{\beta} = \mathbf{HC}\boldsymbol{\beta} = \mathbf{0}$ y el modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, bajo la restricción $\mathbf{H}\boldsymbol{\beta} = \mathbf{0}$, se transforma en (12.10), siendo

$$\widetilde{\mathbf{X}} = \mathbf{XC}.$$

La estimación LS de $\boldsymbol{\theta}$ es

$$\widehat{\boldsymbol{\theta}} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}\mathbf{y}$$

y la suma de cuadrados residual es

$$R_1^2 = (\mathbf{y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}})'(\mathbf{y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}).$$

También se puede probar que la estimación LS de los parámetros $\boldsymbol{\beta}$, bajo la restricción (12.9), es

$$\widehat{\boldsymbol{\beta}}_H = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{H}'(\mathbf{H}(\mathbf{X}'\mathbf{X})^{-}\mathbf{H}')^{-1}\mathbf{H}\widehat{\boldsymbol{\beta}}$$

y la suma de cuadrados del modelo lineal es

$$R_1^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_H)'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_H)$$

El siguiente teorema es conocido como Teorema Fundamental del Análisis de la Varianza.

Teorema 12.5.1 *Sea* $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_p)$ *el modelo lineal normal y* $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$ *una hipótesis lineal demostrable de rango* t. *Consideremos los estadísticos*

$$R_0^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}), \quad R_1^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_H)'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_H).$$

Se verifica:

1. $R_0^2/\sigma^2 \sim \chi^2_{n-r}$.
2. *Si* H_0 *es cierta*
$$\frac{R_1^2}{\sigma^2} \sim \chi^2_{n-r'}, \quad \frac{R_1^2 - R_0^2}{\sigma^2} \sim \chi^2_t,$$
siendo $r' = r - t$.
3. *Si* H_0 *es cierta, los estadísticos* $(R_1^2 - R_0^2)$ *y* R_0^2 *son estocásticamente independientes.*

Demost.: Observemos primero que bajo el modelo lineal normal, $y_1, \ldots, y_n$ son normales independientes, y $z_1, \ldots, z_n$ (véase Teorema 12.3.3) son también normales independientes.

1. Cada z_i es $N(0, \sigma^2)$ para $i > r$. Luego R_0^2/σ^2 es suma de $(n-r)$ cuadrados de $N(0,1)$ independientes.
2. Si la hipótesis lineal es cierta, la matriz de diseño $\mathbf{X}$ se transforma en $\widetilde{\mathbf{X}} = \mathbf{XC}$, es decir, las columnas de $\mathbf{XC}$ son combinación lineal de las columnas de $\mathbf{X}$. Podemos encontrar una matriz ortogonal
$$\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_{r'}, \mathbf{t}_{r'+1}, \ldots, \mathbf{t}_r, \mathbf{t}_{r+1}, \ldots, \mathbf{t}_n]$$
tal que
$$C_{r'}(\mathbf{XC}) = [\mathbf{t}_1, \ldots, \mathbf{t}_{r'}] \subset C_r(\mathbf{X}) = [\mathbf{t}_1, \ldots, \mathbf{t}_r].$$
Siguiendo los mismos argumentos del Teorema 12.3.3, tenemos que
$$R_1^2 = \sum_{i=r'+1}^{n} z_i^2$$
y R_1^2/σ^2 sigue la distribución $\chi^2_{n-r'}$. Por otro lado
$$R_1^2 - R_0^2 = \sum_{i=r'+1}^{r} z_i^2$$
y $(R_1^2 - R_0^2)/\sigma^2$ sigue la distribución χ^2_t, donde $t = r - r'$.

3. Las sumas de cuadrados que intervienen en R_0^2 y en $R_1^2 - R_0^2$ no tienen términos en común, por tanto son independientes.□

Consecuencia inmediata y muy importante de este resultado es que, si H_0 es cierta, entonces el estadístico

$$F = \frac{(R_1^2 - R_0^2)/t\sigma^2}{R_0^2/(n-r)\sigma^2} = \frac{(R_1^2 - R_0^2)}{R_0^2}\frac{n-r}{t} \sim F_{n-r}^t. \tag{12.11}$$

Es decir, F sigue la distribución F con t y $n-r$ grados de libertad y no depende de la varianza (desconocida) del modelo.

12.6 Inferencia en regresión múltiple

Consideremos el modelo de regresión múltiple (12.5). El rango del modelo es $\text{rang}(\mathbf{X}) = m+1$. La hipótesis más interesante en las aplicaciones es

$$H_0 : \beta_1 = \ldots = \beta_m = 0,$$

que equivale a decir que la variable respuesta Y no depende de las variables explicativas $X_1, \ldots, X_m$. La matriz de la hipótesis lineal es

$$\mathbf{H} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad rang(\mathbf{H}) = m.$$

Si H_0 es cierta, solamente interviene el parámetro β_0, evidentemente $\widehat{\beta}_{0H} = \overline{y}$ (media muestral) y las sumas de cuadrados residuales son

$$R_0^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2, \quad R_1^2 = \sum_{i=1}^n (y_i - \overline{y})^2,$$

donde $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_m$ son los estimadores LS bajo el modelo no restringido y $\widehat{y}_i = \widehat{\beta}_0 + x_{i1}\widehat{\beta}_1 + \ldots + x_{im}\widehat{\beta}_m$. Aplicando (12.11), bajo H_0 tenemos que

$$F = \frac{(R_1^2 - R_0^2)}{R_0^2}\frac{n-m-1}{m} \sim F_{n-m-1}^m.$$

El test F se suele expresar en términos de la correlación múltiple. Se demuestra que

$$R_0^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = (1 - R^2)\sum_{i=1}^{n}(y_i - \overline{y})^2,$$

donde R es el coeficiente de correlación múltiple muestral entre Y y $X_1, \ldots, X_m$ (Teorema 4.2.2). Por tanto, si H_0 es cierta, es decir, si la correlación múltiple poblacional es cero, entonces

$$F = \frac{R^2}{1 - R^2}\frac{n - m - 1}{m} \sim F_{n-m-1}^{m}.$$

Rechazaremos H_0 si F es significativa.

12.7 Complementos

Hemos visto los aspectos fundamentales del modelo lineal. Un estudio más completo incluiría:

a) análisis gráfico de los residuos, b) efectos de la colinealidad, c) mínimos cuadrados ponderados, d) errores correlacionados, e) selección de las variables, etc. Ver Peña (1989), Chatterjee y Price (1991).

Para tratar variables explicativas mixtas, podemos definir un modelo lineal considerando las dimensiones principales obtenidas aplicando análisis de coordenadas principales sobre una matriz de distancias entre las observaciones. Consultar Cuadras y Arenas (1990), Cuadras *et al.* (1996).

Capítulo 13

ANÁLISIS DE LA VARIANZA (ANOVA)

El análisis de la varianza comprende un conjunto de técnicas estadísticas que permiten analizar como operan diversos factores, estudiados simultáneamente en un diseño factorial, sobre una variable respuesta.

13.1 Diseño de un factor

Supongamos que las observaciones de una variable Y solamente dependen de un factor con k niveles:

$$\begin{array}{lllll} \text{Nivel 1} & y_{11} & y_{12} & \cdots & y_{1n_1} \\ \text{Nivel 2} & y_{21} & y_{22} & \cdots & y_{2n_2} \\ \cdots & & \cdots & & \\ \text{Nivel k} & y_{k1} & y_{k2} & \cdots & y_{kn_k} \end{array}$$

Si escribimos $\mu_i = \mu + \alpha_i$, en el modelo (12.6) tenemos

$$y_{ih} = \mu_i + e_{ih}, \quad i = 1, \ldots, k;\ h = 1, \ldots, n_i,$$

donde μ_i es la media de la variable en el nivel i. Indiquemos:

$$\begin{array}{lll} \text{Media nivel } i: & y_{i\cdot} & = (1/n_i) \sum_h y_{ih} \\ \text{Media general:} & \overline{y} & = (1/n) \sum_i \sum_h y_{ih} \\ \text{No. total de observaciones:} & n & = n_1 + \ldots + n_k \end{array}$$

También indiquemos:

Suma de cuadrados entre grupos:	Q_E	$= \sum_i n_i(y_{i\cdot} - \overline{y})^2$
Suma de cuadrados dentro de grupos:	Q_D	$= \sum_i \sum_h (y_{ih} - y_{i\cdot})^2$
Suma de cuadrados total:	Q_T	$= \sum_i \sum_h (y_{ih} - \overline{y})^2$

Se verifica la relación fundamental:

$$Q_T = Q_E + Q_D.$$

Las estimaciones LS de las medias μ_i son

$$\widehat{\mu}_i = y_{i\cdot}, \quad i = 1, \ldots, k,$$

y la suma de cuadrados residual es $R_0^2 = Q_D$.

La hipótesis nula de principal interés es la que establece que no existen diferencias entre los niveles de los factores:

$$H_0 : \mu_1 = \ldots = \mu_k,$$

y tiene rango 1. Bajo H_0 solamente existe una media μ y su estimación es $\widehat{\mu} = \overline{y}$. Entonces la suma de cuadrados residual es $R_1^2 = Q_T$ y además se verifica

$$R_1^2 - R_0^2 = Q_E$$

Por tanto, como una consecuencia del Teorema 12.5.1, tenemos que:

1. $Q_D/(n-k)$ es un estimador centrado de σ^2 y $Q_D/\sigma^2 \sim \chi^2_{n-k}$.
2. Si H_0 es cierta, $Q_E/(k-1)$ es también estimador centrado de σ^2 y
$$\frac{Q_T}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{Q_E}{\sigma^2} \sim \chi^2_{k-1}.$$
3. Si H_0 es cierta, los estadísticos Q_E y Q_D son estocásticamente independientes.

