Análisis Multivariante

F. Tusell[1]

26 de octubre de 2016

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Normal multivariante y asociadas

1.1. Introducción.

Consideraremos en lo que sigue variables aleatorias n-variantes, es decir, aplicaciones $\boldsymbol{X} : \Omega \longrightarrow \quad R^n$. A cada $\omega \in \Omega$ corresponderá entonces un $\boldsymbol{X} = \boldsymbol{X}(\omega) \in R^n$. Designaremos por $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in})'$ a la observación i-ésima de la variable aleatoria n-variante $\boldsymbol{X}$, y por $F_{\boldsymbol{X}}(\boldsymbol{x})$ y $f_{\boldsymbol{X}}(\boldsymbol{x})$ a las funciones de distribución y densidad respectivamente de $\boldsymbol{X}$. Emplearemos el convenio de utilizar mayúsculas para las variables aleatorias y minúsculas para sus valores concretos en un muestreo determinado. Llamaremos X_j a la variable aleatoria j-ésima.

¿Por qué no emplear las técnicas habituales (univariantes) sobre cada X_j?. Podríamos en efecto estudiar cada X_j por separado. Si lo hiciéramos, perderíamos sin embargo la posibilidad de extraer partido de la (posible) correlación entre diferentes variables X_j y X_k en $\boldsymbol{X}$. Los métodos de Análisis Multivariante comparten la idea de explotar esta información.

Llamaremos $\boldsymbol{\mu_X}$ al vector de medias de la variable aleatoria $\boldsymbol{X}$, y $\Sigma_{\boldsymbol{X}}$ a su matriz de covarianzas.

$$\boldsymbol{\mu_X} = E\boldsymbol{X} \tag{1.1}$$

$$\Sigma_{\boldsymbol{X}} = E[(\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{X} - \boldsymbol{\mu_X})'] \tag{1.2}$$

Al igual que la distribución normal desempeña un papel destacado en la Estadística univariante, una generalización de ella, la distribución nor-

mal multivariante, constituye un modelo teórico de gran trascendencia en el Análisis Multivariante.

1.2. Distribución normal multivariante.

Se dice que $X \sim N(0,1)$ si:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty$$

y por ende:

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}x^2} dx \qquad -\infty < x < \infty \tag{1.3}$$

$$\psi_X(u) = Ee^{iuX} \tag{1.4}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-iu)^2} e^{-\frac{1}{2}u^2} dx \tag{1.5}$$

$$= e^{-\frac{1}{2}u^2} \tag{1.6}$$

Por transformación lineal de una variable aleatoria $N(0,1) : Y = \sigma X + \mu$ se obtiene una variable aleatoria normal general $N(\mu, \sigma^2)$ cuyas funciones de densidad, distribución y característica son:

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \qquad -\infty < y < \infty \tag{1.7}$$

$$F_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \qquad -\infty < y < \infty \tag{1.8}$$

$$\psi_Y(u) = e^{iu\mu - \frac{1}{2}\sigma^2 u^2} \tag{1.9}$$

Si tenemos p variables aleatorias X_j con distribución $N(0,1)$, independientes unas de otras, la función de densidad conjunta de la variable aleatoria p-variante $\boldsymbol{X} = (X_1, \ldots, X_p)'$ viene dada por el producto de las marginales

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}(x_1^2 + \ldots + x_p^2)} \tag{1.10}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}\boldsymbol{x}' I \boldsymbol{x}}, \tag{1.11}$$

y la función característica por:

$$\psi_{\boldsymbol{X}}(\boldsymbol{u}) = e^{-\frac{1}{2}\boldsymbol{u}'\boldsymbol{u}}. \tag{1.12}$$

Decimos que la variable aleatoria p-variante $\boldsymbol{X}$ cuya función de densidad es (1.10) sigue una distribución $N_p(\vec{0}, I)$, designando el primer argumento el vector de medias y el segundo la matriz de covarianzas. Esta última es

diagonal, en virtud de la independencia entre las distintas componentes de $\boldsymbol{X}$.

Si efectuamos una transformación lineal $\boldsymbol{X} \longrightarrow \boldsymbol{Y}$ como

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p + \mu_1 && (1.13)\\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p + \mu_2 && (1.14)\\ &\vdots \\ Y_p &= a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p + \mu_p && (1.15) \end{aligned}$$

o, en notación matricial, $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{\mu}$, y A es de rango completo, tenemos que $\boldsymbol{X} = A^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$ y la función de densidad de $\boldsymbol{Y}$ se obtiene fácilmente de la de $\boldsymbol{X}$:

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{\boldsymbol{X}}(A^{-1}(\boldsymbol{y} - \boldsymbol{\mu}))\left|\frac{\partial \boldsymbol{X}}{\partial \boldsymbol{Y}}\right| && (1.16)\\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'(A^{-1})'(A^{-1})(\boldsymbol{y}-\boldsymbol{\mu})}|A^{-1}| && (1.17)\\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \frac{1}{|A|} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'(AA')^{-1}(\boldsymbol{y}-\boldsymbol{\mu})} && (1.18) \end{aligned}$$

Como

$$\begin{aligned} \Sigma_{\boldsymbol{Y}} &= E(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})' && (1.19)\\ &= EA\boldsymbol{X}\boldsymbol{X}'A' && (1.20)\\ &= AA', && (1.21) \end{aligned}$$

tenemos que la función de densidad (1.18) puede escribirse así:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \left(\frac{1}{\sqrt{2\pi}}\right)^p \frac{1}{|\Sigma_{\boldsymbol{Y}}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}, \qquad (1.22)$$

ya que $|A| = \sqrt{|A||A|} = \sqrt{|A||A'|} = \sqrt{|\Sigma_{\boldsymbol{Y}}|}$. Por otra parte, la función característica de $\boldsymbol{Y}$ es:

$$\begin{aligned} \psi_{\boldsymbol{Y}}(\boldsymbol{u}) &= Ee^{i\boldsymbol{u}'\boldsymbol{Y}} && (1.23)\\ &= Ee^{i\boldsymbol{u}'(A\boldsymbol{X}+\boldsymbol{\mu})} && (1.24)\\ &= \psi_{\boldsymbol{X}}(A'\boldsymbol{u})e^{i\boldsymbol{u}'\boldsymbol{\mu}} && (1.25)\\ &= e^{i\boldsymbol{u}'\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{u}'AA'\boldsymbol{u}} && (1.26)\\ &= e^{i\boldsymbol{u}'\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{u}'\Sigma_{\boldsymbol{Y}}\boldsymbol{u}} && (1.27) \end{aligned}$$

La expresión (1.22) requiere para estar definida que $\Sigma_{\boldsymbol{Y}}$ sea de rango total –sólo así puede encontrarse la inversa–. La expresión (1.27) por el contrario es una función característica incluso aunque $\Sigma_{\boldsymbol{Y}}$ sea de rango deficiente. Se dice que (1.22) y (1.27) son funciones de densidad y característica de un

vector aleatorio con distribución $N_p(\boldsymbol{\mu}, \Sigma_{\boldsymbol{Y}})$. Si $\Sigma_{\boldsymbol{Y}}$ es de rango deficiente, se dice que estamos ante una distribución *normal singular*, que carece de densidad (1.22).

Observación 1.1 La función de densidad normal multivariante es unimodal, alcanza su máximo para $\boldsymbol{y}$ coincidente con el vector de medias $\boldsymbol{\mu}$, y tiene contornos de igual densidad elípticos (o hiper-elípticos).

Los siguientes hechos son de muy sencilla demostración:

1. Las distribuciones de cualesquiera combinaciones lineales de componentes de $\boldsymbol{Y}$ son normales.

2. Si $\boldsymbol{Y}$ es normal multivariante, cualesquiera marginales son normales uni- o multivariantes.

3. Si $\boldsymbol{X}$ e $\boldsymbol{Y}$ son vectores independientes conjuntamente definidos con distribuciones respectivas $N_p(\boldsymbol{\mu_X}, \Sigma_{\boldsymbol{X}})$ y $N_p(\boldsymbol{\mu_Y}, \Sigma_{\boldsymbol{Y}})$, y A, B son matrices cualesquiera de orden $d \times p$, $(d \leq p)$, y rango d, se verifica:

$$A\boldsymbol{X} + B\boldsymbol{Y} \sim N_d(A\boldsymbol{\mu_X} + B\boldsymbol{\mu_Y}, A\Sigma_{\boldsymbol{X}}A' + B\Sigma_{\boldsymbol{Y}}B')$$

   Como caso particular, $C\boldsymbol{X} \sim N_d(C\boldsymbol{\mu_X}, C\Sigma_{\boldsymbol{X}}C')$.

4. La incorrelación entre cualesquiera componentes X_i, X_j (o grupos de componentes) de $\boldsymbol{X}$, implica su independencia. En el caso de variables aleatorias con distribución normal multivariante, incorrelación e independencia son nociones coextensivas.

5. Transformaciones lineales ortogonales de vectores $N_d(\vec{0}, \sigma^2 I)$ tienen distribución $N_d(\vec{0}, \sigma^2 I)$.

Observación 1.2 Una normal multivariante tiene contornos de igual densidad, cuando esta densidad existe, cuya expresión viene dada por:

$$-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})'\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = k.$$

Como la matriz de covarianzas (en el caso de rango completo, para el que existe la densidad) es definida positiva, la expresión anterior proporciona la superficie de un hiper-elipsoide: una elipse ordinaria en R^2, un elipsoide (similar a un balón de rugby) en R^3, y figuras que ya no podemos visualizar en más de tres dimensiones.

Observación 1.3 Hay versiones multivariantes del Teorema Central del Límite, que sugieren que variables multivariantes que son:

- Suma de muchas otras,
- Aproximadamente independientes, y
- Sin influencia abrumadora de ninguna sobre el conjunto,

siguen distribución aproximadamente normal multivariante. Es un hecho, sin embargo, que el supuesto de normalidad multivariante es sumamente restrictivo, y de rara plausibilidad en la práctica. En particular, el supuesto de normalidad multivariante es *mucho más fuerte* que el de normalidad de las marginales, como el siguiente ejemplo ilustra.

Ejemplo 1.1 Supongamos un vector bivariante (X_1, X_2), en que X_1 y X_2 son respectivamente temperaturas máximas y mínimas de una ubicación. Podemos perfectamente imaginar un caso con normalidad marginal (las mínimas y máximas se distribuyen cada una de modo normal). Sin embargo, el supuesto de normalidad bivariante sería claramente inadecuado: por definición, $X_1 \geq X_2$, y por tanto el vector (X_1, X_2) se distribuye sólo en el semiplano por debajo de la recta $X_1 = X_2$. Una normal bivariante debe estar definida en todo el plano real.

El siguiente teorema será de utilidad:

Teorema 1.1 *Sea $\boldsymbol{X}$ un vector aleatorio con distribución normal $(p+q)$-variante, particionado del modo que se indica:*

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

Entonces la distribución de $\boldsymbol{X}_1$ condicionada por $\boldsymbol{X}_2 = \boldsymbol{x}_2$ es:

$$N_p(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

DEMOSTRACION:

Una demostración conceptualmente simple se limitaría a efectuar el cociente de la densidad conjunta entre la densidad marginal $f(\boldsymbol{X}_1)$, simplificando el cociente hasta encontrar una densidad normal con el vector de medias y matriz de covarianzas que indica el enunciado. Una aproximación más simple es la que sigue (véase **?**, p. 99). Consideremos la variable aleatoria

$$\boldsymbol{Y} = \boldsymbol{X}_1 + M\boldsymbol{X}_2,$$

siendo M una matriz de dimensiones $p \times q$. La matriz de covarianzas entre las $\boldsymbol{Y}$ y las $\boldsymbol{X}_2$ será:

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{X}_2) &= E\left\{[(\boldsymbol{X}_1 - \boldsymbol{\mu}_1) + M(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)](\boldsymbol{X}_2 - \boldsymbol{\mu}_2)'\right\} && (1.28) \\ &= E\left\{(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)' + M(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)'\right\} && (1.29) \\ &= \Sigma_{12} + M\Sigma_{22} && (1.30) \end{aligned}$$

Si hacemos $M = -\Sigma_{12}\Sigma_{22}^{-1}$, la expresión anterior será una matriz de ceros; por tanto, $\boldsymbol{Y} = \boldsymbol{X}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X}_2$ es un vector aleatorio normal multivariante independiente de $\boldsymbol{X}_2$.

Siendo independiente, su distribución incondicionada y condicionada por $\boldsymbol{X}_2 = \boldsymbol{x}_2$ es la misma. Tomando valor medio y matrices de covarianzas en ambos casos, obtenemos los siguientes momentos:

a) Incondicionados:

$$\begin{aligned} E[\boldsymbol{Y}] &= E[\boldsymbol{X}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X}_2] = \boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2 \quad (1.31) \\ \Sigma_{\boldsymbol{Y}} &= E[(\boldsymbol{X}_1 - \boldsymbol{\mu}_1) - \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)][(\boldsymbol{X}_1 - \boldsymbol{\mu}_1) - \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)]' \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{12}{}' = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}{}' \quad (1.32) \end{aligned}$$

b) Condicionados:

$$\begin{aligned} E[\boldsymbol{Y}|\boldsymbol{X}_2 = \boldsymbol{x}_2] &= E[\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2] - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2 \quad (1.33) \\ \Sigma_{\boldsymbol{Y}|\boldsymbol{X}_2=\boldsymbol{x}_2} &= \Sigma_{(\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2)} \quad (1.34) \end{aligned}$$

e igualando (1.31) a (1.33) y (1.32) a (1.34) llegamos a:

$$\begin{aligned} E[\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2] &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \quad (1.35) \\ \Sigma_{\boldsymbol{Y}|\boldsymbol{X}_2=\boldsymbol{x}_2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad (1.36) \end{aligned}$$

Las expresiones (1.35) y (1.36) junto con la normalidad de $\boldsymbol{X}_1$ demuestran el teorema.