Consecuencia inmediata es que, si H_0 es cierta, entonces el estadístico

$$F = \frac{Q_E/(k-1)}{Q_D/(n-k)} \sim F^{k-1}_{n-k}.$$

13.2 Diseño de dos factores

Supongamos que las observaciones de una variable Y dependen de dos factores A, B, denominados factores fila y columna, con a y b niveles $\mathrm{A}_1,\ldots,\mathrm{A}_a$ y $\mathrm{B}_1,\ldots,\mathrm{B}_b$, y que disponemos de una observación para cada combinación de los niveles de los factores:

	B_1	B_2	$\cdots$	B_b	
A_1	y_{11}	y_{12}	$\cdots$	y_{1b}	$y_{1\cdot}$
A_2	y_{21}	y_{22}	$\cdots$	y_{2b}	$y_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	y_{a1}	y_{a2}	$\cdots$	y_{ab}	$y_{a\cdot}$
	$y_{\cdot 1}$	$y_{\cdot 2}$	$\cdots$	$y_{\cdot b}$	$y_{\cdot\cdot}$

siendo

$$y_{i\cdot}=\frac{1}{b}\sum_{j=1}^{b}y_{ij},\quad y_{\cdot j}=\frac{1}{a}\sum_{i=1}^{a}y_{ij},\quad y_{\cdot\cdot}=\overline{y}=\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}y_{ij},$$

las medias por filas, por columnas y general. Supongamos que los datos se ajustan al modelo (12.7) con las restricciones (12.8), donde μ es la media general, α_i es el efecto del nivel A_i del factor fila, β_j es el efecto del nivel B_j del factor columna. El rango del diseño y los g.l. del residuo son

$$r=1+(a-1)+(b-1)=a+b-1,\ n-r=ab-(a+b-1)=(a-1)(b-1).$$

Las estimaciones de los parámetros son

$$\widehat{\mu}=\overline{y},\quad \widehat{\alpha}_i=y_{i\cdot}-\overline{y},\quad \widehat{\beta}_j=y_{\cdot j}-\overline{y},$$

y la expresión de la desviación aleatoria es

$$\widehat{e}_{ij}=y_{ij}-\widehat{\mu}-\widehat{\alpha}_i-\widehat{\beta}_j=(y_{ij}-y_{i\cdot}-y_{\cdot j}+\overline{y}).$$

La suma de cuadrados residual del modelo es

$$R_0^2=\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-y_{i\cdot}-y_{\cdot j}+\overline{y})^2.$$

También consideramos las cantidades:

Suma de cuadrados entre filas:	Q_A	$= b\sum_i (y_{i\cdot} - \overline{y})^2$
Suma de cuadrados entre columnas:	Q_B	$= a\sum_j (y_{\cdot j} - \overline{y})^2$
Suma de cuadrados residual:	Q_R	$= \sum_{i,j} (y_{ij} - y_{i\cdot} - y_{\cdot j} + \overline{y})^2$
Suma de cuadrados total:	Q_T	$= \sum_{i,j} (y_{ij} - \overline{y})^2$

Se verifica la siguiente identidad:

$$Q_T = Q_A + Q_B + Q_R.$$

En el modelo de dos factores, las hipótesis de interés son:

$$\begin{array}{ll} H_0^A: & \alpha_1 = \cdots = \alpha_a = 0 \quad \text{(no hay efecto fila)} \\ H_0^B: & \beta_1 = \cdots = \beta_b = 0 \quad \text{(no hay efecto columna)} \end{array}$$

Supongamos H_0^B cierta. Entonces el modelo se transforma en $y_{ij} = \mu + \alpha_i + e_{ij}$, es decir, actúa solamente un factor, y por tanto

$$R_1^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} (y_{ij} - y_{i\cdot})^2.$$

Ahora bien, desarrollando $(y_{ij} - y_{i\cdot})^2 = ((y_{\cdot j} - \overline{y}) + (y_{ij} - y_{i\cdot} - y_{\cdot j} + \overline{y}))^2$ resulta que

$$R_1^2 = Q_B + Q_R.$$

Análogamente, si H_0^F es cierta, obtendríamos $R_1^2 = Q_A + Q_R$. Por el Teorema 12.5.1 se verifica:

1. $Q_R/(a-1)(b-1)$ es un estimador centrado de σ^2 y $Q_R/\sigma^2 \sim \chi^2_{(a-1)(b-1)}$.

2. Si H_0^A es cierta, $Q_A/(a-1)$ es también estimador centrado de σ^2, $Q_A/\sigma^2 \sim \chi^2_{(a-1)}$ y los estadísticos Q_A y Q_R son estocásticamente independientes.

3. Si H_0^B es cierta, $Q_B/(b-1)$ es también estimador centrado de σ^2, $Q_B/\sigma^2 \sim \chi^2_{(b-1)}$ y los estadísticos Q_B y Q_R son estocásticamente independientes.

Por lo tanto tenemos que para decidir H_0^A utilizaremos el estadístico

$$F_A = \frac{Q_A}{Q_R}\frac{(a-1)(b-1)}{(a-1)} \sim F^{a-1}_{(a-1)(b-1)},$$

y para decidir H_0^B utilizaremos

$$F_B = \frac{Q_B}{Q_R}\frac{(a-1)(b-1)}{(b-1)} \sim F^{b-1}_{(a-1)(b-1)}.$$

13.3 Diseño de dos factores con interacción

Supongamos que las observaciones de una variable Y dependen de dos factores A, B, denominados factores fila y columna, con a y b niveles $\mathrm{A}_1, \ldots \mathrm{A}_a$ y $\mathrm{B}_1, \ldots, \mathrm{B}_b$, y que disponemos de c observaciones (réplicas) para cada combinación de los niveles de los factores:

	B_1	B_2	$\cdots$	B_b	
A_1	$y_{111}, \ldots, y_{11c}$	$y_{121}, \ldots, y_{12c}$	$\cdots$	$y_{1b1}, \ldots, y_{1bc}$	$y_{1\cdot\cdot}$
A_2	$y_{211}, \ldots, y_{21c}$	$y_{221}, \ldots, y_{22c}$	$\cdots$	$y_{2b1}, \ldots, y_{2bc}$	$y_{2\cdot\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	$y_{a11}, \ldots, y_{a1c}$	$y_{a22}, \ldots, y_{a2c}$	$\cdots$	$y_{ab1}, \ldots, y_{abc}$	$y_{a\cdot\cdot}$
	$y_{\cdot 1\cdot}$	$y_{\cdot 2\cdot}$	$\cdots$	$y_{\cdot b\cdot}$	$y_{\ldots}$

siendo

$$y_{i\cdot\cdot} = \frac{1}{bc}\sum_{j,h=1}^{b,c} y_{ijh}, \qquad y_{\cdot j\cdot} = \frac{1}{ac}\sum_{i,h=1}^{a,c} y_{ijh},$$
$$y_{ij\cdot} = \frac{1}{c}\sum_{h=1}^{c} y_{ijh}, \quad \overline{y} = y_{\ldots} = \frac{1}{abc}\sum_{i,j,h=1}^{a,b,c} y_{ij}.$$

El modelo lineal del diseño de dos factores con interacción es

$$y_{ijh} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijh},$$
$$i = 1, \ldots, a;\ j = 1, \ldots, b;\ h = 1, \ldots, c,$$

siendo μ la media general, α_i el efecto del nivel A_i del factor fila, β_j el efecto del nivel B_j del factor columna, γ_{ij} la interacción entre los niveles A_i,B_j. El parámetro γ_{ij} mide la desviación del modelo aditivo $E(y_{ijh}) =$

$\mu + \alpha_i + \beta_j$ y solamente es posible estimar si hay $c > 1$ réplicas. Se suponen las restricciones

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0.$$

Así el número de parámetros independientes del modelo es

$$1 + (a-1) + (b-1) + (a-1)(b-1) = ab$$

y los g.l. del residuo son $abc - ab = ab(c-1)$.

Las estimaciones de los parámetros son

$$\widehat{\mu} = \overline{y}, \quad \widehat{\alpha}_i = y_{i\cdot\cdot} - \overline{y}, \quad \widehat{\beta}_j = y_{\cdot j\cdot} - \overline{y}, \quad \widehat{\gamma}_{ij} = y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + \overline{y},$$

y la expresión de la desviación aleatoria es

$$\widehat{e}_{ijh} = y_{ijh} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j - \widehat{\gamma}_{ij} = (y_{ij} - \overline{y}).$$

La suma de cuadrados residual del modelo es

$$R_0^2 = \sum_{i,j,h=1}^{a,b,c} (y_{ijh} - y_{i\cdot\cdot})^2.$$

También debemos considerar las cantidades:

Suma de cuadrados entre filas:	Q_A	$= bc\sum_i (y_{i\cdot\cdot} - \overline{y})^2$
Suma de cuadrados entre columnas:	Q_B	$= ac\sum_j (y_{\cdot j\cdot} - \overline{y})^2$
Suma de cuadrados de la interacción:	Q_{AB}	$= c\sum_{i,j} (y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + \overline{y})^2$
Suma de cuadrados residual:	Q_R	$= \sum_{i,jh} (y_{ijh} - y_{i\cdot\cdot})^2$
Suma de cuadrados total:	Q_T	$= \sum_{i,j} (y_{ijh} - \overline{y})^2$

Se verifica la siguiente identidad

$$Q_T = Q_A + Q_B + Q_{AB} + Q_R.$$

Las hipótesis de interés son:

$$\begin{array}{ll} H_0^A: & \alpha_1 = \cdots = \alpha_a = 0 \quad \text{(no hay efecto fila)} \\ H_0^B: & \beta_1 = \cdots = \beta_b = 0 \quad \text{(no hay efecto columna)} \\ H_0^{AB}: & \gamma_{11} = \cdots = \gamma_{ab} = 0 \quad \text{(no hay interacción)} \end{array}$$

Como en los casos anteriores, podemos ver que la aceptación o rechazo de las hipótesis se decide mediante el test F:

$$\begin{aligned} F_A &= \frac{Q_A}{Q_R}\frac{ab(c-1)}{a-1} && \sim F^{a-1}_{ab(c-1)} \\ F_B &= \frac{Q_B}{Q_R}\frac{ab(c-1)}{b-1} && \sim F^{b-1}_{ab(c-1)} \\ F_{AB} &= \frac{Q_{AB}}{Q_R}\frac{ab(c-1)}{(a-1)(b-1)} && \sim F^{(a-1)(b-1)}_{ab(c-1)} \end{aligned}$$

13.4 Diseños multifactoriales

Los diseños de dos factores se generalizan a un número mayor de factores. Cada factor representa una causa de variabilidad que actúa sobre la variable observable. Si por ejemplo, hay 3 factores A, B, C, las observaciones son y_{ijkh}, donde i indica el nivel i-ésimo de A, j indica el nivel j-ésimo de B, k indica el nivel k-ésimo de C, y h indica la réplica h para la combinación ijk de los tres factores, que pueden interactuar. Un modelo típico es

$$y_{ijkh} = \mu + \alpha^A_i + \alpha^B_j + \alpha^C_k + \alpha^{AB}_{ij} + \alpha^{AC}_{ik} + \alpha^{BC}_{jk} + \alpha^{ABC}_{ijk} + e_{ijkh},$$

siendo:

$$\begin{aligned} \mu &= \text{media general}, \\ \alpha^A_i, \alpha^B_j, \alpha^C_k &= \text{efectos principales de A,B,C}, \\ \alpha^{AB}_{ij}, \alpha^{AC}_{ik}, \alpha^{BC}_{jk} &= \text{interacciones entre A y B, A y C, B y C}, \\ \alpha^{ABC}_{ijk} &= \text{interacción entre A,B y C}, \\ e_{ijkh} &= \text{desviación aleatoria } N(0,\sigma^2). \end{aligned}$$

Son hipótesis de interés: $H^A_0 : \alpha^A_i = 0$ (el efecto principal de A no es significativo), $H^{AB}_0 : \alpha^{AB}_i = 0$ (la interacción entre A y B no es significativa), etc. Los tests para aceptar o no estas hipótesis se obtienen descomponiendo la variabilidad total en sumas de cuadrados

$$\sum_{i,j,k,h} (y_{ikjh} - \overline{y})^2 = A + B + C + AB + AC + BC + ABC + R,$$

donde R es el residuo. Si los factores tienen a, b, c niveles, respectivamente, y hay d réplicas para cada combinación de los niveles, entonces A tiene $(a-1)$

g.l., AB tiene $(a-1)(b-1)$ g.l. Si interpretamos las réplicas como un factor D, el residuo es

$$R = D + AD + BD + CD + ABD + ACD + BCD + ABCD$$

con

$$q = (d-1) + (a-1)(d-1) \ldots + (a-1)(b-1)(c-1)(d-1) = abc(d-1)$$

g.l. Entonces calcularemos los cocientes F

$$F = \frac{A/(a-1)}{R/q}, \quad F = \frac{AB/(a-1)(b-1)}{R/q},$$

que sirven para aceptar o rechazar H_0^A y H_0^{AB} , respectivamente.

En determinadas situaciones experimentales puede suceder que algunos factoros no interactúen. Entonces las sumas de cuadrados correspondientes se suman al residuo. Por ejemplo, si C no interactúa con A,B, el modelo es

$$y_{ijkh} = \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + \alpha_{ij}^{AB} + e_{ijkh}$$

y la descomposición de la suma de cuadrados es

$$\sum_{i,j,k,h} (y_{ikjh} - \overline{y})^2 = A + B + C + AB + R',$$

donde $R' = AC + BC + ABC + R$ es el nuevo residuo con g.l.

$$q' = (a-1)(c-1) + (b-1)(c-1) + (a-1)(b-1)(c-1) + q.$$

Los cocientes F para las hipótesis anteriores son ahora

$$F = \frac{A/(a-1)}{R'/q'}, \quad F = \frac{AB/(a-1)(b-1)}{R'/q'}.$$

13.5 Modelos log-lineales

Supongamos que tenemos dos variables categóricas A,B con a, b categorías respectivamente, y hemos observado las ab categorias $n = \sum_{ij} f_{ij}$ veces,

donde f_{ij} es el número de veces en que apareció la intersección $\mathrm{A}_i \cap \mathrm{B}_j$, es decir, tenemos la tabla de contingencia $a \times b$:

	B_1	B_2	$\cdots$	B_b	
A_1	f_{11}	f_{12}	$\cdots$	f_{1b}	$f_{1\cdot}$
A_2	f_{21}	f_{22}	$\cdots$	f_{2b}	$f_{2\cdot}$
$\vdots$			$\ddots$		
A_a	f_{a1}	f_{a2}	$\cdots$	f_{ab}	$f_{a\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot b}$	n

donde $f_{i\cdot} = \sum_j f_{ij}, f_{\cdot j} = \sum_i f_{ij}$ son las frecuencias de A_i,B_j respectivamente. Indiquemos las probabilidades

$$p_{ij} = P(\mathrm{A}_i \cap \mathrm{B}_j), \quad p_{i\cdot} = P(\mathrm{A}_i), \quad p_{\cdot j} = P(\mathrm{B}_j).$$

Existe independencia estocástica entre A y B si $p_{ij} = p_{i\cdot}p_{\cdot j}$, es decir, si

$$\ln p_{ij} = \ln p_{i\cdot} + \ln p_{\cdot j}.$$

Si introducimos las frecuencias teóricas

$$F_{ij} = np_{ij}, \quad F_{i\cdot} = np_{i\cdot}, \quad F_{\cdot j} = np_{\cdot j},$$

la condición de independencia es

$$\ln F_{ij} = \ln F_{i\cdot} + \ln F_{\cdot j} - \ln n,$$

que podemos escribir como

$$\ln F_{ij} = \lambda + \lambda_i^A + \lambda_j^B, \tag{13.1}$$

siendo

$$\begin{aligned} \lambda &= (\textstyle\sum_{i=1}^{a} \sum_{j=1}^{b} \ln F_{ij})/ab, \\ \lambda_i^A &= (\textstyle\sum_{j=1}^{b} \ln F_{ij})/b - \lambda, \\ \lambda_j^B &= (\textstyle\sum_{i=1}^{a} \ln F_{ij})/a - \lambda. \end{aligned}$$

El modelo (13.1) es un ejemplo de *modelo log-lineal.*

Generalmente no podemos aceptar la independencia estocástica. Por tanto, hemos de añadir un término a (13.1) y escribir

$$\ln F_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB},$$

donde $\lambda_{ij}^{AB} = \ln F_{ij} - \lambda - \lambda_i^A - \lambda_j^B$ es la desviación del modelo lineal. La similitud con el modelo anova de dos factores es clara.

En las aplicaciones no conocemos las frecuencias esperadas F_{ij}, sino las frecuencias observadas f_{ij}. Entonces la estimación de los parámetros es muy semejante al modelo anova, pero los tests de hipótesis se resuelven mediante ji-cuadrados.

La hipótesis de interés es la independencia entre A,B

$$H_0 : \lambda_{ij}^{AB} = 0,$$

que equivale a decir que los datos se ajustan al modelo (13.1). Sean

$$\widehat{F}_{ij} = n f_{i\cdot} \times f_{\cdot j}$$

las estimaciones máximo-verosímiles de las frecuencias esperadas. El test ji-cuadrado clásico consiste en calcular

$$\sum_{i,j} (f_{ij} - \widehat{F}_{ij})^2 / \widehat{F}_{ij}$$

y el test de la razón de verosimilitud se basa en

$$2 \sum_{i,j} f_{ij} \log(f_{ij} / \widehat{F}_{ij}),$$

que también sigue la distribución ji-cuadrado con $(a-1)(b-1)$ g.l.

El tratamiento de 3 variables categóricas A, B, C es semejante. Partiendo de una tabla de contingencia $a \times b \times c$, puede interesar saber si A, B, C son mútuamente independientes

$$\ln F_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C,$$

si hay dependencia entre A y B, entre A y C, entre B y C

$$\ln F_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC},$$

si además hay dependencia entre A, B, C

$$\ln F_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC},$$

y si A es independiente de B, C, que son dependientes, el modelo es

$$\ln F_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{jk}^{BC}.$$

En cada caso, el test ji-cuadrado o el de razón de verosimilitud nos permiten decidir si los datos se ajustan al modelo. Conviene observar que obtendríamos $\chi^2 = 0$ en el tercer modelo, ya que los datos se ajustan exactamente al modelo.

Ejemplo. Las frecuencias de supervivientes, clasificadas por género (A), supervivencia (B) y clase (C), del hundimiento del vapor Titanic son:

Género	Sobrevivió	1	2	3
Hombre	SI	118	154	422
Mujer		4	13	106
Hombre	NO	62	25	88
Mujer		141	93	90

Los resultados del análisis log-lineal son:

Modelo para $\ln F_{ijk}$	Interpretación	Símbolo	χ^2	g.l.
$\phi = \lambda + \lambda_i^G + \lambda_j^S + \lambda_k^C$	G⊥S⊥C	[G][S][C]	540.7	7
$\phi + \lambda_{ij}^{GS} + \lambda_{ik}^{GC} + \lambda_{jk}^{SC}$	GS⊥GC⊥SC	[GS][GC][SC]	61.5	2
$\phi + \lambda_{ij}^{GS} + \lambda_{ik}^{GC} + \lambda_{jk}^{SC} + \lambda_{ijk}^{GSC}$	dependencia	[GSC]	0	-
$\phi + \lambda_{jk}^{GC}$	S⊥GC	[S][GC]	511.1	5

Salvo el modelo de dependencia completa, ningún modelo se ajusta a los datos. Hemos de aceptar que la supervivencia dependía del género y la clase.

13.6 Complementos

El Análisis de la Varianza fue introducido por R. A. Fisher en 1938, para resolver problemas de diseño experimental en agricultura. Hemos visto que es una aplicación del modelo lineal. Existen muchos diseños diferentes, cuyo estudio dejamos para otro momento.

Los primeros estudios y aplicaciones consideraban factores de efectos fijos. En 1947, C. Eisenhart consideró que algunos efectos podían ser aleatorios. Ciertamente, los efectos que actúan sobre los modelos pueden ser fijos, aleatorios o mixtos, y cuando hay interacciones el cálculo de los cocientes F es diferente. Ver Cuadras (2000), Peña (1989).

Capítulo 14

ANÁLISIS DE LA VARIANZA (MANOVA)

14.1 Modelo

El análisis multivariante de la varianza (MANOVA) es una generalización en $p > 1$ variables del análisis de la varianza (ANOVA).

Supongamos que tenemos n observaciones independientes de p variables observables $Y_1, \ldots, Y_p$, obtenidas en diversas condiciones experimentales, como en el caso univariante. La matriz de datos es

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = [\widetilde{\mathbf{y}}_1, \widetilde{\mathbf{y}}_2, \ldots, \widetilde{\mathbf{y}}_p],$$

donde $\widetilde{\mathbf{y}}_j = (y_{1j}, y_{2j}, \ldots, y_{nj})'$ son las n observaciones de la variable Y_j, que suponemos siguen un modelo lineal $\widetilde{\mathbf{y}}_j = \mathbf{X}\boldsymbol{\beta}_j + \mathbf{e}_j$.

El modelo lineal multivariante es

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{14.1}$$

siendo

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

la matriz de diseño,

$$\mathbf{B} = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{m1} & \beta_{m2} & \cdots & \beta_{mp} \end{pmatrix}$$

la matriz de parámetros de regresión,

$$\mathbf{E} = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{np} \end{pmatrix}$$

la matriz de desviaciones aleatorias. Las matrices $\mathbf{Y}$ y $\mathbf{X}$ son conocidas. Suponemos que las filas de $\mathbf{E}$ son independientes $N_p(\mathbf{0},\Sigma)$.

14.2 Estimación

En el modelo MANOVA hemos de estimar los mp parámetros de regresión y la matriz de covarianzas Σ.

En el modelo univariante $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, la estimación LS $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$ minimiza $\widehat{\mathbf{e}}'\widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$. En el caso multivariante, el estimador LS de $\mathbf{B}$ es $\widehat{\mathbf{B}}$ tal que minimiza la traza

$$\mathrm{tr}(\widehat{\mathbf{E}}'\widehat{\mathbf{E}}) = tr((\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})).$$

Se demuestra que:

1. Las estimaciones LS de los parámetros de regresión $\mathbf{B}$ son

   $$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

   cuando el diseño es de rango máximo $r = \mathrm{rang}(\mathbf{X}) = m$, y

   $$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$$

   cuando $r < m$.

2. La matriz de residuos es la matriz $\mathbf{R}_0 = (R_0(i,j))$ de orden $p \times p$

$$\mathbf{R}_0 = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}),$$

donde $R_0(j,j)$ es la suma de cuadrados residual del modelo univariante $\mathbf{y}_j = \mathbf{X}\boldsymbol{\beta}_j + \mathbf{e}_j$.

3. Una estimación centrada de la matriz de covarianzas Σ es

$$\widehat{\Sigma} = \mathbf{R}_0/(n-r).$$

Teorema 14.2.1 *Sea* $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ *el model lineal multivariante donde las filas de* $\mathbf{E}$ *son* $N_p(\mathbf{0},\Sigma)$ *independientes. Sea* $\mathbf{R}_0$ *la matriz de residuos. Se verifica:*

1. $\mathbf{R}_0 = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}']\mathbf{Y}$.
2. *La distribución de* $\mathbf{R}_0$ *es Wishart* $W_p(\Sigma, n-r)$.