1.3. Regresión lineal.

Supongamos, con la notación de la Sección anterior, que $p = 1$ (con lo que $\boldsymbol{X}_1$ es un escalar), y que nos planteamos el siguiente problema: encontrar $g(\boldsymbol{X}_2)$ aproximando de manera "óptima" a X_1. "Óptima" se entiende en el sentido de minimizar $E[X_1 - g(\boldsymbol{X}_2)]^2$. Demostraremos que la función $g(\boldsymbol{X}_2)$ buscada es precisamente $E[X_1|\boldsymbol{X}_2]$. Para ello precisamos algunos resultados instrumentales.

Lema 1.1 *Si denotamos mediante un superíndice la v.a. con respecto a la cual se toma valor medio (es decir, $E^{(X_1)}[Z] = \int_{-\infty}^{\infty} Z f_{X_1}(x_1)dx_1$), se tiene:*

$$E[\boldsymbol{X}_1] = E^{(\boldsymbol{X}_1)}[\boldsymbol{X}_1] = E^{(\boldsymbol{X}_2)}[E^{(\boldsymbol{X}_1)}(\boldsymbol{X}_1|\boldsymbol{X}_2)]$$

Demostracion:

$$
\begin{aligned}
E^{(\boldsymbol{X}_2)}[E^{(\boldsymbol{X}_1)}(\boldsymbol{X}_1|\boldsymbol{X}_2)] &= \int f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)[E^{(\boldsymbol{X}_1)}(\boldsymbol{X}_1|\boldsymbol{X}_2)]d\boldsymbol{x}_2 && (1.37)\\
&= \int f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)\left[\int \boldsymbol{x}_1 f_{\boldsymbol{X}_1|\boldsymbol{X}_2}(\boldsymbol{x}_1|\boldsymbol{x}_2)d\boldsymbol{x}_1\right]d\boldsymbol{x}_2 && (1.38)\\
&= \int d\boldsymbol{x}_1 \int d\boldsymbol{x}_2 \left[\boldsymbol{x}_1 f_{\boldsymbol{X}_1|\boldsymbol{X}_2}(\boldsymbol{x}_1|\boldsymbol{x}_2) f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)\right] && (1.39)\\
&= \int d\boldsymbol{x}_1 \int d\boldsymbol{x}_2 \left[\boldsymbol{x}_1 f_{\boldsymbol{X}_1,\boldsymbol{X}_2}(\boldsymbol{x}_1,\boldsymbol{x}_2)\right] && (1.40)\\
&= \int \boldsymbol{x}_1 d\boldsymbol{x}_1 \int f_{\boldsymbol{X}_1,\boldsymbol{X}_2}(\boldsymbol{x}_1,\boldsymbol{x}_2)d\boldsymbol{x}_2 && (1.41)\\
&= \int \boldsymbol{x}_1 f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)d\boldsymbol{x}_1 && (1.42)\\
&= E^{(\boldsymbol{X}_1)}[\boldsymbol{X}_1] && (1.43)
\end{aligned}
$$

Lema 1.2 *Sea,*

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

Entonces, $Z = X_1 - E[X_1|\boldsymbol{X}_2]$ es una v.a. incorrelada con cualquier función $\ell(\boldsymbol{X}_2)$.

DEMOSTRACION:

Como, de acuerdo con el lema anterior, $E[Z] = 0$, tenemos que:

$$
\begin{aligned}
\text{cov}[Z, \ell(\boldsymbol{X}_2)] &= E\left[Z(\ell(\boldsymbol{X}_2) - E[\ell(\boldsymbol{X}_2)])\right] && (1.44)\\
&= E[Z\ell(\boldsymbol{X}_2)] && (1.45)\\
&= E[X_1\ell(\boldsymbol{X}_2) - E[X_1|\boldsymbol{X}_2]\ell(\boldsymbol{X}_2)] && (1.46)\\
&= 0 && (1.47)
\end{aligned}
$$

haciendo uso del lema anterior para evaluar la expresión (1.46). Tenemos así el siguiente,

Teorema 1.2 *La mejor aproximación en términos de error cuadrático medio de X_1 en función de $\boldsymbol{X}_2$ es la proporcionada por $g(\boldsymbol{X}_2) = E[X_1|\boldsymbol{X}_2]$.*

DEMOSTRACION: Consideremos cualquier otra función $h(\boldsymbol{X}_2)$. Entonces:

$$
\begin{aligned}
E[X_1 - h(\boldsymbol{X}_2)]^2 &= E[X_1 - g(\boldsymbol{X}_2) + g(\boldsymbol{X}_2) - h(\boldsymbol{X}_2)]^2\\
&= E[X_1 - g(\boldsymbol{X}_2)]^2 + E[g(\boldsymbol{X}_2) - h(\boldsymbol{X}_2)]^2\\
&\quad +2\text{cov}[\underbrace{X_1 - g(\boldsymbol{X}_2)}_{Z}, \underbrace{g(\boldsymbol{X}_2) - h(\boldsymbol{X}_2)}_{\ell(\boldsymbol{X}_2)}]\\
&= E[X_1 - g(\boldsymbol{X}_2)]^2 + E[g(\boldsymbol{X}_2) - h(\boldsymbol{X}_2)]^2\\
&\geq E[X_1 - g(\boldsymbol{X}_2)]^2
\end{aligned}
$$

Es interesante observar que $E[X_1|\boldsymbol{X}_2]$ es una función lineal de $\boldsymbol{X}_2$ en el caso que consideramos de distribución normal multivariante conjunta de $X_1, \boldsymbol{X}_2$. La expresión de $E[X_1|\boldsymbol{X}_2]$ es reminiscente de la de $X\hat{\beta}$ en regresión lineal, pero aquí la linealidad no es un supuesto, sino un resultado.

Definición 1.1 *Llamamos* varianza generalizada *de una distribución multivariante al determinante de su matriz de covarianzas,* $|\Sigma|$. *Llamamos* varianza total *a traza*(Σ).

Lema 1.3 *Las varianzas generalizadas de la distribución de* $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$ *y las correspondientes a las distribuciones de* $\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2$ *y* $\boldsymbol{X}_2$ *están relacionadas por:*

$$|\Sigma| = |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}||\Sigma_{22}|$$

Demostracion: Basta tomar determinantes en la igualdad matricial,

$$\begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{12}' & I \end{pmatrix} = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

Emplearemos la notación $\Sigma_{11,2}$ para designar la matriz de covarianzas $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Algunas cosas merecen resaltarse. La matriz de covarianzas de la distribución condicionada por $\boldsymbol{X}_2 = \boldsymbol{x}_2$ no depende de $\boldsymbol{x}_2$. Por otra parte, la expresión que da el valor medio de $\boldsymbol{X}_1$ condicionado por $\boldsymbol{X}_2 = \boldsymbol{x}_2$ es formalmente similar a la que se obtendría regresando los valores centrados de $\boldsymbol{X}_1$ sobre los valores centrados de $\boldsymbol{X}_2$. Es una función lineal en $\boldsymbol{x}_2$.

Una tercera observación de interés es que las varianzas de las $\boldsymbol{X}_1$ en la distribución condicionada son no mayores que en la distribución no condicionada; esto es fácil de ver si reparamos en que los elementos diagonales de $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (que se restan de sus homólogos de Σ_{11}) resultan de evaluar una forma cuadrática de matriz Σ_{22}^{-1} definida no negativa. Esto es lógico: conocido $\boldsymbol{X}_2 = \boldsymbol{x}_2$, disminuye la incertidumbre acerca de los valores que puede tomar $\boldsymbol{X}_1$. El único caso en que las varianzas –condicionadas e incondicionadas– serían idénticas es aquél en que $\Sigma_{12} = 0$.

1.4. Correlación simple, parcial y múltiple.

Sean X_i y X_j dos variables aleatorias conjuntamente definidas. Sean σ_i^2 y σ_j^2 sus varianzas respectivas, y λ_{ij} su covarianza. Se denomina *coeficiente de correlación simple* entre ambas a:

$$\rho_{ij} \stackrel{\text{def}}{=} \frac{\lambda_{ij}}{+\sqrt{\sigma_i^2\sigma_j^2}}. \tag{1.48}$$

Se demuestra fácilmente haciendo uso de la desigualdad de Schwartz que $-1 \leq \rho_{ij} \leq +1$. Un coeficiente de correlación simple igual a 1 en valor absoluto (+1 ó -1) indica una perfecta asociación lineal entre las variables aleatorias X_i y X_j (véase **?**, Cap. 14, por ej.).

Imaginemos que X_i, X_j son variables aleatorias de entre las que componen el vector $\boldsymbol{X}_1$. Si las varianzas y covarianzas en (1.48), en lugar de proceder de Σ_{11}, proceden de los lugares homólogos en $\Sigma_{11,2}$, tenemos el llamado *coeficiente de correlación parcial* entre X_i y X_j controlado el efecto de $\boldsymbol{X}_2$:

$$\rho_{ij.\boldsymbol{X}_2} \stackrel{\text{def}}{=} \frac{\lambda_{ij,2}}{+\sqrt{\sigma^2_{i,2}\sigma^2_{j,2}}}.$$

Podemos interpretar $\rho_{ij.\boldsymbol{X}_2}$ como el coeficiente de correlación entre X_i y X_j una vez que de ambas se ha eliminado la parte que cabe expresar como combinación lineal de las variables aleatorias en $\boldsymbol{X}_2$.

Definimos *coeficiente de correlación múltiple al cuadrado* entre la variable X_j (en $\boldsymbol{X}_1$) y $\boldsymbol{X}_2$ así:

$$R^2_{j.\boldsymbol{X}_2} = \left(\frac{\sigma^2_j - \sigma^2_{j.\boldsymbol{X}_2}}{\sigma^2_j}\right),$$

o en forma reminiscente del $R^2 = 1 - \text{SSE/SST}$ habitual en regresión,

$$R^2_{j.\boldsymbol{X}_2} = 1 - \frac{\sigma^2_{j.\boldsymbol{X}_2}}{\sigma^2_j}.$$

El coeficiente de correlación múltiple al cuadrado es aquella parte de la varianza de X_j "explicada" linealmente por las variables aleatorias $\boldsymbol{X}_2$.

Ejemplo 1.2 Consideremos una matriz de covarianzas[1] entre las tres variables X_1 = "Tensión arterial", X_2 = "Renta disponible" y X_3 = "Edad".

$$\Sigma = \begin{pmatrix} 1{,}00 & 0{,}60 & 0{,}90 \\ 0{,}60 & 1{,}00 & 0{,}80 \\ 0{,}90 & 0{,}80 & 1{,}00 \end{pmatrix};$$

Una apreciación superficial podría llevar a concluir que hay una abultada correlación de 0.60 entre la variable X_2 (Renta) y la variable X_1 (Tensión arterial). Si efectuamos el análisis controlando el efecto de la variable X_3, el resultado cambia drásticamente. En efecto, tendríamos:

$$\begin{aligned} \Sigma_{11} &= \begin{pmatrix} 1{,}00 & 0{,}60 \\ 0{,}60 & 1{,}00 \end{pmatrix} \\ \Sigma_{22} &= \begin{pmatrix} 1{,}00 \end{pmatrix} \\ \Sigma_{12} &= \begin{pmatrix} 0{,}90 \\ 0{,}80 \end{pmatrix} \end{aligned}$$

[1] Valores ficticios. El ejemplo es puramente ilustrativo.

Por consiguiente, la matriz de covarianzas de las variables X_1, X_2 controlado el efecto de X_3, en aplicación del Teorema 1.1, resulta ser:

$$\Sigma_{11\cdot 2} = \begin{pmatrix} 1{,}00 & 0{,}60 \\ 0{,}60 & 1{,}00 \end{pmatrix} - \begin{pmatrix} 0{,}90 \\ 0{,}80 \end{pmatrix} (1{,}00) \begin{pmatrix} 0{,}90 & 0{,}80 \end{pmatrix} \quad (1.49)$$

$$\approx \begin{pmatrix} 0{,}19 & -0{,}12 \\ -0{,}12 & 0{,}30 \end{pmatrix} \quad (1.50)$$

El coeficiente de correlación *parcial* (eliminado el efecto de X_3 entre X_1 y X_2 sería ahora:

$$\rho_{12,3} \approx \frac{-0{,}12}{\sqrt{0{,}19 \times 0{,}30}} \approx -0{,}4588;$$

es decir, una correlación apreciable y de signo contrario al inicial.