Demost.: Sea $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_r, \mathbf{t}_{r+1}, \ldots, \mathbf{t}_n]$ una matriz ortogonal tal que sus columnas formen una base ortonormal de R^n, de manera que las r primeras generen el subespacio $C_r(\mathbf{X})$ y por tanto las otras $n-r$ sean ortogonales a $C_r(\mathbf{X})$. Consideremos $\widehat{\mathbf{E}} = \mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}$ y definamos $\mathbf{Z} = \mathbf{T}'\mathbf{Y}$. Como en el modelo lineal (ver Teorema 12.3.3), se verifica

$$\mathbf{T}'\widehat{\mathbf{E}} = \begin{bmatrix} \mathbf{0} \\ \mathbf{Z}_{n-r} \end{bmatrix},$$

donde $\mathbf{Z}_{n-r}$ es una matriz $(n-r) \times p$ con filas $N_p(\mathbf{0},\Sigma)$ independientes. En consecuencia

$$\mathbf{R}_0^2 = \widehat{\mathbf{E}}'\widehat{\mathbf{E}} = \widehat{\mathbf{E}}'\mathbf{T}\mathbf{T}'\widehat{\mathbf{E}} = \mathbf{Z}'_{n-r}\mathbf{Z}_{n-r} \sim W_p(\Sigma, n-r). \square$$

14.3 Tests de hipótesis lineales

Una hipótesis lineal demostrable de rango t y matriz $\mathbf{H}$ es

$$H_0 : \mathbf{HB} = \mathbf{0}$$

donde las filas de $\mathbf{H}$ son combinación lineal de las filas de $\mathbf{X}$.

Como en el caso univariante (Sección 12.5), si H_0 es cierta, el modelo se transforma en

$$\mathbf{Y} = \widetilde{\mathbf{X}}\Theta + \mathbf{E},$$

la estimación de los parámetros $\mathbf{B}$ restringidos a H_0 viene dada por

$$\widehat{\mathbf{B}}_H = \widehat{\mathbf{B}} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{H}'(\mathbf{H}(\mathbf{X}'\mathbf{X})^{-}\mathbf{H}')^{-1}\mathbf{H}\widehat{\mathbf{B}}$$

y la matriz residual es

$$\mathbf{R}_1 = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}_H)'(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}_H).$$

Teorema 14.3.1 *Sea* $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ *el modelo lineal multivariante, donde las filas de* $\mathbf{E}$ *son* $N_p(\mathbf{0},\Sigma)$ *independientes,* $\mathbf{R}_0$ *la matriz de residuos,* $H_0 : \mathbf{HB} = \mathbf{0}$ *una hipótesis lineal demostrable y* $\mathbf{R}_1$ *la matriz de residuos bajo* H_0. *Se verifica:*

1. $\mathbf{R}_0 \sim W_p(\Sigma, n - r)$.
2. *Si* H_0 *es cierta,*

$$\mathbf{R}_1 \sim W_p(\Sigma, n - r'), \quad \mathbf{R}_1 - \mathbf{R}_0 \sim W_p(\Sigma, t),$$

   siendo $t = ran(H), r' = r - t$.

3. *Si* H_0 *es cierta, las matrices* $\mathbf{R}_0$ *y* $\mathbf{R}_1 - \mathbf{R}_0$ *son estocásticamente independientes.*

Demost.: Si la hipótesis H_0 es cierta, la matriz de diseño $\mathbf{X}$ se transforma en $\mathbf{XC}$, donde las columnas de $\mathbf{XC}$ son combinación lineal de las columnas de $\mathbf{X}$. Podemos encontrar una matriz ortogonal

$$\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_{r'}, \mathbf{t}_{r'+1}, \ldots, \mathbf{t}_r, \mathbf{t}_{r+1}, \ldots, \mathbf{t}_n]$$

tal que

$$C_{r'}(\mathbf{XC}) = [\mathbf{t}_1, \ldots, \mathbf{t}_{r'}] \subset C_r(\mathbf{X}) = [\mathbf{t}_1, \ldots, \mathbf{t}_r].$$

Siguiendo los mismos argumentos del teorema anterior, tenemos que

$$\mathbf{T}'\widehat{\mathbf{E}} = \begin{bmatrix} \mathbf{0} \\ \mathbf{Z}_{n-r'} \end{bmatrix}$$

donde las $n - r'$ filas de $\mathbf{Z}_{n-r'}$ son $N_p(\mathbf{0},\Sigma)$ independientes. Por tanto $\mathbf{R}_1^2 = \mathbf{Z}'_{n-r'}\mathbf{Z}_{n-r'}$ es Wishart $W_p(\Sigma, n-r')$. Por otro lado podemos escribir

$$\mathbf{Z}_{n-r'} = \begin{bmatrix} \mathbf{Z}_t \\ \mathbf{Z}_{n-r} \end{bmatrix}$$

donde las $t = r - r'$ filas de $\mathbf{Z}_t$ son independientes de las $n - r$ filas de $\mathbf{Z}_{n-r}$. Entonces es fácil ver que

$$\mathbf{R}_1^2 - \mathbf{R}_0^2 = \mathbf{Z}'_t\mathbf{Z}_t,$$

es decir, $\mathbf{R}_1^2 - \mathbf{R}_0^2$ es Wishart $W_p(\Sigma, n-r')$ e independiente de $\mathbf{R}_0$.□

La consecuencia más importante de este teorema es que, si H_0 es cierta, entonces

$$\Lambda = \frac{|\mathbf{R}_0|}{|(\mathbf{R}_1 - \mathbf{R}_0) + \mathbf{R}_0|} = \frac{|\mathbf{R}_0|}{|\mathbf{R}_1|} \sim \Lambda(p, n-r, t),$$

es decir, $0 \leq \Lambda \leq 1$ sigue la distribución de Wilks. Aceptaremos H_0 si Λ no es significativo y rechazaremos H_0 si Λ es pequeño y significativo.

Tabla general MANOVA

	g. l.	matriz Wishart	lambda de Wilks
Desviación hipótesis	t	$\mathbf{R}_1 - \mathbf{R}_0$	$\Lambda = \lvert\mathbf{R}_0\rvert/\lvert\mathbf{R}_1\rvert$
Residuo	$n - r$	$\mathbf{R}_0$	

Criterio decisión: Si $\Lambda < \Lambda_\alpha$ es rechazada H_0, donde $P(\Lambda(p, n-r, t) < \Lambda_\alpha) = \alpha$.

14.4 Manova de un factor

El modelo del diseño de un único factor o causa de variabilidad es

$$\mathbf{y}_{ih} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \mathbf{e}_{ih}, \quad i = 1, \ldots, k;\ h = 1, \ldots, n_i,$$

donde $\boldsymbol{\mu}$ es un vector de medias general, $\boldsymbol{\alpha}_i$ es el efecto del nivel y del factor, $\mathbf{y}_{ih}$ es la observación multivariante h en la situación (o población) i, correspondiendo a la misma situación experimental del análisis canónico de poblaciones (Capítulo 7), con $n = n_1 + \ldots + n_k$. Por tanto

$$\mathbf{W} = \mathbf{R}_0, \quad \mathbf{B} = \mathbf{R}_1 - \mathbf{R}_0, \quad \mathbf{T} = \mathbf{R}_1 = \mathbf{B} + \mathbf{W},$$

son las matrices de dispersión "dentro grupos", "entre grupos" y "total", respectivamente (Sección 3.3.3).

MANOVA de un factor			
	g. l.	matriz Wishart	lambda de Wilks
Entre grupos	$k-1$	$\mathbf{B}$	$\Lambda = \lvert\mathbf{W}\rvert/\lvert\mathbf{W}+\mathbf{B}\rvert$
Dentro grupos	$n-k$	$\mathbf{W}$	$\sim \Lambda(p, n-k, k-1)$
Total	$n-1$	$\mathbf{T}$	

14.5 Manova de dos factores

Si suponemos que las $n = a \times b$ observaciones multivariantes dependen de dos factores fila y columna, con a y b niveles respectivamente, el modelo es

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \mathbf{e}_{ij}, \quad i = 1, \ldots, a; j = 1, \ldots, b,$$

donde $\boldsymbol{\mu}$ es la media general, $\boldsymbol{\alpha}_i$ es el efecto aditivo del nivel i del factor fila, $\boldsymbol{\beta}_j$ es el efecto aditivo del nivel j del factor columna. Como generalización del caso univariante, intervienen las matrices $\mathbf{A} = (a_{uv}), \mathbf{B} = (b_{uv}), \mathbf{T} = (t_{uv}), \mathbf{R}_0 = (r_{uv})$ con elementos

$$\begin{aligned}
a_{uv} &= a \textstyle\sum_j (y_{\cdot ju} - \overline{y}_u)(y_{\cdot jv} - \overline{y}_v) \\
b_{uv} &= b \textstyle\sum_i (y_{i\cdot u} - \overline{y}_u)(y_{i\cdot v} - \overline{y}_v) \\
r_{uv} &= \textstyle\sum_{ij} (y_{iju} - y_{i\cdot u} - y_{\cdot ju} + \overline{y}_u)(y_{ijv} - y_{i\cdot v} - y_{\cdot jv} + \overline{y}_v) \\
t_{uv} &= \textstyle\sum_{ij} (y_{iju} - \overline{y}_u)(y_{ijv} - \overline{y}_v), \qquad u, v = 1, \ldots, p,
\end{aligned}$$

siendo, para cada variable Y_u, $\overline{y}_u$ la media, $y_{\cdot ju}$ la media fijando el nivel j del factor columna, etc. Se verifica

$$\mathbf{T} = \mathbf{A} + \mathbf{B} + \mathbf{R}_0.$$

Indicando $q = (a-1)(b-1)$, obtenemos la tabla

MANOVA de dos factores			
	g. l.	matriz Wishart	lambda de Wilks
Filas	$a-1$	$\mathbf{A}$	$\lvert\mathbf{A}\rvert/\lvert\mathbf{T}\rvert \sim \Lambda(p, q, a-1)$
Columnas	$b-1$	$\mathbf{B}$	$\lvert\mathbf{B}\rvert/\lvert\mathbf{T}\rvert \sim \Lambda(p, q, b-1)$
Residuo	q	$\mathbf{R}_0$	
Total	$ab-1$	$\mathbf{T}$	

14.6 Manova de dos factores con interacción

En el diseño de dos factores con interacción suponemos que las $n = a \times b \times c$ observaciones multivariantes dependen de dos factores fila y columna, con a y b niveles respectivamente, y que hay c observaciones (réplicas) para cada una de las $a \times b$ combinaciones de los niveles. El modelo lineal es

$$\mathbf{y}_{ijh} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \mathbf{e}_{ijh}, \quad i = 1, \ldots, a; j = 1, \ldots, b; h = 1, \ldots, c,$$

donde $\boldsymbol{\mu}$ es la media general, $\boldsymbol{\alpha}_i$ es el efecto aditivo del nivel y del factor fila, $\boldsymbol{\beta}_j$ es el efecto aditivo del nivel j del factor columna, $\boldsymbol{\gamma}_{ij}$ es la interacción, parámetro que mide la desviación de la aditividad del efecto de los factores, e $\mathbf{y}_{ijh} = (y_{ijh1}, \ldots, y_{ijhp})'$ es la réplica multivariante h de las variables observables. También, como en el caso univariante, intervienen las matrices $\mathbf{A} = (a_{uv}), \mathbf{B} = (b_{uv}), \mathbf{AB} = (c_{uv}), \mathbf{R}_0 = (r_{uv}), \mathbf{T} = (t_{uv})$, donde

$$\begin{array}{ll} a_{uv} & = bc \sum_i (y_{i\cdot\cdot u} - \overline{y}_u)(y_{i\cdot\cdot v} - \overline{y}_v) \\ b_{uv} & = ac \sum_j (y_{\cdot j\cdot u} - \overline{y}_u)(y_{\cdot j\cdot v} - \overline{y}_v) \\ c_{uv} & = c \sum_{i,j} (y_{ij\cdot u} - y_{i\cdot\cdot u} - y_{\cdot j\cdot v} + \overline{y}_u)(y_{ij\cdot v} - y_{i\cdot\cdot v} - y_{\cdot j\cdot v} + \overline{y}_v) \\ r_{uv} & = \sum_{i,jh} (y_{ijhu} - y_{i\cdot\cdot u})(y_{ijhv} - y_{i\cdot\cdot v}) \\ t_{uv} & = \sum_{i,j} (y_{iju} - \overline{y}_u)(y_{iju} - \overline{y}_u), \qquad u, v = 1, \ldots, p, \end{array}$$

que verifican

$$\mathbf{T} = \mathbf{A} + \mathbf{B} + \mathbf{AB} + \mathbf{R}_0.$$

Obtenemos la tabla:

MANOVA de dos factores con interacción

	g. l.	matriz Wishart	lambda de Wilks
Filas	$a-1$	$\mathbf{A}$	$\lvert\mathbf{A}\rvert/\lvert\mathbf{T}\rvert \sim \Lambda(p, r, a-1)$
Columnas	$b-1$	$\mathbf{B}$	$\lvert\mathbf{B}\rvert/\lvert\mathbf{T}\rvert \sim \Lambda(p, r, b-1)$
Interacción	$(a-1)(b-1) = q$	$\mathbf{AB}$	$\lvert\mathbf{AB}\rvert/\lvert\mathbf{T}\rvert \sim \Lambda(p, r, q)$
Residuo	$ab(c-1) = r$	$\mathbf{R}_0$	
Total	$abc-1$	$\mathbf{T}$	

14.7 Ejemplos

Exemple 14.7.1 *Ratas experimentales.*

En un experimento para inhibir un tumor, se quiere investigar el efecto del sexo (S) y de la temperatura ambiental (T). Se consideran las variables:
Y_1 =peso inicial, Y_2 =peso final, Y_3 =peso del tumor.

	Machos			Hembras		
Temp	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
4	18.15	16.51	0.24	19.15	19.49	0.16
	18.68	19.50	0.32	18.35	19.81	0.17
	19.54	19.84	0.20	20.58	19.44	0.22
20	21.27	23.30	0.33	18.87	22.00	0.25
	19.57	22.30	0.45	20.66	21.08	0.20
	20.15	18.95	0.35	21.56	20.34	0.20
34	20.74	16.69	0.31	20.22	19.00	0.18
	20.02	19.26	0.41	18.38	17.92	0.30
	17.20	15.90	0.28	20.85	19.90	0.17

Los resultados MANOVA son:

	g. l.	matriz dispersión	lambda	F	g.l.
T	2	$\begin{pmatrix} 4.81 & 9.66 & .284 \\ & 32.5 & .376 \\ & & .019 \end{pmatrix}$	.261	3.18	6,20
S	1	$\begin{pmatrix} .642 & 1.27 & -.19 \\ & 2.51 & -.38 \\ & & .006 \end{pmatrix}$	.337	6.55	3,10
T×S	2	$\begin{pmatrix} .275 & .816 & .038 \\ & 32.5 & .088 \\ & & .006 \end{pmatrix}$	.772	0.46	6,20
Residuo	12	$\begin{pmatrix} 19.3 & 7.01 & -.19 \\ & 26.7 & .208 \\ & & .039 \end{pmatrix}$			
Total	17	$\begin{pmatrix} 25.0 & 18.7 & -.06 \\ & 32.5 & .284 \\ & & .125 \end{pmatrix}$			

Son significativos los efectos S y T, pero la interacción no es significativa. Una representación canónica de los $3 \times 2 = 6$ grupos (Figura 14.1) ayuda a visualizar las diferencias. Podemos ver que la pequeña diferencia entre la representación de las tres temperatures de los machos y de las hembras es indicio de una cierta interacción, aunque no significativa.

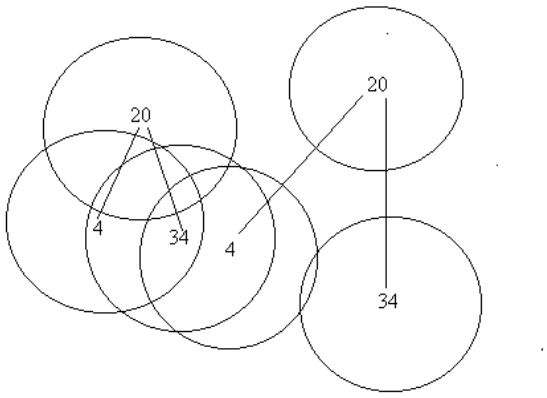

Figura 14.1: Representación canónica de los datos de las ratas hembras (izquierda) y machos (derecha).

Exemple 14.7.2 *Coleópteros.*

Continuando con el ejemplo 7.5.1, vamos a estudiar 8 especies (factor E) de coleópteros del género Timarcha, pero teniendo en cuenta el sexo, machos y hembras (factor S), en relación a 5 variables biométricas.

Las matrices de dispersión entre especies, entre sexos, debidas a la interacción, residual y los estadísticos Λ y F son:

$$\text{E}=\begin{pmatrix} 14303 & 24628 & 17137 & 48484 & 36308 \\ & 43734 & 31396 & 85980 & 64521 \\ & & 23610 & 61519 & 46405 \\ & & & 169920 & 126980 \\ & & & & 95395 \end{pmatrix} \quad \begin{matrix} \Lambda = .0068 \\ F_{35,2353} = 152.8 \end{matrix}$$

$$\text{S}=\begin{pmatrix} 675.94 & 1613.0 & 1644.5 & 4520.0 & 3270.6 \\ & 3849.3 & 3924.4 & 10786. & 7804.9 \\ & & 4001.0 & 10997. & 7957.2 \\ & & & 30225. & 21871. \\ & & & & 15825. \end{pmatrix} \quad \begin{matrix} \Lambda = .1944 \\ F_{5,559} = 463.2 \end{matrix}$$

$$\text{E}\times\text{S}=\begin{pmatrix} 96.470 & 81.532 & 63.559 & 92.035 & 20.554 \\ & 97.205 & 85.554 & 157.28 & 102.31 \\ & & 86.405 & 127.66 & 108.25 \\ & & & 428.97 & 236.53 \\ & & & & 282.30 \end{pmatrix} \quad \begin{matrix} \Lambda = .7692 \\ F_{35,2353} = 4.329 \end{matrix}$$

$$\mathrm{R}_0=\begin{pmatrix} 1546.7 & 1487.8 & 1346.4 & 2452.6 & 1924.0 \\ & 3498.5 & 3078.4 & 4206.6 & 3415.6 \\ & & 3082.9 & 3888.2 & 3159.4 \\ & & & 9178.6 & 6038.0 \\ & & & & 5950.3 \end{pmatrix}$$

14.8 Otros criterios

Sea $\lambda_1 \geq \ldots \geq \lambda_p$ los valores propios de $\mathbf{R}_0$ respecto de $\mathbf{R}_1$. Podemos expresar el criterio de Wilks como

$$\Lambda = \frac{|\mathbf{R}_0|}{|\mathbf{R}_1|} = \lambda_1 \times \ldots \times \lambda_p.$$

Este criterio es especialmente interesante, teniendo en cuenta que si λ es la razón de verosimilitud en el test de hipótesis, entonces $\lambda = \Lambda^{n/2}$.

Se demuestra que cualquier estadístico que sea invariante por cambios de origen y de escala de los datos, debe ser función de estos valores propios. Así otros tests propuestos son:

1. Traza de Hotelling:

$$\mathrm{tr}((\mathbf{R}_1 - \mathbf{R}_0)\mathbf{R}_o^{-1}) = \sum_{i=1}^{p} \frac{1-\lambda_i}{\lambda_i}.$$

2. Traza de Pillai:

$$\mathrm{tr}((\mathbf{R}_1 - \mathbf{R}_0)\mathbf{R}_1^{-1}) = \sum_{i=1}^{p} 1-\lambda_i.$$

3. Raíz mayor de Roy: $(1-\lambda_p)/\lambda_p$.

En el ejemplo 14.7.2, para contrastar las diferencias entre localidades, obtenemos los siguientes valores de los estadísticos de Wilks, Hotelling, Pillai y Roy, y sus transformaciones a una F:

		F	g.l.	g.l.
Wilks	0.007	152.8	35	2354
Hotelling	28.02	446.2	35	2787
Pillai	2.090	57.78	35	2815
Roy	24.90	2002	7	563

14.9 Complementos

El Análisis Multivariante de la Variancia es muy similar al Análisis de la Variancia, sólo que interviene más de una variable cuantitativa observable. Esta extensión multivariante se inicia en 1930 con los trabajos de H. Hotelling, J. Wishart y S.S. Wilks. Posteriormente S.N. Roy propuso un planteo basado en el principio de unión-intersección.

Los cuatro criterios que hemos visto son equivalentes para $p = 1$, y diferentes para $p > 1$. No está claro cual es el mejor criterio, depende de la hipótesis alternativa. Por ejemplo, en el diseño de un factor, si los vectores de medias están prácticamente alineados, entonces el criterio de Roy es el más potente. Ver Rencher (1998).

Capítulo 15

FUNCIONES ESTIMABLES MULTIVARIANTES

15.1 Funciones estimables

En el modelo lineal univariante $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, además de la estimación de los parámetros de regresión $\boldsymbol{\beta}$, tiene también interés la estimación de ciertas combinaciones lineales de los parámetros $\boldsymbol{\beta}$.

Definició 15.1.1 *Una función paramétrica ψ es una combinación lineal de los parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$*

$$\psi = p_1\beta_1 + \cdots + p_m\beta_m = \mathbf{p}'\boldsymbol{\beta},$$

donde $\mathbf{p} = (p_1, \ldots, p_m)'$. Una función paramétrica ψ es estimable si existe una combinación lineal $\widehat{\psi}$ de $\mathbf{y} = (y_1, \ldots, y_n)'$

$$\widehat{\psi} = a_1y_1 + \cdots + a_ny_n = \mathbf{a}'\mathbf{y},$$

donde $\mathbf{a} = (a_1, \ldots, a_n)'$, tal que

$$E(\widehat{\psi}) = \psi.$$

La caracterización de que una función paramétrica ψ es estimable es la siguiente

Proposició 15.1.1 *Una función paramétrica $\psi = \mathbf{p}'\boldsymbol{\beta}$ es estimable si y sólo si el vector fila $\mathbf{p}'$ es combinación lineal de las filas de la matriz de diseño $\mathbf{X}$.*

Demost.: $E(\widehat{\psi}) = E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'E(\mathbf{y}) = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \mathbf{p}'\boldsymbol{\beta}$, que vale para todo $\boldsymbol{\beta}$. Por lo tanto $\mathbf{a}'\mathbf{X} = \mathbf{p}'$, es decir, $\mathbf{p}'$ es combinación lineal de las filas de $\mathbf{X}$.□

15.2 Teorema de Gauss-Markov

La estimación óptima de una función paramétrica estimable $\psi = \mathbf{p}'\boldsymbol{\beta}$ se obtiene sustituyendo $\boldsymbol{\beta}$ por la estimación LS $\widehat{\boldsymbol{\beta}}$. Esto es el famoso teorema de Gauss-Markov.