No cuesta imaginar el origen de la aparente paradoja. Las dos variables X_1 y X_2 aparecen altamente correladas con la X_3 (Edad), y ello induce una correlación espúrea entre ellas. Al eliminar el efecto (lineal) de la variable X_3, la aparente relación directa entre X_1 y X_2 desaparece por completo (de hecho, se torna de relación inversa).

1.5. Distribución de Wishart.

Definición 1.2 *Sean $\boldsymbol{X}_i$ $(i = 1, \ldots, n)$ vectores aleatorios independientes, con distribución común $N_d(\vec{0}, \Sigma)$. Entonces, la matriz aleatoria*

$$A = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i{}'$$

con $\frac{1}{2}d(d+1)$ elementos distintos –dado que es simétrica– sigue la distribución conocida como distribución de Wishart, $W_d(n, \Sigma)$, con n grados de libertad y matriz de parámetros Σ.

La distribución de Wishart puede en cierto modo considerarse como una generalización de la χ^2; en efecto, si $X_i \sim N_1(0, \sigma^2)$ se verifica que: $A = \sum_{i=1}^{n} X_i^2 \sim \sigma^2 \chi_n^2 = W_1(n, \sigma^2)$. De la definición se deducen de modo inmediato las siguientes propiedades:

1. Si $S \sim W_d(n, \Sigma)$, $T \sim W_d(m, \Sigma)$ y ambas son independientes, $S + T \sim W_d(m + n, \Sigma)$.

2. Si $S \sim W_d(n, \Sigma)$ y C es una matriz $q \times d$ de rango q, entonces:

   $$CSC' \sim W_q(n, C\Sigma C')$$

   DEMOSTRACION: $S \sim W_d(n, \Sigma) \Leftrightarrow S = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i{}'$ con $\boldsymbol{X}_i \sim N_d(\vec{0}, \Sigma)$.

   Por consiguiente,

   $$CSC' = C \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i{}' \right) C' = \sum_{i=1}^{n} (C\boldsymbol{X}_i)(C\boldsymbol{X}_i)'$$

   Pero $C\boldsymbol{X}_i \sim N_q(\vec{0}, C\Sigma C')$, lo que muestra que $CSC' \sim W_q(n, C\Sigma C')$.

3. Como caso particular de la propiedad anterior, si $\vec{a}$ es un vector de constantes y $S \sim W_d(n, \Sigma)$ tenemos:

   $$\boldsymbol{a}'S\boldsymbol{a} \sim W_1(n, \boldsymbol{a}'\Sigma\boldsymbol{a}) \sim (\boldsymbol{a}'\Sigma\boldsymbol{a})\chi_n^2 \tag{1.51}$$

   o, lo que es igual,

   $$\frac{\boldsymbol{a}'S\boldsymbol{a}}{\boldsymbol{a}'\Sigma\boldsymbol{a}} \sim \chi_n^2 \qquad \forall \boldsymbol{a} \neq 0 \tag{1.52}$$

4. Como caso particular de (1.52), si $\boldsymbol{a}' = (0 \ldots 0 \ 1 \ 0 \ldots 0)$ (un único "uno" en posición i-ésima) se verifica que cuando $S \sim W_d(n, \Sigma)$,

   $$\boldsymbol{a}'S\boldsymbol{a} = s_{ii}^2 \sim \sigma_{ii}^2 \chi_n^2. \tag{1.53}$$

   Es decir, el cociente entre un elemento diagonal de una matriz de Wishart y la correspondiente varianza poblacional, se distribuye como una χ_n^2, con los mismos grados de libertad que la Wishart.

1.6. Formas cuadráticas generalizadas.

Sea X una matriz $N \times d$, que representaremos alternativamente de una de las siguientes formas:

$$X = \begin{pmatrix} \boldsymbol{X}_1{}' \\ \boldsymbol{X}_2{}' \\ \vdots \\ \boldsymbol{X}_N{}' \end{pmatrix} = \left(\boldsymbol{X}^{(1)} \boldsymbol{X}^{(2)} \dots \boldsymbol{X}^{(d)} \right)$$

Entonces, la "suma de cuadrados" $W = \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i{}'$ puede escribirse como: $W = X'X$. Es una matriz $d \times d$. Llamaremos *forma cuadrática generalizada* a una expresión como:

$$X'AX = \sum_i \sum_j a_{ij} \boldsymbol{X}_i \boldsymbol{X}_j{}'.$$

Es, como la "suma de cuadrados" anterior, una matriz $d \times d$.

Lema 1.4 *Si las filas de X siguen una distribución $\boldsymbol{X}_i \overset{iid}{\sim} N_d(\vec{0}, \Sigma)$, se verifica lo siguiente:*

1. $\boldsymbol{X}^{(j)} \sim N_N(\vec{0}, \sigma_{jj}^2 I_N)$.
2. $X'\boldsymbol{a} \sim N_d(\vec{0}, ||\boldsymbol{a}||^2 \Sigma)$.
3. *Si $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$, $r \leq N$, son vectores en R^N mutuamente ortogonales, $\vec{u}_i = X'\boldsymbol{a}_i \quad (i = 1, \dots, r)$ son mutuamente independientes. Si $||\boldsymbol{a}_i||^2 = 1$, $\vec{u}_i \sim N_d(\vec{0}, \Sigma)$.*

Demostracion: Solo (3) requiere demostración, siendo inmediatos los restantes apartados. Consideremos $\vec{u}_i, \vec{u}_j \quad (i \neq j)$. Claramente, $E[\vec{u}_i] = E[\vec{u}_j] = \vec{0}$, y:

$$\begin{aligned} E[\boldsymbol{u_i} \boldsymbol{u_j}{}'] &= E\left[\left(\sum_k a_{ik} \boldsymbol{X}_k \right) \left(\sum_l a_{jl} \boldsymbol{X}_l \right)' \right] \\ &= \sum_k \sum_l a_{ik} a_{jl} E[\boldsymbol{X}_k \boldsymbol{X}_l{}'] \\ &= \sum_k a_{ik} a_{jk} \Sigma \\ &= \begin{cases} 0_{d \times d} & \text{si } i \neq j \text{ (de donde se sigue la independencia)} \\ \Sigma & \text{si } i = j \text{ y } ||\vec{a}_i||^2 = 1 \end{cases} \end{aligned}$$

Lema 1.5 *Sea X una matriz aleatoria $N \times d$ cuyas filas $\boldsymbol{X}_i{}'$ son independientes con distribución común $N_d(\vec{0}, \Sigma)$. Sea U una matriz ortogonal $N \times N$, e $Y = UX$. Entonces, $Y'Y = X'X$ se distribuye como una $W_d(N, \Sigma)$.*

DEMOSTRACION:

Es inmediata: $Y'Y = X'U'UX = X'X$. Es claro además que $X'X = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i{}'$ sigue la distribución indicada.

Teorema 1.3 *Sea X una matriz aleatoria $N \times d$ cuyas filas $\boldsymbol{X}_i'$ son independientes con distribución común $N_d(\vec{0}, \Sigma)$. Los estimadores habituales del vector de medias y matriz de covarianzas:*

$$S = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})' \tag{1.54}$$

$$\overline{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i \tag{1.55}$$

verifican:

1. *S es independiente de $\overline{\boldsymbol{X}}$.*
2. *$NS \sim W_d(N-1, \Sigma)$.*

DEMOSTRACION: Consideremos una matriz U ortogonal $N \times N$ cuya última fila sea:

$$\left(\tfrac{1}{\sqrt{N}} \quad \cdots \quad \tfrac{1}{\sqrt{N}} \quad \tfrac{1}{\sqrt{N}}\right).$$

Sea $Y = UX$. Su última fila es: $\boldsymbol{Y_N} = \sum_{i=1}^{N} u_{Ni}\boldsymbol{X}_i = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{X}_i = \overline{\boldsymbol{X}}\sqrt{N}$. Por tanto, $\boldsymbol{Y_N}\boldsymbol{Y_N}{}' = N\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}{}'$. Por otra parte,

$$\begin{aligned}
NS &= \sum_{i=1}^{N}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})' \\
&= \sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i{}' - N\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}{}' - N\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}{}' + N\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}{}' \\
&= \sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i{}' - N\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}{}' \\
&= \sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i{}' - \boldsymbol{Y_N}\boldsymbol{Y_N}{}' \\
&= \sum_{i=1}^{N}\boldsymbol{Y_i}\boldsymbol{Y_i}{}' - \boldsymbol{Y_N}\boldsymbol{Y_N}{}' \\
&= \sum_{i=1}^{N-1}\boldsymbol{Y_i}\boldsymbol{Y_i}{}'
\end{aligned}$$

Como las filas $\vec{Y_i}$ son independientes unas de otras, y $\overline{\boldsymbol{X}}$ y NS dependen de filas diferentes, son claramente independientes. Es de destacar que, aunque

hemos supuesto $E[\boldsymbol{X}] = \mathbf{0}$, este supuesto es innecesario. Puede comprobarse fácilmente que si sumamos una constante cualquiera a cada columna $X^{(j)}$, S no se altera.

1.7. Distribución T^2 de Hotelling.

Sea $W \sim W_d(n, \Sigma)$ y $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, ambas independientes. Entonces:

$$n(\boldsymbol{X} - \boldsymbol{\mu})' W^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$$

sigue la distribución conocida como T^2 de Hotelling, de dimensión d y con n grados de libertad. La denotaremos por $T^2_{d,n}$. Esta distribución puede verse como una generalización de la $\mathcal{F}_{1,n}$ (y, por tanto, T como una generalización de la t de Student). En efecto, cuando $d = 1$,

$$W \sim W_1(n, \sigma^2) = \sigma^2 \chi^2_n \tag{1.56}$$
$$X \sim N(\mu, \sigma^2) \tag{1.57}$$

y:

$$n(\boldsymbol{X} - \boldsymbol{\mu})' W^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = \frac{(X-\mu)^2}{W/n} = \frac{\left(\frac{X-\mu}{\sigma}\right)^2}{W/n\sigma^2} \sim \mathcal{F}_{1,n}$$

No es preciso contar con tablas de la distribución de Hotelling, pues una relación muy simple la liga con la distribución $\mathcal{F}$ de Snedecor. Para su establecimiento necesitaremos los lemas a continuación. La presentación sigue de modo bastante ajustado a **?**, p. 29 y siguientes, donde se puede acudir para más detalles.

Lema 1.6 *Si $\boldsymbol{Y} \sim N_d(\boldsymbol{0}, \Sigma)$ y Σ es de rango completo, entonces: $\boldsymbol{Y}' \Sigma^{-1} \boldsymbol{Y} \sim \chi^2_d$.*

Demostracion: Siendo Σ definida positiva, Σ^{-1} existe y es también definida positiva. Entonces puede encontrarse $\Sigma^{-\frac{1}{2}}$ tal que: $\Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} = \Sigma^{-1}$. Por otra parte, $\boldsymbol{X} = \Sigma^{-\frac{1}{2}} \boldsymbol{Y}$ se distribuye como $N_d(\boldsymbol{0}, I_d)$. Entonces,

$$\boldsymbol{Y}' \Sigma^{-1} \boldsymbol{Y} = \boldsymbol{Y}' \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \boldsymbol{Y} = \boldsymbol{X}' \boldsymbol{X} \sim \chi^2_d$$

Lema 1.7 *Sea $\boldsymbol{X}' = (X_1 \vdots \boldsymbol{X}_2{}')$ un vector $N_d(\boldsymbol{\mu}, \Sigma)$, con $\boldsymbol{\mu} = (\mu_1 \vdots \boldsymbol{\mu_2}{}')$ y $\Sigma = \begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Sea σ^{ij} el elemento genérico en el lugar ij–ésimo de la matriz Σ^{-1}. Entonces,*

$$\mathrm{Var}[X_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2] = \frac{1}{\sigma^{11}}.$$

DEMOSTRACION: De acuerdo con el Teorema 1.1, p. 13,

$$\sigma_{X_1|\boldsymbol{X}_2=\boldsymbol{x}_2} = \sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{1.58}$$

Por otra parte, por el Lema 1.3, p. 16, sabemos que:

$$|\Sigma| = |\sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}||\Sigma_{22}|. \tag{1.59}$$

De (1.58) y (1.59) se deduce entonces que $\sigma_{X_1|\boldsymbol{X}_2=\boldsymbol{x}_2} = \frac{|\Sigma|}{|\Sigma_{22}|} = 1/\sigma^{11}$.

Lema 1.8 *Sea* $\boldsymbol{Y} = Z\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *con* Z *de orden* $n \times p$ *y* $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$. *Sea* $Q = \text{mín}_{\boldsymbol{\beta}} ||\boldsymbol{Y} - Z\boldsymbol{\beta}||^2 = ||\boldsymbol{Y} - Z\hat{\beta}||^2$. *Entonces:*

$$Q \sim \sigma^2\chi^2_{n-p} \tag{1.60}$$
$$Q = 1/w^{11} \tag{1.61}$$

siendo $W^{-1} = [w^{ij}]$ *y* $W = \begin{pmatrix} \boldsymbol{Y}'\boldsymbol{Y} & \boldsymbol{Y}'Z \\ Z'\boldsymbol{Y} & Z'Z \end{pmatrix}$.

DEMOSTRACION: Que $Q \sim \sigma^2\chi^2_{n-p}$ lo sabemos por teoría de regresión lineal; Q no es otra cosa que SSE, la suma de cuadrados de los residuos al ajustar $\boldsymbol{Y}$ sobre las Z. Por consiguiente,

$$Q = ||(I - Z(Z'Z)^{-1}Z')\boldsymbol{Y}||^2 \tag{1.62}$$
$$= \boldsymbol{Y}'(I - Z(Z'Z)^{-1}Z')\boldsymbol{Y} \tag{1.63}$$
$$= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'Z(Z'Z)^{-1}Z'\boldsymbol{Y} \tag{1.64}$$

Por otra parte, de la definición de W se tiene (empleando el mismo procedimiento que en la demostración del Lema 1.3, p. 16) que:

$$|W| = |\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'Z(Z'Z)^{-1}Z'\boldsymbol{Y}||Z'Z| \tag{1.65}$$

De (1.64) y (1.65) se deduce entonces que $Q = \frac{|W|}{|Z'Z|} = 1/w^{11}$.

Lema 1.9 *Sea* $W \sim W_d(n, \Sigma)$, $n \geq d$. *Entonces:*

1. $\frac{\sigma^{11}}{w^{11}} \sim \chi^2_{n-d+1}$ *es independiente de* w_{ij}, $i, j = 2, \ldots, d$.

2. $\frac{\boldsymbol{\ell}'\Sigma^{-1}\boldsymbol{\ell}}{\boldsymbol{\ell}'W^{-1}\boldsymbol{\ell}} \sim \chi^2_{n-d+1}$, *para cualquier* $\boldsymbol{\ell} \neq \mathbf{0}$.

DEMOSTRACION: $W \sim W_d(n,\Sigma) \Longleftrightarrow W = X'X = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i{}'$ con $\boldsymbol{X}_i \sim N_d(\mathbf{0},\Sigma)$. Si regresáramos la primera variable sobre todas las restantes, de acuerdo con el Lema 1.7, p. 23 anterior,

$$Q = ||\boldsymbol{X}^{(1)} - \sum_{i=2}^{d} \hat{\beta}_i \boldsymbol{X}^{(i)}||^2 \sim \frac{1}{\sigma^{11}} \chi^2_{n-(d-1)}$$

Además, Q es independiente de las columnas de X empleadas como regresores: $\boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(d)}$. Por otra parte, $Q = 1/w^{11}$. Por consiguiente,

$$1/w^{11} \sim (1/\sigma^{11})\chi^2_{n-(d-1)} \tag{1.66}$$

$$\sigma^{11}/w^{11} \sim \chi^2_{n-(d-1)}. \tag{1.67}$$

Para demostrar la segunda parte, sea L una matriz ortogonal $d \times d$ cuya fila superior fuera: $\boldsymbol{\ell}'/||\boldsymbol{\ell}||$. Siempre puede encontrarse una matriz así. Entonces, $LWL' \sim W_d(n, L\Sigma L')$. Como,

$$(LWL')^{-1} = LW^{-1}L' \tag{1.68}$$

$$(L\Sigma L')^{-1} = L\Sigma^{-1}L' \tag{1.69}$$

se tiene que:

$$\frac{\boldsymbol{\ell}'\Sigma^{-1}\boldsymbol{\ell}}{\boldsymbol{\ell}'W^{-1}\boldsymbol{\ell}} = \frac{\boldsymbol{\ell}'\Sigma^{-1}\boldsymbol{\ell}/||\boldsymbol{\ell}||^2}{\boldsymbol{\ell}'W^{-1}\boldsymbol{\ell}/||\boldsymbol{\ell}||^2} \tag{1.70}$$

$$= \frac{(L\Sigma^{-1}L')_{11}}{(LW^{-1}L')_{11}} \tag{1.71}$$

$$= \frac{(L\Sigma L')^{11}}{(LWL')^{11}} \tag{1.72}$$

$$= \chi^2_{n-d+1} \tag{1.73}$$

aplicando (1.53). Es de resaltar que la distribución no depende de $\boldsymbol{\ell}$.

Teorema 1.4 *Si $Z^2 = n\boldsymbol{Y}'W^{-1}\boldsymbol{Y}$ con $\boldsymbol{Y} \sim N_d(\mathbf{0},\Sigma)$, $n \geq d$ y $W \sim W_d(n,\Sigma)$, siendo $\boldsymbol{Y}$ y W independientes (y siguiendo por tanto Z^2 una distribución $T^2_{d,n}$), entonces:*

$$\frac{n-d+1}{d}\frac{Z^2}{n} \sim \mathcal{F}_{d,n-d+1}$$

DEMOSTRACION:

$$\frac{Z^2}{n} = \boldsymbol{Y}'W^{-1}\boldsymbol{Y} = \frac{\boldsymbol{Y}'\Sigma^{-1}\boldsymbol{Y}}{\boldsymbol{Y}'\Sigma^{-1}\boldsymbol{Y}/\boldsymbol{Y}'W^{-1}\boldsymbol{Y}} \tag{1.74}$$

El numerador de (1.74) se distribuye como una χ^2 con d grados de libertad, y el denominador como una χ^2 con $n-d+1$ grados de libertad. Además, como ponía de manifiesto el lema anterior, ambos son independientes, de donde se sigue la distribución $\mathcal{F}$ de Snedecor del cociente.

1.8. Distribución de Wilks y asociadas

Multitud de contrastes univariantes resultan de efectuar cocientes de sumas de cuadrados, que debidamente normalizadas siguen, bajo el supuesto de normalidad de las observaciones, distribución $\mathcal{F}$ de Snedecor. Cuando las observaciones son multivariantes, las "sumas de cuadrados" son formas cuadráticas generalizadas, con distribuciones de Wishart, y el cociente entre determinantes de las mismas puede verse como generalización de los contrastes univariantes.

Definición 1.3 *Supongamos dos matrices aleatorias E y H con distribuciones respectivas,*

$$H \sim W_p(\nu_H, \Sigma) \tag{1.75}$$
$$E \sim W_p(\nu_E, \Sigma) \tag{1.76}$$

independientes. Entonces, el cociente:

$$\frac{|E|}{|E+H|}$$

sigue la distribución conocida como lambda de Wilks *de dimensión p y con grados de libertad ν_H y ν_E, que denotaremos por $\Lambda(p, \nu_H, \nu_E)$.*

La distribución anterior se conoce también como distribución U.

En las aplicaciones surgen de modo muy natural matrices de Wishart E y H asociadas a "suma de cuadrados de los residuos" y "suma de cuadrados atribuible a la hipótesis H". La Tabla 1.1 muestra el paralelismo existente entre algunos productos de matrices Wishart y cocientes de sumas de cuadrados habituales en regresión y ANOVA univariantes.

Cuadro 1.1: Equivalencia entre estadísticos uni- y multivariantes.

Matriz	**Distribución multivariante**	**Análogo univariante**	**Distribución univariante**
$E^{-\frac{1}{2}}HE^{-\frac{1}{2}}$	Beta tipo II multivariante	$\hat{\sigma}_H^2/\hat{\sigma}_E^2$	$\frac{\nu_E}{\nu_H}\mathcal{F}_{\nu_E,\nu_H}$
$(E+H)^{-\frac{1}{2}}H(E+H)^{-\frac{1}{2}}$	Beta tipo I multivariante	$\frac{\hat{\sigma}_H^2}{\hat{\sigma}_H^2+\hat{\sigma}_E^2}$	Beta$(\frac{\nu_E}{2}, \frac{\nu_H}{2})$

Los siguientes teoremas sobre los valores propios de las matrices en la Tabla 1.1 y sus análogas no simétricas HE^{-1} y $H(E+H)^{-1}$ son de utilidad.

Teorema 1.5 *Sean E y H matrices simétricas y definidas positivas. Entonces los valores propios de HE^{-1} son no negativos y los de $H(E+H)^{-1}$ no negativos y menores que 1.*

DEMOSTRACION:

$$\begin{aligned} |HE^{-1} - \phi I| = 0 &\Leftrightarrow |HE^{-\frac{1}{2}} - \phi E^{\frac{1}{2}}| = 0 \\ &\Leftrightarrow |E^{-\frac{1}{2}} H E^{-\frac{1}{2}} - \phi I| = 0 \end{aligned}$$

Es claro que $E^{-\frac{1}{2}} H E^{-\frac{1}{2}}$ es semidefinida positiva, pues para cualquier $\boldsymbol{x}$ tenemos que $\boldsymbol{x}' E^{-\frac{1}{2}} H E^{-\frac{1}{2}} \boldsymbol{x} = \boldsymbol{z}' H \boldsymbol{z}$, en que $\boldsymbol{z} = E^{-\frac{1}{2}} \boldsymbol{x}$.

Sean entonces $\phi_1, \ldots, \phi_d$ los valores propios de HE^{-1}. Tenemos de manera enteramente similar que los de $H(E+H)^{-1}$ son soluciones de

$$\begin{aligned} |H(E+H)^{-1} - \theta I| = 0 &\Leftrightarrow |H - \theta(E+H)| = 0 \\ &\Leftrightarrow |(1-\theta)H - \theta E| = 0 \\ &\Leftrightarrow \left| HE^{-1} - \frac{\theta}{1-\theta} I \right| = 0 \end{aligned}$$

lo que evidencia que

$$\phi_i = \frac{\theta_i}{1-\theta_i}, \qquad (i = 1, \ldots, d)$$

y por tanto

$$\theta_i = \frac{\phi_i}{1+\phi_i}. \qquad (i = 1, \ldots, d)$$

claramente comprendido entre 0 y 1.

Hay diversas tabulaciones de funciones de interés de dichos valores propios cuando las matrices E y H son Wishart independientes: del mayor de ellos, de la suma, del producto, etc., funciones todas ellas que se presentan de modo natural como posibles estadísticos de contraste en las aplicaciones. Un examen de las relaciones entre los diversos estadísticos se posterga a las Secciones 3.3 y 4.3.

1.9. Contrastes en la distribución normal

El supuesto de normalidad encuentra parcial justificación en el teorema central del límite: si las influencias sobre un sistema son múltiples, aproximadamente incorreladas entre sí, y sin ninguna que tenga una importancia dominadora del total, cabe esperar que el resultado se distribuirá de modo aproximadamente normal.

En la práctica, ello resulta mucho más problemático con variables multivariantes que univariantes. Tiene interés disponer de contrastes que permitan evaluar el ajuste a una normal tanto en el caso uni- como multivariante. En lo que sigue se introducen algunos de esos contrastes.

Debe tenerse presente que, incluso aunque el supuesto de normalidad parezca claramente inadecuado, muchos de los procedimientos desarrollados bajo el mismo continúan dando resultados aceptables. En lo sucesivo trataremos de indicar en cada caso como afecta el incumplimiento del supuesto de normalidad a los contrastes y estimaciones.

1.9.1. Diagnósticos de normalidad univariante

Podría, desde luego, emplearse un contraste de ajuste "todo terreno", como la prueba χ^2 o el test de Kolmogorov-Smirnov, descritos en cualquier texto básico de Estadística (por ej., **?**, p. 249). Pero hay contrastes especializados que dan habitualmente mejor resultado cuando la hipótesis de ajuste a contrastar es la de normalidad.

Gráficos QQ. Una de las pruebas más simples e ilustrativas para evaluar el ajuste de una muestra $y_1, \ldots, y_n$ a una distribución normal consiste en construir su gráfico QQ. Se hace de la siguiente manera:

1. Se ordena la muestra, obteniendo $y_{(1)} \leq \ldots \leq y_{(n)}$. Entonces $y_{(i)}$ es el cuantil $\frac{i}{n}$ muestral —deja a su izquierda o sobre él una fracción $\frac{i}{n}$ de la muestra—. Habitualmente se considera como el cuantil $\frac{(i-\frac{1}{2})}{n}$ (corrección de continuidad).

2. Se obtienen (mediante tablas o por cualquier otro procedimiento) los cuantiles $\frac{(i-\frac{1}{2})}{n}$ de una distribución $N(0,1)$, es decir, los valores $q_1 \leq \ldots \leq q_n$ verificando:
$$\int_{-\infty}^{q_i} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = \frac{(i-\frac{1}{2})}{n}.$$

3. Se hace la gráfica de los puntos $(q_i, y_{(i)})$, $i = 1, \ldots, n$.

Es fácil ver que en el supuesto de normalidad los puntos deberían alinearse aproximadamente sobre una recta. Si no presentara forma aproximadamente rectilínea, tendríamos motivo para cuestionar la normalidad.

Contraste de Shapiro-Wilk. Está basado en el cociente del cuadrado de la mejor, o aproximadamente mejor, estimación lineal insesgada de la desviación standard dividida por la varianza muestral. El numerador se construye tomando una combinación lineal de los valores ordenados de la muestra, con coeficientes proporcionados en **?**. Lógicamente, cada tamaño de muestra requiere unos coeficientes diferentes. En su formulación original, era de aplicación sólo a muestras reducidas —con $n \leq 50$ aproximadamente—. No obstante, trabajo posterior (ver **?**) ha permitido extenderlo a tamaños muestrales tan grandes como $n \leq 5000$. Una alternativa para n muy grande es el contraste de D'Agostino a continuación.