Teorema 15.2.1 *Sea $\psi = \mathbf{p}'\boldsymbol{\beta}$ una función paramétrica estimable. Se verifica:*

1. *Si $\widehat{\boldsymbol{\beta}}$ es estimador LS de $\boldsymbol{\beta}$, entonces $\widehat{\psi} = \mathbf{p}'\widehat{\boldsymbol{\beta}}$ es único.*
2. *$\widehat{\psi} = \mathbf{p}'\widehat{\boldsymbol{\beta}}$ es estimador lineal insesgado de ψ y, dentro de los estimadores lineales insesgados de ψ, tiene varianza mínima.*

Demost.: Existe un estimador insesgado $\widehat{\psi} = \mathbf{a}'\mathbf{y}$ de $\psi = \mathbf{p}'\boldsymbol{\beta}$. Sea $C_r(\mathbf{X})$ el subespacio generado por las columnas de $\mathbf{X}$. Entonces $\mathbf{a} = \widetilde{\mathbf{a}} + \mathbf{b}$, donde $\widetilde{\mathbf{a}} \in C_r(\mathbf{X})$ y $\mathbf{b}$ es ortogonal a $C_r(\mathbf{X})$. Consideremos al estimador $\widetilde{\mathbf{a}}'\mathbf{y}$. Tenemos

$$E(\widehat{\psi}) = E(\mathbf{a}'\mathbf{y}) = E(\widetilde{\mathbf{a}}'\mathbf{y} + \mathbf{b}'\mathbf{y}) = E(\widetilde{\mathbf{a}}'\mathbf{y}) + \mathbf{b}'\mathbf{X}\boldsymbol{\beta} = E(\widetilde{\mathbf{a}}'\mathbf{y}) = \psi,$$

puesto que $\mathbf{b}'\mathbf{X} = \mathbf{0}$. Luego $\widetilde{\mathbf{a}}'\mathbf{y}$ es estimador centrado. Si $\mathbf{a}_1'\mathbf{y}$ es otro estimador centrado con $\mathbf{a}_1 \in C_r(\mathbf{X})$, entonces $E(\widetilde{\mathbf{a}}'\mathbf{y}) - E(\mathbf{a}'\mathbf{y}) = (\widetilde{\mathbf{a}}' - \mathbf{a}')\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ $\Rightarrow \widetilde{\mathbf{a}} = \mathbf{a}_1$, es decir, $\widetilde{\mathbf{a}}'\mathbf{y}$ es único.

Por otro lado, $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ es ortogonal a $C_r(\mathbf{X})$ y $\widetilde{\mathbf{a}}'\mathbf{e} = \widetilde{\mathbf{a}}'\mathbf{y} - \widetilde{\mathbf{a}}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0}$ $\Rightarrow \widetilde{\mathbf{a}}'\mathbf{y} = \widetilde{\mathbf{a}}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{p}'\widehat{\boldsymbol{\beta}}$. Así $\widehat{\psi} = \widetilde{\mathbf{a}}'\mathbf{y} = \mathbf{p}'\widehat{\boldsymbol{\beta}}$ es único y centrado.

Finalmente, indicando

$$\|\mathbf{a}\|^2 = a_1^2 + \ldots + a_n^2,$$

tenemos que

$$\mathrm{var}(\mathbf{a}'\mathbf{y}) = \|\mathbf{a}\|^2 \sigma^2 = (\|\widetilde{\mathbf{a}}\|^2 + \|\mathbf{b}\|^2)\sigma^2 \leq \|\widetilde{\mathbf{a}}\|^2 \sigma^2 = \mathrm{var}(\widetilde{\mathbf{a}}'\mathbf{y}),$$

que prueba que $\widehat{\psi} = \mathbf{p}'\widehat{\boldsymbol{\beta}}$ tiene varianza mínima.□

Un criterio para saber si $\mathbf{p}'\boldsymbol{\beta}$ es función paramétrica estimable es

$$\mathbf{p}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{p}'.$$

15.3 Funciones estimables multivariantes

En el modelo lineal multivariante (14.1), también tiene interés la estimación de ciertas combinaciones lineales de los parámetros $\mathbf{B}$. Indiquemos por $\mathbf{y}_1, \ldots, \mathbf{y}_n$ los vectores fila de $\mathbf{Y}$, y $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ los vectores fila de $\mathbf{B}$,es decir:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix}.$$

Definició 15.3.1 *Una función paramétrica multivariante $\boldsymbol{\psi}$ es una combinación lineal de las filas de $\mathbf{B}$,*

$$\boldsymbol{\psi}' = p_1\boldsymbol{\beta}_1 + \cdots + p_m\boldsymbol{\beta}_m = \mathbf{p}'\mathbf{B},$$

donde $\mathbf{p} = (p_1, \ldots, p_m)'$. Una función paramétrica multivariante $\boldsymbol{\psi}$ es estimable (fpem) si existe una combinación lineal $\widehat{\boldsymbol{\psi}}'$ de las filas de $\mathbf{Y}$

$$\widehat{\boldsymbol{\psi}}' = a_1\mathbf{y}_1 + \cdots + a_n\mathbf{y}_n = \mathbf{a}'\mathbf{Y},$$

donde $\mathbf{a} = (a_1, \ldots, a_n)'$, tal que

$$E(\widehat{\boldsymbol{\psi}}) = \boldsymbol{\psi}.$$

La caracterización de que una función paramétrica ψ es fpem es la siguiente:

Proposició 15.3.1 *Una función paramétrica $\boldsymbol{\psi}' = \mathbf{p}'\mathbf{B}$ es estimable si y sólo si el vector fila $\mathbf{p}'$ es combinación lineal de las filas de la matriz de diseño $\mathbf{X}$.*

La demostración es similar al caso univariante. La estimación óptima de una fpem $\boldsymbol{\psi}' = \mathbf{p}'\mathbf{B}$ viene dada por

$$\widehat{\boldsymbol{\psi}}' = \mathbf{p}'\widehat{\mathbf{B}}.$$

Sólo hay que sustituir $\mathbf{B}$ por sus estimaciones LS $\widehat{\mathbf{B}}$.

Teorema 15.3.2 *Sea $\boldsymbol{\psi}' = (\psi_1, \ldots, \psi_p) = \mathbf{p}'\mathbf{B}$ una función paramétrica estimable. Se verifica:*

1. *Si* $\widehat{\mathbf{B}}$ *es estimador LS de* $\mathbf{B}$*, entonces* $\widehat{\boldsymbol{\psi}}' = (\widehat{\psi}_1, \ldots, \widehat{\psi}_p) = \mathbf{p}'\widehat{\mathbf{B}}$ *es único.*

2. *Cada* $\widehat{\psi}_j$ *es estimador lineal insesgado de* ψ_j *y de varianza mínima entre los estimadores lineales insesgados de* ψ_j*.*

Observemos que este teorema vale sin necesidad de una hipótesis de normalidad. El estimador LS de $\boldsymbol{\psi}$ es

$$\widehat{\boldsymbol{\psi}}' = \mathbf{p}'\widehat{\mathbf{B}} = \mathbf{p}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} = g_1\mathbf{y}_1 + \cdots + g_n\mathbf{y}_n$$

donde $\mathbf{y}_1, \ldots, \mathbf{y}_n$ son las filas de la matriz de datos $\mathbf{Y}$. El vector $\mathbf{g} = (g_1, \ldots, g_n)'$ es único, y podemos definir la dispersión de $\widehat{\boldsymbol{\psi}}$, que es mínima, como la cantidad

$$\delta^2_{\psi} = g_1^2 + \cdots + g_n^2. \tag{15.1}$$

La versión del Teorema 14.3.1 para fpem es:

Teorema 15.3.3 *En el modelo MANOVA normal, si* $\widehat{\boldsymbol{\psi}} = \mathbf{p}'\widehat{\mathbf{B}}$ *es la estimación LS de* $\boldsymbol{\psi}$*, entonces:*

1. *La distribución de* $\widehat{\boldsymbol{\psi}}$ *es la de una combinación lineal de variables normales independientes.*

2. *La distribución de* $\mathbf{R}_0$ *es* $W_p(\Sigma, n-r)$*.*

3. $\widehat{\boldsymbol{\psi}}$ *y* $\mathbf{R}_0$ *son estocásticamente independientes.*

15.4 Análisis canónico de fpem

Supongamos que $\boldsymbol{\psi}'_1 = \mathbf{p}'_1\mathbf{B}, \ldots, \boldsymbol{\psi}'_s = \mathbf{p}'_s\mathbf{B}$ es un sistema de s fpem. Podemos plantear la representación canónica del sistema como una generalización del análisis canónico de poblaciones.

15.4.1 Distancia de Mahalanobis

Sean $\widehat{\boldsymbol{\psi}}_1, \ldots, \widehat{\boldsymbol{\psi}}_s$ las estimaciones LS de los fpem, $\widehat{\Sigma} = \mathbf{R}_0/(n-r)$ la estimación de la matriz de covarianzas. Podemos definir la distancia de Mahalanobis (estimada) entre las funciones $\boldsymbol{\psi}_i, \boldsymbol{\psi}_j$ como

$$M(i,j)^2 = (\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j)'\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j).$$

Observemos que si $\widehat{\boldsymbol{\psi}}_i' = \mathbf{g}_i'\mathbf{Y}$ es independiente de $\widehat{\boldsymbol{\psi}}_j' = \mathbf{g}_j'\mathbf{Y}$ y se verifica la hipótesis $H_0 : \boldsymbol{\psi}_i = \boldsymbol{\psi}_j$, entonces $\delta_{ij}^{-1}(\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j)$ es $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, donde $\delta_{ij} = \|\mathbf{g}_i - \mathbf{g}_j\|$, y $(n-r)\widehat{\Sigma}$ es $W_p(\Sigma, n-r)$, por lo tanto $\delta_{ij}^{-1}M(i,j)$ es Hotelling $T^2(p, n-r)$ y

$$\frac{n-r-p+1}{(n-r)p}\delta_{ij}^{-1}M(i,j)^2 \ \sim \ F^p_{n-r-p+1}.$$

Análogamente vemos que la distribución de

$$\frac{n-r-p+1}{(n-r)p}\frac{1}{\delta_\psi^2}(\widehat{\boldsymbol{\psi}}_i - \boldsymbol{\psi}_i)'\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\psi}}_i - \boldsymbol{\psi}_i)$$

es también $F^p_{n-r-p+1}$, donde δ_ψ^2 es la dispersión mínima (15.1).

15.4.2 Coordenadas canónicas

Si $\widehat{\boldsymbol{\psi}}_i = (\widehat{\psi}_{i1}, \ldots, \widehat{\psi}_{ip})', i = 1, \ldots, s$, consideremos las medias

$$\overline{\psi}_j = \frac{1}{s}\sum_{i=1}^{s}\widehat{\psi}_{ij}, \quad j = 1, \ldots, s,$$

y la matriz

$$\mathbf{U} = \begin{pmatrix} \widehat{\psi}_{11} - \overline{\psi}_1 & \cdots & \widehat{\psi}_{1p} - \overline{\psi}_p \\ \vdots & \ddots & \vdots \\ \widehat{\psi}_{s1} - \overline{\psi}_1 & \cdots & \widehat{\psi}_{sp} - \overline{\psi}_p \end{pmatrix}.$$

Sea $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ la matriz de vectores propios de $\mathbf{U}'\mathbf{U}$ respecto de $\widehat{\Sigma}$, con la normalización $\mathbf{v}_j'\widehat{\Sigma}\mathbf{v}_j = 1$, es decir,

$$\mathbf{U}'\mathbf{U}\mathbf{V} = \widehat{\Sigma}\mathbf{V}\mathbf{D}_\lambda, \quad \mathbf{V}'\widehat{\Sigma}\mathbf{V} = \mathbf{I},$$

donde $\mathbf{D}_\lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ es la matriz diagonal con los valores propios. Las coordenadas canónicas de $\widehat{\boldsymbol{\psi}}_1, \ldots, \widehat{\boldsymbol{\psi}}_s$ son las filas $\mathbf{w}_1', \ldots, \mathbf{w}_s'$ de la matriz

$$\mathbf{W} = \mathbf{U}\mathbf{V}.$$

La distancia euclídea entre las filas coincide con la distancia de Mahalanobis entre las fpem

$$(\mathbf{w}_i - \mathbf{w}_j)'(\mathbf{w}_i - \mathbf{w}_j) = (\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j)'\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j).$$

De manera análoga podemos definir la variabilidad geométrica de las fpem, probando que es

$$V_\psi = \frac{1}{2s^2}\sum_{i,j=1}^{s} M(i,j)^2 = \frac{1}{s}\sum_{i=1}^{p}\lambda_i,$$

y que es máxima en dimensión reducida q. El porcentaje de variabilidad explicada por las q primeras coordenadas canónicas es

$$P_q = 100\frac{V(\mathbf{Y})_q}{V_\psi} = 100\frac{\lambda_1+\cdots+\lambda_q}{\lambda_1+\cdots+\lambda_p}.$$

15.4.3 Regiones confidenciales

Sean $\mathbf{w}_i' = \widehat{\boldsymbol{\psi}}_i'\mathbf{V}$, $i = 1,\ldots,s$, las proyecciones canónicas de las estimaciones de las fpem. Podemos entender $\mathbf{w}_i'$ como una estimación de $\boldsymbol{\psi}_i^{*\prime} = \boldsymbol{\psi}_i'\mathbf{V}$, la proyección canónica de $\boldsymbol{\psi}_i$. Podemos también encontrar regiones confidenciales para las $\boldsymbol{\psi}_i^*, i = 1,\ldots,g$.

Sea $1-\alpha$ el coeficiente de confianza, F_α tal que $P(F > F_\alpha) = \alpha$, donde F sigue la distribución F con p y $(n-g-p+1)$ g.l., y consideremos:

$$R_\alpha^2 = F_\alpha\frac{(n-r)p}{(n-r-p+1)}.$$

Luego las proyecciones canónicas $\boldsymbol{\psi}_i^*$ de las fpem pertenecen a regiones confidenciales que son hiperesferas (esferas en dimensión 3, círculos en dimensión 2) de centros y radios

$$(\mathbf{w}_i, \delta_i R_\alpha)$$

donde δ_i es la dispersión mínima (15.1) de la estimación LS de $\boldsymbol{\psi}_i$.

15.5 Ejemplos

Ejemplo 1. Se quiere hacer una comparación de dos fármacos ansiolíticos (Diazepan y Clobazan) con un placebo, que indicaremos D, C, P. Las variables observables son efectos secundarios en la conducción de automóbiles: Y_1 =tiempos de reacción (segundos) a la puesta en rojo de un semáforo, Y_2 =distancia mínima (cm.) entre dos puntos que el conductor necesitaba

para poder pasar por el medio. Los datos sobre 8 individuos (media de varias pruebas) eran:

	Placebo		Clobazan		Diazepan	
Ind.	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
1	.548	177.8	.519	203.0	.637	194.8
2	.619	184.4	.776	164.8	.818	175.2
3	.641	247.2	.678	215.8	.701	205.8
4	.628	163.4	.595	153.6	.687	152.2
5	.846	173.6	.858	171.6	.855	189.2
6	.517	167.2	.493	166.0	.618	181.0
7	.876	174.0	.741	170.2	.849	189.0
8	.602	158.6	.719	157.2	.731	184.6

Los datos se ajustan a un diseño de dos factores sin interacción:

$$\mathbf{y}_{ij}= \boldsymbol{\mu} + \boldsymbol{\alpha}_i+\boldsymbol{\beta}_j+\mathbf{e}_{ij}.$$

Interesa estudiar si hay diferencias significativas entre los fármacos, y si las hay, representarlos y compararlos. Es decir, queremos hacer un test sobre la hipótesis $H_0 : \boldsymbol{\alpha}_1= \boldsymbol{\alpha}_2= \boldsymbol{\alpha}_3$ y representar las funciones estimables

$$\boldsymbol{\psi}_1 = \boldsymbol{\mu} + \boldsymbol{\alpha}_1, \quad \boldsymbol{\psi}_2 = \boldsymbol{\mu} + \boldsymbol{\alpha}_2, \quad \boldsymbol{\psi}_3 = \boldsymbol{\mu} + \boldsymbol{\alpha}_3.$$

La tabla MANOVA es:

	g. l.	matriz dispersión	lambda	F	g.l.
Fármacos	2	$\begin{pmatrix} .0275 & 1.97 \\ & 309 \end{pmatrix}$	.482	2.86	4,26
Individuos	7	$\begin{pmatrix} .258 & -1.23 \\ & 8474 \end{pmatrix}$	.025	9.84	14,26
Residuo	14	$\begin{pmatrix} .037 & -1.96 \\ & 2221 \end{pmatrix}$			

Las diferencias entre fármacos y entre individuos son significativas

Las estimaciones LS son:

$$\widehat{\boldsymbol{\psi}}_1 = (.659, 180.8)', \quad \widehat{\boldsymbol{\psi}}_2 = (.672, 175.3)', \quad \widehat{\boldsymbol{\psi}}_3 = (.737, 184.0)',$$

con dispersión (15.1): $\delta_1 = \delta_2 = \delta_3 = \sqrt{1/8} = 0.354$. Los dos valores propios de $\mathbf{U}'\mathbf{U}$ respecto de $\widehat{\Sigma}$ son $1.684, 0.108$ y explican el 100% de la variabilidad

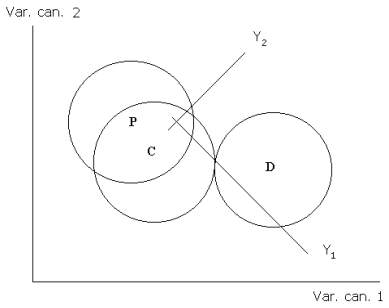

Figura 15.1: Representación canonica de tres fármacos en un diseño de dos factores.

geométrica en dimensión 2. Las coordenadas y los radios de la representación canónica (izquierda) y las correlaciones entre variables observables Y_1, Y_2, Y_3 y canónicas W_1, W_2 (derecha) son:

Fármaco	Y_1	Y_2	radio
Placebo	19.73	8.91	0.86
Clobazan	19.75	8.44	0.86
Diazepan	21.32	8.68	0.86

	W_1	W_2
Y_1	.869	-.494
Y_2	.296	.955

La representación canónica indica que no hay diferencias entre P y C. En cambio D se diferencia significativamente de P. Puesto que las variables miden efectos secundarios, resulta que C no los tiene, pero D sí (Fig. 15.1).

Ejemplo 2. Continuando con el ejemplo 14.7.1, queremos hacer la representación canónica de los tres niveles de la temperatura. Los valores propios de $\mathbf{U}'\mathbf{U}$ respecto de $\widehat{\Sigma}$ son 2.529, 1.375, que explican el 100% de la variabilidad geométrica (Fig. 15.2). Las coordenadas y los radios de la representación canónica (izquierda) y las correlaciones entre variables observables Y_1, Y_2, Y_3 y canónicas W_1, W_2 (derecha) son:

temp	W_1	W_2	radio
4	-.539	-.871	1.29
20	1.29	.091	1.29
34	-.753	.779	1.29

	W_1	W_2
Y_1	.395	.278
Y_2	.961	-.276
Y_3	.405	.653

Ejemplo 3. Continuando con el ejemplo 14.7.2, podemos hacer la representación canónica de las ocho especies, eliminando el efecto del sexo y

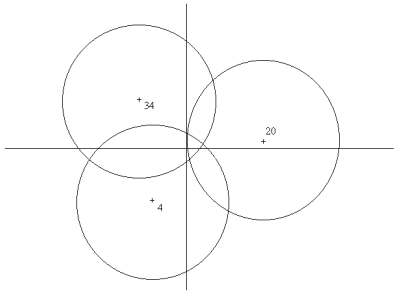

Figura 15.2: Representación canónica de los efectos principales de las temperaturas.

de la interacción. Los dos primeros valores propios de $\mathbf{U}'\mathbf{U}$ respecto de $\widehat{\Sigma}$ son 201.67, 28.054, que explican el 98.2% de la variabilidad geométrica (Fig. 13.3). Las coordenadas y los radios de la representación canónica (izquierda) y las correlaciones entre variables observables y canónicas (derecha) son:

Especie	W_1	W_2	radio
1	-4.567	-1.164	.342
2	-3.760	-.5129	.342
3	-1.944	-1.031	.418
4	-2.613	1.536	.342
5	-2.299	1.731	.342
6	-1.705	.6381	.342
7	6.828	-3.671	.503
8	10.06	2.475	.342

	W_1	W_2
Y_1	.600	.115
Y_2	.661	.450
Y_3	.453	.698
Y_4	.804	.522
Y_5	.748	.522

Esta representación permite visualizar las diferencias entre las especies, sin la influencia del dimorfismo sexual y de la interacción especie×sexo.

15.6 Complementos

El teorema de Gauss-Markov se puede generalizar de diversas maneras al caso multivariante. Ver Mardia *et al.* (1979), Rencher (1998).

La representación de funciones paramétricas estimables multivariantes fue propuesta por Cuadras (1974). Ver Cuadras *et al.* (1996) y otras generalizaciones en Lejeune y Calinski (2000), Arenas y Cuadras (2003).

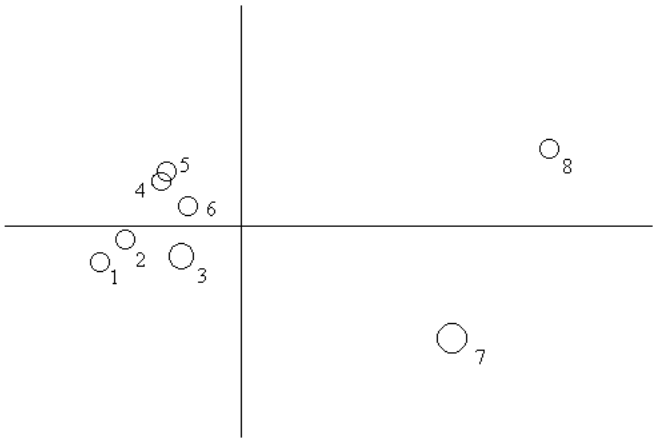

Figura 15.3: Representación canonica de 8 especies de coleópteros, eliminando el efecto del dimorfismo sexual y de la interacción.

Bibliografia

[1] Anderson, T.W. (1958) *An introduction to multivariate analysis.* J. Wiley, N. York.