Observación 1.4

Contraste de D'Agostino. El contraste de D'Agostino (ver **?**; tablas en **?** reproducidas en **?** y en el Apéndice) emplea el estadístico

$$D = \frac{\sum_{i=1}^{n} \left[i - \frac{1}{2}(n+1) \right] y_{(i)}}{\sqrt{n^3 \sum_{i=1}^{n} (y_{(i)} - \overline{y})^2}} \tag{1.77}$$

o alternativamente su expresión aproximadamente centrada y tipificada

$$Y = \frac{\sqrt{n}\left(D - (2\sqrt{\pi})^{-1}\right)}{0{,}02998598}. \tag{1.78}$$

Requiere $n > 50$. Su distribución para diferentes n está tabulada. Es un contraste "ómnibus", sin una alternativa predefinida. No obstante, el valor de Y proporciona información acerca de la naturaleza de la desviación de la muestra analizada respecto al comportamiento normal: cuando la kurtosis es más de la esperada bajo una hipótesis normal, Y tiende a tomar valores negativos. Lo contrario sucede cuando la muestra presenta menos kurtosis de la esperable en una normal.

Hay otros varios contrastes, explotando una idea similar o comparando la simetría y kurtosis de la muestra con las esperables bajo la hipótesis de normalidad: véase **?**, Sec. 4.4 para un resumen.

1.9.2. Diagnósticos de normalidad multivariante

Un paso previo consistirá en examinar la normalidad de las distribuciones marginales unidimensionales: esta es necesaria, pero no suficiente, para la normalidad multivariante, que es más restrictiva que la mera normalidad de las marginales. Hay un caso, no obstante, en que la normalidad de las marginales si implica normalidad multivariante: el caso de independencia, como resulta fácil comprobar.

Puede pensarse en explotar las ideas en los contrastes univariantes descritos, pero hay que hacer frente a problemas adicionales: no hay una ordenación natural en el espacio p-dimensional, y tropezamos rápidamente con la

"maldición de la dimensionalidad" *(dimensionality curse)*. Lo primero es claro; para adquirir alguna intuición sobre la "maldición de la dimensionalidad" es bueno considerar el siguiente ejemplo.

Ejemplo 1.3 *(en un espacio de elevada dimensionalidad, los puntos quedan casi siempre "lejos")* Consideremos un espacio de dimensión dos; los puntos cuyas coordenadas no difieran en más de una unidad, distan a lo sumo (en distancia euclídea) $\sqrt{2}$. En R^3, la distancia sería $\sqrt{3}$ y, en general, $\sqrt{p}$ en R^p. Alternativamente podríamos pensar en los siguientes términos. El volumen de una hiper-esfera de radio r en p dimensiones tiene por expresión

$$S_p = \frac{\pi^{p/2} r^p}{\Gamma(\frac{p}{2}+1)}. \tag{1.79}$$

Esta fórmula da para $p = 2$ y $p = 3$ las familiares fórmulas de la superficie del círculo y volumen de la esfera[2]. Cuando $p = 3$, la esfera de radio unidad ocupa un volumen de $4\pi/3 = 4{,}1887$; el cubo circunscrito (de lado 2, por tanto) tiene un volumen de 8. De los puntos en el cubo, más de la mitad quedan a distancia menos de 1 del centro de la esfera. Cuando la dimensión p crece, la razón de volúmenes de la hiper-esfera y el hiper-cubo circunscritos es

$$\frac{\pi^{p/2}}{2^p \Gamma(\frac{p}{2}+1)}, \tag{1.80}$$

rápidamente decreciente a cero. Casi todo el volumen de un cubo en $p \gg 3$ dimensiones está en las "esquinas". No hay apenas puntos a corta distancia del centro de la esfera.

Lo que el ejemplo sugiere es que una muestra, salvo de tamaño descomunal, será siempre escasa si el número de dimensiones es alto, y ello no permite concebir muchas esperanzas en cuanto a la potencia que podamos obtener.

Contraste de Gnanadesikan y Kettenring. Dada una muestra $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ proponen construir los estadísticos,

$$u_i = \frac{n}{(n-1)^2} (\boldsymbol{y}_i - \overline{\boldsymbol{y}})' S^{-1} (\boldsymbol{y}_i - \overline{\boldsymbol{y}}) \tag{1.81}$$

que se demuestra siguen una distribución $B(\alpha, \beta)$ con α y β definidos así:

$$\alpha = \frac{p-1}{2p} \tag{1.82}$$

$$\beta = \frac{n-p-2}{2(n-p-1)}. \tag{1.83}$$

[2]Basta recordar que $\Gamma(r) = (r-1)\Gamma(r-1)$, $\Gamma(1) = 1$ y $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Los cuantiles de una $B(\alpha, \beta)$ vienen dados por

$$v_i \quad = \quad \frac{i-\alpha}{n-\alpha-\beta+1}, \tag{1.84}$$

lo que sugiere hacer la gráfica de los puntos $(v_i, u_{(i)})$ y comprobar su alineación sobre una recta. La separación de la recta es indicativa de violación de la hipótesis de normalidad multivariante.

Al igual que en la sección anterior, cabe pensar en contrastes formales que ayuden a nuestro juicio subjetivo sobre la falta de linealidad o no de los puntos mencionados. Como estadístico puede utilizarse

$$D^2_{(n)} \quad = \quad \max_i D^2_i, \tag{1.85}$$

en que $D^2_i = (\boldsymbol{y}_i - \overline{\boldsymbol{y}})' S^{-1} (\boldsymbol{y}_i - \overline{\boldsymbol{y}})$. Los valores críticos están tabulados en **?**.

Un hecho de interés es que el contraste está basado en las cantidades D_i, que son de interés en si mismas como medida de la "rareza" de puntos muestrales —miden la lejanía de cada punto al vector de medias estimado de la muestra en distancia de Mahalanobis—. El contraste reseñado puede por tanto verse también como un contraste de presencia de puntos extraños o *outliers*.

Otros contrastes. Se han propuesto otros contrastes, como el de **?**, que investiga la asimetría y kurtosis en la muestra en relación con la esperable en una normal multivariante.

1.9.3. Búsqueda de *outliers*

Es en general mucho más difícil en espacios de elevada dimensionalidad que en una, dos o tres dimensiones, donde es posible la visualización.

Un método atrayente es el siguiente: sea S la estimación habitual de la matriz de covarianzas basada en una muestra de tamaño n y sea S_{-i} el mismo estimador prescindiendo de la observación i-ésima. Consideremos el estadístico:

$$W = \max_i \frac{|(n-2)S_{-i}|}{|(n-1)S|} \tag{1.86}$$

Si hubiera alguna observación que fuera un *outlier*, "hincharía" mucho la estimación de la matriz de covarianzas, y esperaríamos que W tuviera un valor "pequeño"; por tanto, W tendrá su región crítica por la izquierda. Se puede demostrar que

$$W = 1 - \frac{n D^2_{(n)}}{(n-1)^2} \tag{1.87}$$

con $D_{(n)}$ definido con en (1.85), p. 31, lo que permite emplear para el contraste basado en W las tablas en **?**.

Alternativamente, definamos

$$F_i = \frac{n-p-1}{p}\left(1 - \frac{nD_i^2}{(n-1)^2}\right)^{-1} \qquad (i = 1, \ldots, n) \tag{1.88}$$

Entonces, $F_i \overset{\text{iid}}{\sim} F_{p,n-p-1}$ y

$$P\left(\max_i F_i > f\right) = 1 - [P(F < f)]^n \tag{1.89}$$

en que F es una variable con distribución $\mathcal{F}$ de Snedecor. Obsérvese que ambos contrastes están relacionados:

$$F_{(n)} \overset{\text{def}}{=} \max_i F_i = \frac{n-p-1}{p}\left(\frac{1}{W} - 1\right). \tag{1.90}$$

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

1.1 Las funciones de R `qqnorm` y `shapiro.test` (ésta última en el paquete `ctest`) permiten realizar con comodidad gráficas QQ y el contraste de Shapiro-Wilk respectivamente.

Capítulo 2

Inferencia en poblaciones normales multivariantes.

2.1. Inferencia sobre el vector de medias.

Como estimador de $\boldsymbol{\mu}$ empleamos habitualmente $\overline{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X_i}$, que es el estimador máximo verosímil si la distribución es normal multivariante. Como estimador de la matriz de covarianzas puede emplearse $S = (1/N)\sum_{i=1}^{N}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})(\boldsymbol{X_i}-\overline{\boldsymbol{X}})'$ (máximo verosímil, sesgado) o $N(N-1)^{-1}S = (N-1)^{-1}\sum_{i=1}^{N}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})(\boldsymbol{X_i}-\overline{\boldsymbol{X}})'$ (insesgado). Es habitualmente irrelevante cual de ellos se emplee, en especial si N es moderadamente grande. En los desarrollos que siguen emplearemos S.

2.1.1. Contraste sobre el vector de medias conocida Σ.

Como $\overline{\boldsymbol{X}} \sim N_d(\boldsymbol{\mu}, \frac{1}{N}\Sigma)$, tenemos que:

$$N(\overline{\boldsymbol{X}}-\boldsymbol{\mu})'\Sigma^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}) \sim \chi^2_d$$

Para contrastar $H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu_0}$ calcularíamos el valor del estadístico

$$Q_0 = N(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0})'\Sigma^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0}),$$

rechazando la hipótesis al nivel de significación α si $Q_0 > \chi^2_{d,\alpha}$.

2.1.2. Contraste sobre el vector de medias con Σ desconocida.

Como,

$$NS \sim W_d(N-1,\Sigma) \tag{2.1}$$
$$\sqrt{N}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}) \sim N_d(\mathbf{0},\Sigma) \tag{2.2}$$

y además son independientes, podemos asegurar que bajo la hipótesis nula $H_0\colon \boldsymbol{\mu}=\boldsymbol{\mu_0}$ se verifica

$$N(N-1)(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0})'(NS)^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0}) \sim T^2_{d,N-1},$$

o sea,

$$(N-1)(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0})'S^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0}) \sim T^2_{d,N-1}.$$

Por consiguiente,

$$\frac{N-1-d+1}{d}\frac{T^2_{d,N-1}}{N-1} \sim \mathcal{F}_{d,N-1-d+1} \tag{2.3}$$
$$\frac{N-d}{d}(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0})'S^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0}) \sim \mathcal{F}_{d,N-d} \tag{2.4}$$

El rechazo se producirá al nivel de significación α si el estadístico supera $\mathcal{F}^{\alpha}_{d,N-d}$.

2.1.3. Contraste de igualdad de medias en dos poblaciones con matriz de covarianzas común.

Si tenemos dos muestras,

$$\text{Muestra 1}: \quad \boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_{N_1} \tag{2.5}$$
$$\text{Muestra 2}: \quad \boldsymbol{Y}_1,\boldsymbol{Y}_2,\ldots,\boldsymbol{Y}_{N_2} \tag{2.6}$$

procedentes de sendas poblaciones normales multivariantes con matriz de covarianzas común Σ, entonces:

$$\overline{\boldsymbol{X}} = \frac{1}{N_1}\sum_{i=1}^{N_1}\boldsymbol{X_i} \tag{2.7}$$
$$\overline{\boldsymbol{Y}} = \frac{1}{N_2}\sum_{j=1}^{N_2}\boldsymbol{Y_j} \tag{2.8}$$
$$\tag{2.9}$$
$$N_1S_1 = \sum_{i=1}^{N_1}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})(\boldsymbol{X_i}-\overline{\boldsymbol{X}})' \sim W_d(N_1-1,\Sigma) \tag{2.10}$$
$$N_2S_2 = \sum_{j=1}^{N_2}(\boldsymbol{Y_j}-\overline{\boldsymbol{Y}})(\boldsymbol{Y_j}-\overline{\boldsymbol{Y}})' \sim W_d(N_2-1,\Sigma) \tag{2.11}$$

Por consiguiente, $S = (N_1S_1 + N_2S_2)/(N_1 + N_2)$ es un estimador de Σ que hace uso de información en ambas muestras, y $(N_1+N_2)S \sim W_d(N_1+N_2-2, \Sigma)$. Bajo la hipótesis H_0: $E[\boldsymbol{X}] = E[\boldsymbol{Y}] = \boldsymbol{\mu_0}$, $E(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}) = \mathbf{0}$. Por otra parte,

$$\Sigma_{(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})} = \frac{1}{N_1}\Sigma + \frac{1}{N_2}\Sigma = \frac{(N_1+N_2)}{N_1N_2}\Sigma.$$

Por consiguiente, bajo H_0,

$$\begin{aligned}
\sqrt{\frac{N_1N_2}{N_1+N_2}}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}) &\sim N_d(\mathbf{0}, \Sigma) \\
(N_1+N_2-2)\frac{N_1N_2}{(N_1+N_2)^2}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}})'S^{-1}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}) &\sim T^2_{d,N_1+N_2-2} \\
\frac{N_1+N_2-d-1}{d}\frac{N_1N_2}{(N_1+N_2)^2}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}})'S^{-1}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}) &\sim \mathcal{F}_{d,N_1+N_2-d-1}.
\end{aligned}$$

Como en el caso anterior, se producirá el rechazo de la hipótesis nula de igualdad de medias al nivel de significación α cuando el estadístico anterior supere $\mathcal{F}^{\alpha}_{d,N_1+N_2-d-1}$.

2.1.4. Contraste de hipótesis lineales generales sobre el vector de medias de una única población.

Supongamos que la hipótesis que deseamos contrastar es expresable en la forma H_0: $C\boldsymbol{\mu} = \boldsymbol{\delta}$, siendo δ un vector $q \times 1$ y C una matriz $q \times d$ de rango q.

De acuerdo con la teoría en la Sección anterior, bajo H_0 : $\sqrt{N}(C\overline{\boldsymbol{X}} - \boldsymbol{\delta}) \sim N_q(\mathbf{0}, C\Sigma C')$, y $NCSC' \sim W_q(N-1, C\Sigma C')$. Por consiguiente:

$$N(N-1)(C\overline{\boldsymbol{X}} - \boldsymbol{\delta})'(NCSC')^{-1}(C\overline{\boldsymbol{X}} - \boldsymbol{\delta}) \sim T^2_{q,N-1} \tag{2.12}$$

$$(N-1)(C\overline{\boldsymbol{X}} - \boldsymbol{\delta})'(CSC')^{-1}(C\overline{\boldsymbol{X}} - \boldsymbol{\delta}) \sim T^2_{q,N-1} \tag{2.13}$$

$$\frac{N-q}{q}(C\overline{\boldsymbol{X}} - \boldsymbol{\delta})'(CSC')^{-1}(C\overline{\boldsymbol{X}} - \boldsymbol{\delta}) \sim \mathcal{F}_{q,N-q} \tag{2.14}$$

siendo de nuevo la región crítica la formada por la cola derecha de la distribución (valores grandes del estadístico producen el rechazo de la hipótesis de contraste).

Ejemplo 2.1 Supongamos que estamos interesados en contrastar si la resistencia al desgaste de dos diferentes marcas de neumáticos es la misma o no. Este es un problema típico de Análisis de Varianza: montaríamos los dos tipos de neumáticos en diferentes coches y, dentro de cada coche, en diferentes ruedas, y diseñaríamos el experimento de modo que hasta donde fuera posible ningún factor ajeno al tipo de neumático influyera en su duración. Por ejemplo, nos abstendríamos

de probar el primer tipo de neumático siempre en ruedas traseras, y el segundo en ruedas delanteras, etc.

Sin embargo, no siempre podemos controlar todos los factores en presencia. Supongamos que los dos tipos de neumáticos se montan por pares en cada coche, cada tipo en una rueda delantera y una trasera. Obtendríamos de cada coche un vector $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ de valores, los dos primeros correspondiendo al primer tipo de neumático y los dos siguientes al segundo. Salvo que hayamos diseñado el experimento con total control del tipo de conductor, estilo de conducción, trayecto, tiempo atmosférico, etc., *no es prudente dar por supuesta la independencia entre las componentes de cada vector,* como sería necesario para hacer un análisis de varianza univariante ordinario. En efecto, todas ellas han sido influenciadas por factores comunes —como coche, conductor, trayecto recorrido—.

Si $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_4)$ es el vector de medias, la hipótesis de interés podría expresarse así:

$$C\boldsymbol{\mu} = \mathbf{0}$$

con

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

El contraste haría entonces uso de (2.14).

2.1.5. Contraste de hipótesis lineales sobre los vectores de medias de dos poblaciones.

Sean dos poblaciones normales multivariantes, con matriz de covarianzas común Σ, de las que poseemos sendas muestras aleatorias simples:

$$\text{Muestra 1}: \qquad \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{N_1} \tag{2.15}$$

$$\text{Muestra 2}: \qquad \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{N_2} \tag{2.16}$$

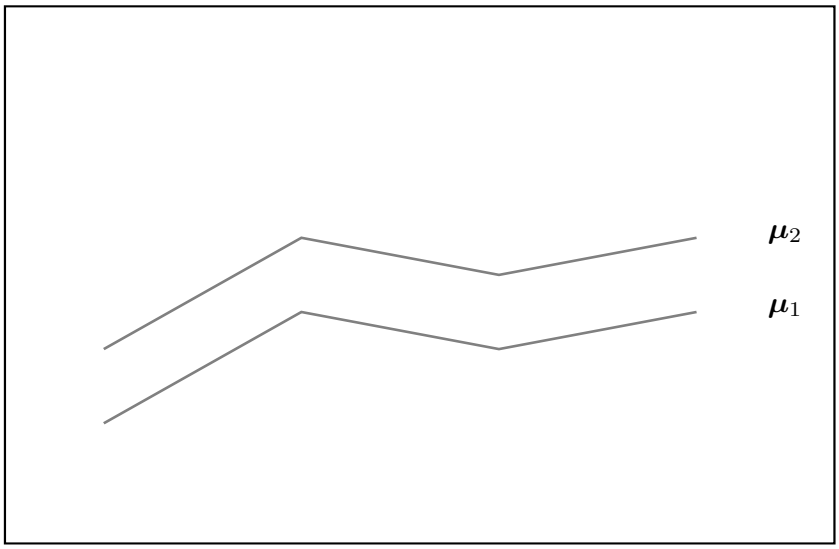
Si la hipótesis $H_0\colon C\boldsymbol{\mu}_1 - C\boldsymbol{\mu}_2 = \boldsymbol{\delta}$ es cierta y C es una matriz $q \times d$ de rango q, se verifica,

$$\begin{aligned} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}(C\overline{\boldsymbol{X}} - C\overline{\boldsymbol{Y}} - \boldsymbol{\delta}) &\sim N_q(\mathbf{0}, C\Sigma C') \\ (N_1 + N_2)S = N_1 S_1 + N_2 S_2 &\sim W_d(N_1 + N_2 - 2, \Sigma) \\ (N_1 + N_2)CSC' &\sim W_q(N_1 + N_2 - 2, C\Sigma C'), \end{aligned}$$

y por tanto,

$$\ell(C\overline{\boldsymbol{X}} - C\overline{\boldsymbol{Y}} - \boldsymbol{\delta})'[(N_1 + N_2)CSC']^{-1}(C\overline{\boldsymbol{X}} - C\overline{\boldsymbol{Y}} - \boldsymbol{\delta}) \sim T^2_{q, N_1+N_2-2}$$

Figura 2.1: Disposición de dos vectores de medias paralelos



con

$$\ell \;=\; \frac{N_1 N_2}{N_1+N_2}(N_1+N_2-2),$$

que tras simplificar proporciona:

$$k(C\overline{\boldsymbol{X}}-C\overline{\boldsymbol{Y}}-\boldsymbol{\delta})'(CSC')^{-1}(C\overline{\boldsymbol{X}}-C\overline{\boldsymbol{Y}}-\boldsymbol{\delta}) \;\sim\; \mathcal{F}_{q,N_1+N_2-q-1} \quad (2.17)$$

con

$$k \;=\; \frac{N_1+N_2-q-1}{q}\frac{N_1 N_2}{(N_1+N_2)^2}.$$

Ejemplo 2.2 Contrastes de esta naturaleza surgen de forma habitual. Hay veces en que la hipótesis de interés no se refiere a la igualdad de los vectores de medias, sino a su forma. Por ejemplo, sean $\boldsymbol{X}_i$ e $\boldsymbol{Y}_j$ vectores aleatorios dando para los sujetos i-ésimo (respectivamente, j-ésimo) de dos poblaciones las sensibilidades auditivas a sonidos de diferentes frecuencias.

Si una de las poblaciones agrupa a jóvenes y otra a ancianos, la hipótesis de igualdad de medias no tendría mayor interés: podemos esperar menor sensibilidad en los mayores. Podría interesarnos en cambio contrastar si los vectores de medias son paralelos (véase Figura 2.1). Es decir, si la esperable pérdida de audición de los ancianos se produce de forma uniforme sobre todas las frecuencias consideradas, o si por el contrario se pierde más sensibilidad para sonidos graves, agudos, u otros. Tal hipótesis se traduciría a una hipótesis de desplazamiento uniforme del vector de medias de una población respecto al de la otra.

Es fácil ver como llevar a cabo dicho contraste con ayuda de (2.17): bastaría tomar

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

y $\boldsymbol{\delta} = \mathbf{0}$.

2.2. Inferencia sobre el coeficiente de correlación entre dos v.a. normales X_1, X_2.

Si $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}' \sim N_2(\boldsymbol{\mu}, \Sigma)$, $Z = \sum_{i=1}^{n} (\boldsymbol{X_i} - \overline{\boldsymbol{X}})(\boldsymbol{X_i} - \overline{\boldsymbol{X}})'$ se distribuye como $W_2(n-1, \Sigma)$. El coeficiente de correlación muestral al cuadrado, $R^2_{X_1,X_2}$, es entonces $Z^2_{12}/Z_{11}Z_{22}$, y su función de densidad puede obtenerse por transformación de la de la Z. Omitimos los detalles[1]. Puede comprobarse que la función de densidad de $R = R_{X_1,X_2}$ (prescindimos de los subíndices por comodidad notacional) es:

$$\begin{aligned} f_R(r) &= \frac{(1-\rho^2)^{n/2}}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}(1-r^2)^{(n-3)/2} \\ &\quad \times \left[\left(\Gamma\left(\frac{n}{2}\right)\right)^2 + \sum_{p=1}^{\infty} \frac{(2\rho r)^p}{p!}\left(\Gamma\left(\frac{n+p}{2}\right)\right)^2\right] \qquad (|r| < 1) \end{aligned}$$

De ella se deduce que:

$$E[R] = \rho + O\left(\frac{1}{n}\right) \tag{2.18}$$

$$\mathrm{Var}[R] = \frac{(1-\rho^2)^2}{n} + O\left(\frac{1}{n^{3/2}}\right). \tag{2.19}$$

Bajo la hipótesis nula $H_0 : \rho = 0$ la densidad se simplifica notablemente:

$$f_R(r) = \frac{1}{B\left(\frac{1}{2}, \frac{n-1}{2}\right)}(1-r^2)^{(n-3)/2} \qquad (|r| < 1)$$

y $T^2 = (n-1)R^2/(1-R^2)$ sigue una distribución $\mathcal{F}_{1,n-1}$, lo que permite contrastar fácilmente la hipótesis de nulidad. Por otra parte, Fisher mostró que

$$Z = \frac{1}{2}\log_e \frac{1+R}{1-R} = \tanh^{-1} R$$

[1] Pueden consultarse en **?** p. 135.

se distribuye aproximadamente como:

$$Z \sim N\left[\frac{1}{2}\log_e \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right]$$

para n "grande", lo que permite construir intervalos de confianza para ρ. La aproximación anterior es válida en el caso normal, y resulta fuertemente afectada por la kurtosis.

2.3. Inferencia sobre la matriz de covarianzas.

Existen contrastes para una gran variedad de hipótesis sobre la matriz de covarianzas de una población normal, o sobre las matrices de covarianzas de más de una población: **?** y **?** son referencias adecuadas. Sólo a título de ejemplo, señalaremos los estadísticos empleados en el contraste de dos hipótesis particulares.

2.3.1. Contraste de igualdad de matrices de covarianzas en dos poblaciones normales.

Sean dos poblaciones normales multivariantes de las que poseemos sendas muestras:

$$\text{Muestra 1}: \quad \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{N_1} \sim N_d(\boldsymbol{\mu}_1, \Sigma_1) \tag{2.20}$$

$$\text{Muestra 2}: \quad \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{N_2} \sim N_d(\boldsymbol{\mu}_2, \Sigma_2) \tag{2.21}$$

Sean,

$$S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(\boldsymbol{X_i} - \overline{\boldsymbol{X}})(\boldsymbol{X_i} - \overline{\boldsymbol{X}})' \tag{2.22}$$

$$S_2 = \frac{1}{N_2}\sum_{j=1}^{N_2}(\boldsymbol{Y_j} - \overline{\boldsymbol{Y}})(\boldsymbol{Y_j} - \overline{\boldsymbol{Y}})' \tag{2.23}$$

$$S = \frac{1}{N_1+N_2}(N_1S_1 + N_2S_2) \tag{2.24}$$

$$N = N_1 + N_2 \tag{2.25}$$

los estimadores habituales de las matrices de covarianzas en cada población y de la matriz de covarianzas conjunta. Sea,

$$\ell = \frac{|S|^{-N/2}}{|S_1|^{-N_1/2}|S_2|^{-N_2/2}} \tag{2.26}$$

Bajo la hipótesis nula $H_0\colon \Sigma_1 = \Sigma_2$, $-2\log_e \ell \sim \chi^2_{\frac{1}{2}d(d+1)}$ asintóticamente.

2.3.2. Contraste de diagonalidad por bloques de la matriz de covarianzas de una única población normal.

Bajo la hipótesis $H_0\colon \Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$, y con la notación habitual, se tiene:

$$\Lambda \overset{\text{def}}{=} \frac{|S|}{|S_{11}||S_{22}|} = \frac{|S_{11} - S_{12}S_{22}^{-1}S_{21}||S_{22}|}{|S_{11}||S_{22}|} = \frac{|S_{11,2}|}{|S_{11}|}. \tag{2.27}$$

Bajo la hipótesis nula, la matriz en el numerador es una Wishart $W_p(N - q - 1, \Sigma_{11})$ y la del denominador $W_p(N-1, \Sigma_{11})$. Por otra parte, como $\boldsymbol{X}_1 = E[\boldsymbol{X}_1|\boldsymbol{X}_2] + (\boldsymbol{X}_1 - E[\boldsymbol{X}_1|\boldsymbol{X}_2])$ es una descomposición de $\boldsymbol{X}_1$ en sumandos independientes, tenemos que: $S_{11} = S_{11,2} + (S_{11} - S_{11,2})$ descompone S_{11} en la suma de dos Wishart independientes. Por tanto,

$$\Lambda = \frac{|S_{11,2}|}{|S_{11,2} + (S_{11} - S_{11,2})|} \sim \Lambda_{p,q,N-q-1}$$

lo que sugiere un modo de hacer el contraste.