[2] Anderson, T.W. and H. Rubin (1956) Statistical inference in factor analysis. *Proc. of the Third Berkeley Symposium on Math. Stat. and Prob.*, vol. **5**, 111-150.

[3] Arenas, C. and Cuadras, C. M. (2004) Comparing two methods for joint representation of multivariate data. *Comm. Stat. Comp. Simul.*, **33**, 415-430.

[4] Batista, J.M. and G. Coenders (2000) *Modelos de Ecuaciones Estructurales.* La Muralla, Madrid.

[5] Benzecri, J.P. (1976) *L'Analyse des Données. I. La Taxinomie. II. L'Analyse des Correspondances.* Dunod, Paris.

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

[7] Cooley, W.W. and P.R. Lohnes (1971) *Multivariate data analysis.* J. Wiley, N. York.

[8] Cox, T.F. and M.A.A. Cox (1964) *Multidimensional Scaling.* Chapman and Hall, London.

[9] Critchley, F. and W. Heiser (1988) Hierarchical trees can be scaled perfectly in one dimension. *J. of Classification,* **5**, 5-20.

[10] Cuadras, C.M. (1974) Análisis discriminante de funciones paramétricas estimables. *Trab. Esta. Inv. Oper.,* **25**, 3-31.

[11] Cuadras, C.M. (1981) *Métodos de Análisis Multivariante.* Eunibar, Barcelona. 3a Ed. EUB, Barcelona, 1996.

[12] Cuadras, C.M. (1988) Distancias estadísticas (con discusión) . *Estadística Española,* **30**, 295-378.

[13] Cuadras, C.M. (1989) Distance analysis in discrimination and classification using both continuous and categorical variables. In: Y. Dodge (Ed.), *Statistical Data Analysis and Inference*, pp. 459–473. Elsevier Science Publishers B. V. (North–Holland), Amsterdam.

[14] Cuadras, C.M. (1991) Ejemplos y aplicaciones insólitas en regresión y correlación. *Qüestió*, **15**, 367-382.

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

[16] Cuadras, C.M (1992b) Some examples of distance based discrimination. *Biometrical Letters*, **29**, 3-20.

[17] Cuadras, C.M. (1993) Interpreting an inequality in multiple regression. *The American Statistician,* **47**, 256-258.

[18] Cuadras, C.M. (1998) Multidimensional dependencies in ordination and classification. In: K. Fernández and E. Morinneau (Eds.), *Analyses Multidimensionnelles des Données,* pp.15-26, CISIA-Ceresta, Saint Mandé (France).

[19] Cuadras, C.M. (2000) Problemas de probabilidades y estadística. Vol. 2. *EUB*, Barcelona.

[20] Cuadras, C.M. (2002a) On the covariance between functions. *J. of Multivariate Analysis,* **81**, 19-27.

[21] Cuadras, C.M. (2002b) Correspondence analysis and diagonal expansions in terms of distribution functions. *J. of Statistical Planning and Inference,* **103**, 137-150.

[22] Cuadras, C. M. (2006) The importance of being the upper bound in the bivariate family. *SORT*, **30**, 55-84.

[23] Cuadras, C.M. and C. Arenas (1990) A distance based regression model for prediction with mixed data. *Comm. Stat.-Theor. Meth.,* **19**, 2261-2279.

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

[25] Cuadras, C.M. and J. Augé (1981) A continuous general multivariate distribution and its properties. *Commun. Stat.-Theor. Meth,* **A10,** 339-353.

[26] Cuadras. C. M., Cuadras, D. (2006) A parametric approach to correspondence analysis. *Linear Algebra and its Applications*, **417**, 64-74.

[27] Cuadras, C.M., Arenas, C. and J. Fortiana (1996) Some computational aspects of a distance-based model for prediction. *Comm. Stat.-Simul. Comp.,* **25**, 593-609.

[28] Cuadras, C.M. and J. Fortiana (1993a) Continuous metric scaling and prediction. In: C.M. Cuadras and C.R. Rao (Eds.), *Multivariate Analysis, Future Directions 2*, pp. 47–66. Elsevier Science Publishers B. V. (North–Holland), Amsterdam.

[29] Cuadras, C. M. and J. Fortiana (1993b) Aplicación de las distancias en estadística. *Questió,* **17**, 39-74.

[30] Cuadras, C. M. and J. Fortiana (1994) Ascertaining the underlying distribution of a data set. In: R. Gutierrez and M.J. Valderrama (Eds.), *Selected Topics on Stochastic Modelling*, pp. 223-230. World-Scientific, Singapore.

[31] Cuadras, C. M. and J. Fortiana (1995) A continuous metric scaling solution for a random variable. *J. of Multivariate Analysis,* **52**, 1–14.

[32] Cuadras, C. M. and J. Fortiana (1996) Weighted continuous metric scaling. In: Gupta, A. K. and V. L. Girko (Eds.), *Multidimensional Statistical Analysis and Theory of Random Matrices*, pp. 27–40. VSP, Zeist, The Netherlands.

[33] Cuadras, C.M. and J. Fortiana (1998) 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.

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

[35] Cuadras, C. M., Fortiana, J. and M.J. Greenacre (2000) Continuous extensions of matrix formulations in correspondence analysis, with applications to the FGM family of distributions. In: R.D.H. Heijmans, D.S.G. Pollock and A. Satorra, (Eds.), *Innovations in Multivariate Statistical Analysis*, pp. 101-116. Kluwer Ac. Publ., Dordrecht.

[36] Cuadras, C. M., Cuadras, D., Greenacre, M. A. (2006) Comparison of different methods for representing categorical data. *Communications in Statistics-Simul. and Comp.*, **35** (2), 447-459.

[37] Cuadras, C. M., Fortiana, J. and F. Oliva (1996) Representation of statistical structures, classification and prediction using multidimensional scaling. In: W. Gaul, D. Pfeifer (Eds.), *From Data to Knowledge,* pp. 20-31. Springer, Berlin.

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

[39] Cuadras, C.M. and Y. Lahlou (2000) Some orthogonal expansions for the logistic distribution. *Comm. Stat.-Theor. Meth.,* **29**, 2643-2663.

[40] Cuadras, C.M. and J. M. Oller (1987) Eigenanalysis and metric multidimensional scaling on hierarchical structures. *Questio,* **11**, 37-57.

[41] Cuadras, C.M. and M. Sánchez-Turet (1975) Aplicaciones del análisis multivariante canónico en la investigación psicológica. *Rev. Psicol. Gen. Aplic.*, **30**, 371-382.

[42] Chatterjee, S. and B. Price (1991) *Regression analysis by example*. Wiley, N. York.

[43] Everitt, B.S. (1993). *Cluster analysis*. Edward Arnold, London.

[44] Flury, B. (1997) *A first course in multivariate statistics*. Springer, N. York.

[45] Fortiana, J. and C. M. Cuadras (1997) A family of matrices, the discretized Brownian Bridge and distance-based regression. *Linear Algebra and its Applications*, **264**, 173-188.

[46] Gittings, R. (1985) *Canonical Analysis. A Review with Applications in Ecology*. Springer-Verlag, Berlin.

[47] Gordon, A.D. (1999) *Classification*. Chapman and Hall, London.

[48] Gower, J.C. (1966) Some distance properties of latent roots and vector methods in multivariate analysis. *Biometrika,* **53,** 315-328**.**

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

[50] Hastie, T. and R.J. Tibshirani (1990) *Generalized Additive Models*. Chapman and Hall, London.

[51] Harman, H. H. (1976) *Modern Factor Analysis*. The Univ. Chicago Press, Chicago, 3a edic.

[52] Hill, M.O. (1973) Reciprocal averaging: an eigenvector method of ordination. *J. of Ecology*, **61**, 237-249.

[53] Holman, E.W. (1972) The relation between Hierarchical and Euclidean models for psychological distances. *Psychometrika,* **37,** 417-423.

[54] Hutchinson, T.P. and C.D. Lai (1991) *The Engineering Statistician's Guide to Continuous Bivariate Distributions*. Rumsby Scientific Pub., Adelaide.

[55] Joe, H. (1997) *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

[56] Joreskog, K. (1967) Some contributions to maximum likelihood factor analysis. *Psychometrika,* **32**, 443-482.

[57] Joreskog, K. (1969) A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika,* **34**, 183-202.

[58] Joreskog, K. (1970) A general method for analysis of covarianvce structures. *Biometrika,* **57**, 239-251.

[59] Joreskog, K, Sorbom, D. (1999) *LISREL 8: A Guide to the Program and Applications.* Scientific Sotware International, Inc., Chicago.

[60] Krzanowski, W.J. and D. Radley (1989) Nonparametric confidence and tolerance regions in canonical variate analysis. *Biometrics*, **45**, 1163-1173.

[61] Lancaster, H.O. (1969) *The Chi-Squared Distribution.* J. Wiley, N. York.

[62] Lebart, L., Morineau, A. and Tabard, N. (1977) *Techniques de la description statistique.* Dunod, Paris.

[63] Lawley, D.N. and A.E. Maxwell. (1971) *Factor analysis as a statistical method.* Butterworth, London.

[64] Leujene, M. and Calinski, T. (2000) Canonical analysis applied to multivariate analysis of variance. *J. of Multivariate Analysis,* **72**, 100-119.

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

[66] Muirhead, R.J. (1982) *Aspects of multivariate statistical theory.* Wiley, N. York.

[67] Peña, D. (1989) *Estadística Modelos y Métodos 2. Modelos lineales y series temporales.* Alianza Universidad Textos, 2a Ed., Madrid.

[68] McLachlan, G.J. (1992) *Discriminant analysis and pattern recognition.* Wiley, N. York.

[69] Oller, J.M. (1987) Information metric for extreme values and logistic distributions. *Sankhya,* **49** A, 17-23.

[70] Oller, J.M. and C.M. Cuadras (1985) Rao's distance for negative multinomial distributions. *Sankhya,* **47** A, 75-83.

[71] Rao, C.R. (1952) *Advanced statistical methods in biometric research.* Wiley, N. York.

[72] Rao, C.R. (1973) *Linear statistical inference and their applications.* Wiley, N. York.

[73] Rao, C. R. (1995) A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió,* 19, 23-63.

[74] Rencher, A.C. (1998) *Multivariate statistical inference and applications.* Wiley, N. York,.

[75] Rummel, R. J. (1963) The dimensions of conflict behavior within and between nations. *General Systems Yearbook*, **8**, 1-50.

[76] Sánchez.-Turet, M. and Cuadras, C. M. (1972) Adaptación española del cuestionario E.P.I. de Eysenck. *Anuario de Psicología,* **6**, 31-59.

[77] Satorra, A. (1989) Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika,* **54**, 131-151.

[78] Seal, H.L. (1964) *Multivariate Statistical Analysis for Biologists.* Methuen and Co. Ltd., London.

[79] Seber, G.A.F. (1977) *Linear Regression Analysis.* J. Wiley, N. York.

[80] Spearman, Ch. (1904) General intelligence objetively determined and measured. *American Journal of Psychology*, **15**, 201-293.

[81] Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B,* **63**, 411-423.

[82] Torrens-Ibern, J. (1972) *Modéles et méthodes de l'analyse factorielle.* Dunod, Paris.

[83] van der Heijden, PG.M. and J. de Leuw (1985) Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, **50**, 429-447.