Existen diferentes aproximaciones para la distribución Λ. Para valores ausentes en tablas, puede emplearse la aproximación

$$-(N - \frac{1}{2}(p + q + 3)) \log_e \Lambda \sim \chi^2_{pq},$$

o alternativamente

$$\frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{gl_2}{gl_1} \quad \sim \quad \mathcal{F}_{gl_1, gl_2}$$

en que

$$\begin{aligned} gl_1 &= pq \\ gl_2 &= wt - \frac{1}{2}pq + 1 \\ w &= N - \frac{1}{2}(p + q + 3) \\ t &= \sqrt{\frac{p^2q^2 - 4}{p^2 + q^2 - 5}}. \end{aligned}$$

Observación 2.1 $\lambda = \Lambda^{\frac{N}{2}}$ con Λ definida en (2.27) sería la razón generalizada de verosimilitudes bajo las hipótesis respectivas: $H_0 : \Sigma_{12} = 0$ versus $H_a : \Sigma$ general. Un resultado asintótico utilizable en general cuando las hipótesis son (como en este caso) anidadas, establece que

$$-2 \log_e \lambda \sim \chi^2_n$$

siendo n la diferencia de parámetros adicionales que especifica la hipótesis nula respecto de la alternativa. En nuestro caso, $n = pq$, porque la hipótesis nula prescribe pq parámetros nulos (las covarianzas contenidas en el bloque Σ_{12}).

El mismo resultado asintótico se ha empleado en el apartado anterior para aproximar la distribución de ℓ en (2.26). Más detalles sobre contrastes razón generalizada de verosimilitudes pueden encontrarse en **?**, p. 84 y **?**.

2.3.3. Contraste de esfericidad

Sea $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ una muestra procedente de una población $N_p(\boldsymbol{\mu}, \Sigma)$. Estamos interesados en contrastar si la matriz de covarianzas es de la forma $\Sigma = \sigma^2 I$, lo que se traduciría en contornos de igual densidad que serían superficies o hiper-superficies esféricas.

El contraste se efectúa haciendo uso de la técnica de la razón de verosimilitudes (Observación 2.1), que en este caso proporciona:

$$L = \left[\frac{|S|}{(\text{traza}(S)/p)^p}\right]^{\frac{N}{2}}. \tag{2.28}$$

Por tanto, asintóticamente,

$$-2\log_e L = -N\log_e\left[\frac{|S|}{(\text{traza}(S)/p)^p}\right] \sim \chi^2_{\frac{p(p+1)}{2}-1}.$$

Los grados de libertad de la χ^2 son la diferencia de parámetros entre una matriz de covarianzas general ($\frac{p(p+1)}{2}$, habida cuenta de la simetría) y los de otra con estructura escalar $\sigma^2 I$ (sólamente uno).

El estadístico en (2.28) puede escribirse en función de los valores propios de S así:

$$L = \left[\frac{|\prod_{i=1}\lambda_i|}{(\sum_{i=1}^p \lambda_i/p)^p}\right]^{\frac{N}{2}}.$$

El cociente en la expresión anterior es (la potencia de orden p) de la media geométrica a la media aritmética de los autovalores, y por tanto un índice de su disimilaridad, tanto más pequeño cuanto más desiguales sean éstos; lo que es acorde con la intuición.

Una mejor aproximación a la distribución χ^2 se logra sustituyendo $-2\log_e L$ por el estadístico

$$L' = -\left(\nu - \frac{2p^2+p+2}{6p}\right)\log_e\left[\frac{|\prod_{i=1}^p \lambda_i|}{(\sum_{i=1}^p \lambda_i/p)^p}\right],$$

en que ν es el número de grados de libertad de la Wishart que ha dado lugar a S: $N-1$ si ha sido estimada a partir de una sóla muestra con media

desconocida, y $N-k$ si ha sido estimada a partir de k muestras en cada una de las cuales se ha ajustado una media.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

2.1 Mostrar que el estadístico T^2 de Hotelling

$$(N-1)(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0})'S^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu_0}) \tag{2.29}$$

empleado para el contraste multivariante de $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0}$, tomará un valor significativo al nivel α sólo si existe un vector de coeficientes $\boldsymbol{a}$ tal que $H_0 : \boldsymbol{a}'\boldsymbol{\mu} = \boldsymbol{a}'\boldsymbol{\mu_0}$ resulta rechazada al mismo nivel α por un contraste t de Student univariante ordinario.

Capítulo 3

Análisis de varianza multivariante

3.1. Introducción

Los modelos de Análisis de Varianza Multivariante (MANOVA) son una generalización directa de los univariantes. Lo único que varía es que la respuesta que se estudia es un vector *para cada observación*, en lugar de una variable aleatoria escalar. Ello conlleva que las sumas de cuadrados cuyos cocientes proporcionan los contrastes de las diferentes hipótesis, sean ahora formas cuadráticas generalizadas. Los estadísticos de contraste, por su parte, serán cocientes de determinantes (con distribución Λ de Wilks) o diferentes funciones de valores propios de ciertas matrices.

Un descripción del modelo univariante puede encontrarse en casi cualquier texto de regresión: **?**, **?** o **?**, por mencionar sólo algunos. **?**, Cap. 20 y 21 contiene una presentación autocontenida de los modelos ANOVA y MANOVA.

La exposición que sigue presupone familiaridad con el modelo de análisis de varianza univariante.

3.2. Modelo MANOVA con un tratamiento

Estudiamos una característica multivariante $\boldsymbol{Y_{ij}}$ que suponemos generada así:

$$\begin{aligned} \boldsymbol{Y_{ij}} &= \boldsymbol{\mu}_i + \boldsymbol{\epsilon_{ij}} = \boldsymbol{\mu} + \boldsymbol{\alpha_i} + \boldsymbol{\epsilon_{ij}} && (3.1) \\ \boldsymbol{\epsilon}_{ij} &\sim N(\mathbf{0}, \Sigma) && (3.2) \end{aligned}$$

En (3.1), $\boldsymbol{Y_{ij}}$ es el vector de valores que toma la v.a. multivariante estudiada para el caso j-ésimo sujeto al tratamiento i-ésimo. De existir un efecto atribuible al nivel i-ésimo del tratamiento, éste vendría recogido por el vector $\boldsymbol{\alpha}_i$. Supondremos el mismo número de casos estudiados con cada nivel del único tratamiento (es decir, consideraremos sólo el caso de diseño equilibrado): hay k niveles y la muestra incluye n casos tratados con cada nivel.

La hipótesis de interés más inmediato sería:

$$\begin{aligned} H_0 : &\quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k \qquad (\Leftrightarrow \boldsymbol{\alpha}_i = \mathbf{0} \quad \forall i) \\ \text{versus } H_a : &\quad \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \quad \text{para algún } i, j. \end{aligned}$$

De un modo enteramente similar a como sucede en el caso ANOVA univariante, la suma generalizada de cuadrados en torno a la media $\boldsymbol{Y}_{\cdot\cdot}$ se descompone así:

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{Y_{ij}} &- \boldsymbol{Y}_{\cdot\cdot})(\boldsymbol{Y_{ij}} - \boldsymbol{Y}_{\cdot\cdot})' \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{Y_{ij}} - \boldsymbol{Y_{i\cdot}} + \boldsymbol{Y_{i\cdot}} - \boldsymbol{Y}_{\cdot\cdot})(\boldsymbol{Y_{ij}} - \boldsymbol{Y_{i\cdot}} + \boldsymbol{Y_{i\cdot}} - \boldsymbol{Y}_{\cdot\cdot})' \\ &= \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{Y_{ij}} - \boldsymbol{Y_{i\cdot}})(\boldsymbol{Y_{ij}} - \boldsymbol{Y_{i\cdot}})'}_{E} + \underbrace{n\sum_{i=1}^{k}(\boldsymbol{Y_{i\cdot}} - \boldsymbol{Y}_{\cdot\cdot})(\boldsymbol{Y_{i\cdot}} - \boldsymbol{Y}_{\cdot\cdot})'}_{H} \end{aligned}$$

Ahora bien, la teoría anterior (en particular, el Teorema 1.3, p. 21), muestra que las matrices aleatorias E y H en la expresión anterior tienen distribuciones respectivas,

$$\begin{aligned} E &\sim W(k(n-1), \Sigma) && (3.3) \\ H &\overset{H_0}{\sim} W(k-1, \Sigma). && (3.4) \end{aligned}$$

La distribución de E se sigue de los supuestos; la de H es correcta cuando la hipótesis nula es cierta. Además, hay independencia entre ambas matrices Wishart, en virtud del Teorema 1.3. En consecuencia, bajo la hipótesis nula,

$$\Lambda = \frac{|E|}{|E+H|} \sim \Lambda_{p,k-1,k(n-1)}.$$

Si H_0 no se verifica, H "engordará": será una Wishart no central. Son valores pequeños del estadístico Λ anterior los que cabe interpretar como evidencia contra la hipótesis nula.

3.3. Relación entre diversos contrastes

Observemos que si $\delta_1, \ldots, \delta_p$ son los valores propios de $E^{-1}H$,

$$\Lambda = \frac{|E|}{|E+H|} = \prod_{i=1}^{p} \left\{ \frac{1}{1+\delta_i} \right\}. \tag{3.5}$$

El estadístico de contraste es una particular función de los autovalores de $E^{-1}H$. No es la única elección posible: hay otras que mencionamos brevemente.

Estadístico máxima raíz de Roy.

$$\theta = \frac{\delta_1}{1+\delta_1}.$$

Estadístico de Pillai.

$$V = \sum_{i=1}^{p} \frac{\delta_i}{1+\delta_i}.$$

Estadístico de Lawley–Hotelling.

$$U = \sum_{i=1}^{p} \delta_i.$$

De todos ellos hay tabulaciones que permiten contrastar H_0 con comodidad. Su comportamiento es diferente dependiendo del tipo de incumplimiento de la hipótesis H_0. Por ejemplo, el estadístico de Roy está particularmente indicado cuando los vectores de medias $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ están aproximadamente alineados: esto hace crecer el primer valor propio de H y de $E^{-1}H$. En cambio, cuando los vectores de medias son diferentes y no están alineados, los otros estadísticos proporcionarán en general más potencia. Volveremos sobre esta cuestión en la Sección 4.3, p. 52.

3.4. Modelos MANOVA con dos o más tratamientos

De modo análogo a como sucede en el caso univariante, un modelo MANOVA con dos tratamientos supone que la respuesta (multivariante) Y_{ijk} (correspondiente al k-ésimo caso, tratado con los niveles i y j de los tratamientos A y B respectivamente) se genera alternativamente de una de las

Cuadro 3.1: Tabla de Análisis de Varianza para un modelo con dos tratamientos e interacción

Fuente	**Suma cuadrados**	**G.L.**
A	$H_A = KJ\sum_{i=1}^{I}(\boldsymbol{Y_{i..}} - \boldsymbol{Y_{...}})(\boldsymbol{Y_{i..}} - \boldsymbol{Y_{...}})'$	$I-1$
B	$H_B = KI\sum_{j=1}^{J}(\boldsymbol{Y_{.j.}} - \boldsymbol{Y_{...}})(\boldsymbol{Y_{.j.}} - \boldsymbol{Y_{...}})'$	$J-1$
AB	$H_{AB} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\boldsymbol{Y_{ij.}} - \boldsymbol{Y_{i..}} - \boldsymbol{Y_{.j.}} + \boldsymbol{Y_{...}})$ $\times(\boldsymbol{Y_{ij.}} - \boldsymbol{Y_{i..}} - \boldsymbol{Y_{.j.}} + \boldsymbol{Y_{...}})'$	$(I-1)(J-1)$
Error	$E = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\boldsymbol{Y_{ijk}} - \boldsymbol{Y_{ij.}})(\boldsymbol{Y_{ijk}} - \boldsymbol{Y_{ij.}})'$	$IJ(K-1)$
Total	$T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\boldsymbol{Y_{ijk}} - \boldsymbol{Y_{...}})(\boldsymbol{Y_{ijk}} - \boldsymbol{Y_{...}})'$	$IJK-1$

siguientes formas (sin y con interacción, respectivamente):

$$\begin{aligned}\boldsymbol{Y_{ijk}} &= \boldsymbol{\mu} + \boldsymbol{\alpha_i} + \boldsymbol{\beta_j} + \boldsymbol{\epsilon_{ijk}} \\ \boldsymbol{Y_{ijk}} &= \boldsymbol{\mu} + \boldsymbol{\alpha_i} + \boldsymbol{\beta_j} + \boldsymbol{\gamma_{ij}} + \boldsymbol{\epsilon_{ijk}}\end{aligned}$$

El análisis es entonces reminiscente del que se realiza en el caso univariante. Las sumas de cuadrados del análisis univariante son ahora sumas de cuadrados generalizadas: matrices que, bajo los supuestos de normalidad multivariante y de vigencia de las respectivas hipótesis de contraste, se distribuyen como Wishart. A título puramente ilustrativo transcribimos en la Tabla 3.1 la partición de la suma generalizada de cuadrados para un modelo con dos tratamientos e interacción.

Podemos ahora construir contrastes para las hipótesis de nulidad de cada uno de los efectos, empleando el estadístico Λ de Wilks, o cualquiera de los presentados en la Sección 3.3. Si empleamos el primero tendríamos, por ejemplo, que bajo la hipótesis $H_A : \boldsymbol{\alpha}_i = \mathbf{0}$ para $i = 1, \ldots, I$,

$$\Lambda_A = \frac{|E|}{|E + H_A|} \sim \Lambda_{p,I-1,IJ(K-1)}$$

y valores suficientemente pequeños de Λ_A conducirían al rechazo de la hipótesis. Similares cocientes de sumas de cuadrados generalizadas permitirían contrastar cada una de las restantes hipótesis de interés.

Salvo el contraste basado en el estadístico de Roy, los demás son bastante robustos a la no normalidad y a la heterogeneidad en las matrices de covarianzas de los vectores de observaciones. Son bastante sensibles, en cambio, a la no independencia de las observaciones. La robustez al incumplimiento de las hipótesis es en general menor cuando aumenta la dimensión.

3.5. Extensiones y bibliografía

Cada modelo ANOVA univariante encuentra una generalización multivariante. Métodos introducidos en el Capítulo 2 tienen también generalización al caso de más de dos poblaciones, en el contexto de modelos MANOVA. Por ejemplo, el modelo MANOVA con un único tratamiento puede verse como una generalización del contraste en la Sección 2.1.3, p. 34. Del mismo modo otros.

Pueden consultarse sobre este tema **?**, Cap. 20 y 21 y **?**, Cap. 6.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

3.1 En S-PLUS, puede realizarse análisis de varianza multivariante mediante la función `manova`. La sintaxis es muy similar a la de la función `lm`, pero la respuesta *debe ser una matriz,* cuya filas son las observaciones. Por ejemplo, podría invocar `manova` así:

```
solucion <- manova(resp ~ diseño,data=frame).
```

La función devuelve (en `solución`) un objeto de tipo `maov`, cuyas componentes pueden examinarse mediante

```
summary(solucion).
```

Los contrastes relacionados en la Sección 3.2 pueden obtenerse mediante la opción `test=` de `summary`, que admite como valores `"wilks lambda"`, `"pillai"`, `"roy largest"` y `"hotelling-lawley"`. Por ejemplo,

```
summary(solucion, test="pillai")
```

realizaría el contraste de Pillai.

Capítulo 4

Análisis de correlación canónica

4.1. Introducción.

Supongamos que tenemos un vector aleatorio $\boldsymbol{X}$ con $(p+q)$ componentes, que particionamos así: $\boldsymbol{X}' = (\boldsymbol{X}_1{}'|\boldsymbol{X}_2{}')$. Sean,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad\qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

la matriz de covarianzas y el vector de medias particionados consecuentemente. Desconocemos la matriz Σ, pero con ayuda de una muestra hemos obtenido su estimador:

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

Estamos interesados en contrastar la hipótesis $H_0\colon \Sigma_{12} = 0$ frente a la alternativa $H_a\colon \Sigma_{12} \neq 0$; es decir, queremos saber si el primer grupo de p variables $(\boldsymbol{X}_1)$ está o no correlado con el segundo grupo de q variables $\boldsymbol{X}_2$. Podríamos enfrentar este problema directamente, contrastando si Σ es o no diagonal por bloques (para lo que hay teoría disponible). Seguiremos una aproximación diferente que, entre otras cosas, hará emerger el concepto de variable canónica y el principio de unión-intersección de Roy.

4.2. Variables canónicas y coeficientes de correlación canónica.

Consideremos variables auxiliares,

$$x = \boldsymbol{a}'\boldsymbol{X}_1 \qquad\qquad y = \boldsymbol{b}'\boldsymbol{X}_2.$$

El coeficiente de correlación entre ambas es:

$$\rho_{x,y}(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}'\Sigma_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\Sigma_{11}\boldsymbol{a}\ \boldsymbol{b}'\Sigma_{22}\boldsymbol{b}}}$$

una estimación del cual es proporcionada por:

$$r_{x,y}(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}'S_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'S_{11}\boldsymbol{a}\boldsymbol{b}'S_{22}\boldsymbol{b}}}$$

Si ambos vectores $\boldsymbol{X}_1, \boldsymbol{X}_2$ fueran independientes, para cualesquiera vectores $\boldsymbol{a}, \boldsymbol{b}$ tendríamos que $\rho_{x,y}(\boldsymbol{a},\boldsymbol{b}) = 0$. De un modo intuitivo, parece pues evidente que debieran ser valores cercanos a cero de $r^2_{x,y}(\boldsymbol{a},\boldsymbol{b})$ los que condujeran a la aceptación de la hipótesis de independencia, en tanto la región crítica estaría formada por los valores $r^2_{x,y}(\boldsymbol{a},\boldsymbol{b})$ superando un cierto umbral (se emplea el cuadrado del coeficiente de correlación para que tenga signo positivo en todo caso).

Obsérvese, sin embargo, que $r^2_{x,y}(\boldsymbol{a},\boldsymbol{b})$ depende de $\boldsymbol{a}$ y de $\boldsymbol{b}$. El método de unión-intersección de Roy maximiza primero $r^2_{x,y}(\boldsymbol{a},\boldsymbol{b})$ respecto de $\boldsymbol{a}, \boldsymbol{b}$ y compara el valor resultante con la distribución del máximo bajo la hipótesis nula. La idea es sustancialmente la misma que cuando se contrastan muchas hipótesis simultáneas.

El problema de maximización de $r^2_{x,y}(\boldsymbol{a},\boldsymbol{b})$ está insuficientemente especificado; multiplicando $\boldsymbol{a}$, $\boldsymbol{b}$, o ambos por una constante cualquiera, $r^2_{x,y}(\boldsymbol{a},\boldsymbol{b})$ no altera su valor. Utilizaremos por ello restricciones de normalización:

$$\boldsymbol{a}'S_{11}\boldsymbol{a} = 1 \qquad\qquad \boldsymbol{b}'S_{22}\boldsymbol{b} = 1$$

Si formamos el lagrangiano,

$$\Phi(\boldsymbol{a},\boldsymbol{b}) = (\boldsymbol{a}'S_{12}\boldsymbol{b})^2 - \lambda(\boldsymbol{a}'S_{11}\boldsymbol{a} - 1) - \mu(\boldsymbol{b}'S_{22}\boldsymbol{b} - 1),$$

derivamos, e igualamos las derivadas a cero, obtenemos:

$$\left(\frac{\partial\Phi(\boldsymbol{a},\boldsymbol{b})}{\partial\boldsymbol{a}}\right)' = 2(\boldsymbol{a}'S_{12}\boldsymbol{b})S_{12}\boldsymbol{b} - 2\lambda S_{11}\boldsymbol{a} = \boldsymbol{0_{p\times 1}} \tag{4.1}$$

$$\frac{\partial\Phi(\boldsymbol{a},\boldsymbol{b})}{\partial\boldsymbol{b}} = 2(\boldsymbol{a}'S_{12}\boldsymbol{b})S_{12}{}'\boldsymbol{a} - 2\mu S_{22}\boldsymbol{b} = \boldsymbol{0_{q\times 1}}. \tag{4.2}$$

Reordenando las anteriores ecuaciones:

$$\begin{aligned} -\lambda S_{11}\boldsymbol{a} + (\boldsymbol{a}'S_{12}\boldsymbol{b})S_{12}\boldsymbol{b} &= \mathbf{0} \quad (4.3)\\ (\boldsymbol{a}'S_{12}\boldsymbol{b})S_{21}\boldsymbol{a} - \mu S_{22}\boldsymbol{b} &= \mathbf{0} \quad (4.4)\end{aligned}$$

Premultiplicando (4.3)–(4.4) por $\boldsymbol{a}'$ y $\boldsymbol{b}'$ obtenemos: $\lambda = \mu = (\boldsymbol{a}'S_{12}\boldsymbol{b})^2 = r^2_{x,y}(\boldsymbol{a},\boldsymbol{b})$, valores que llevados a dichas ecuaciones proporcionan

$$\begin{aligned} -\lambda S_{11}\boldsymbol{a} + \lambda^{\frac{1}{2}}S_{12}\boldsymbol{b} &= \mathbf{0}\\ \mu^{\frac{1}{2}}S_{21}\boldsymbol{a} - \mu S_{22}\boldsymbol{b} &= \mathbf{0}\end{aligned}$$

o sea,

$$\begin{aligned} -\lambda^{\frac{1}{2}}S_{11}\boldsymbol{a} + S_{12}\boldsymbol{b} &= \mathbf{0} \quad (4.5)\\ S_{21}\boldsymbol{a} - \mu^{\frac{1}{2}}S_{22}\boldsymbol{b} &= \mathbf{0} \quad (4.6)\end{aligned}$$

Para que este sistema tenga solución distinta de la trivial ha de verificarse

$$\begin{vmatrix} -\lambda^{\frac{1}{2}}S_{11} & S_{12} \\ S_{21} & -\mu^{\frac{1}{2}}S_{22}\end{vmatrix} = 0, \quad (4.7)$$

o sea, haciendo uso del Lema 1.3,

$$|-\mu^{\frac{1}{2}}S_{22}||-\lambda^{\frac{1}{2}}S_{11} + S_{12}S_{22}^{-1}S_{21}\mu^{-\frac{1}{2}}| = 0 \quad (4.8)$$

Como suponemos S_{22} definida positiva, el primer factor es no nulo, por lo que de (4.8) se deduce:

$$|-\lambda^{\frac{1}{2}}S_{11} + S_{12}S_{22}^{-1}S_{21}\mu^{-\frac{1}{2}}| = |S_{11}||S_{12}S_{22}^{-1}S_{21}S_{11}^{-1} - \lambda I| = 0. \quad (4.9)$$

De nuevo suponiendo que S_{11} es definida positiva, concluimos de (4.9) que

$$|S_{12}S_{22}^{-1}S_{21}S_{11}^{-1} - \lambda I| = 0, \quad (4.10)$$

y por tanto las soluciones de λ son los valores propios de $S_{12}S_{22}^{-1}S_{21}S_{11}^{-1}$. Puesto que λ es también $r^2_{x,y}(\boldsymbol{a},\boldsymbol{b})$, es claro que debemos tomar el *mayor* de los valores propios para resolver nuestro problema de maximización.

El contraste deseado, por tanto, se reduce a comparar dicho λ máximo con su distribución bajo la hipótesis nula. Esta distribución tiene interesantes propiedades: para nada depende de Σ_{11} ni Σ_{22}. Detalles teóricos pueden obtenerse de **?**, p. 301.

Una particularidad del contraste propuesto es que si efectuáramos transformaciones lineales cualesquiera de las variables aleatorias en ambos subvectores, los resultados no se alterarían[1].

[1]Se dice que el contraste es invariante frente a transformaciones lineales no degeneradas. La idea de invariancia es importante en Estadística; es uno de los procedimientos más habituales para restringir la clase de contrastes merecedores de atención. Véase una discusión más completa en **?**, p. 41 y **?**, Sec. 7.3.

En efecto, si $\boldsymbol{Y_1} = A\boldsymbol{X_1}$ e $\boldsymbol{Y_2} = B\boldsymbol{X_2}$ siendo A y B matrices cualesquiera, tenemos que la matriz cuyos valores propios hemos de computar es, en función de las matrices de covarianzas muestrales de $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$,

$$AS_{12}B'(B')^{-1}S_{22}^{-1}B^{-1}BS_{21}A'(A')^{-1}S_{11}^{-1}A^{-1} = AS_{12}S_{22}^{-1}S_{21}S_{11}^{-1}A^{-1} \quad (4.11)$$

Como los valores propios no nulos de CD y de DC son idénticos (supuesto que ambos productos pueden realizarse), los valores propios de la última matriz en (4.11) son idénticos a los de $S_{12}S_{22}^{-1}S_{21}S_{11}^{-1}$.

Calculado λ podemos regresar a (4.5)–(4.6) y obtener $\boldsymbol{a}$ y $\boldsymbol{b}$. Las variables $x = \boldsymbol{a}'\boldsymbol{X}_1$ e $y = \boldsymbol{b}'\boldsymbol{X}_2$, combinaciones lineales de las originales con $\boldsymbol{a}$ y $\boldsymbol{b}$ correspondientes al máximo λ, se denominan *primeras variables canónicas*; son las combinaciones lineales de variables en $\boldsymbol{X}_1$ y en $\boldsymbol{X}_2$ con máxima correlación muestral. Los siguientes valores de λ solución de (6) proporcionan las segundas, terceras, etc. variables canónicas. Hay $s = \text{mín}(p, q)$ pares de variables canónicas, y consecuentemente s coeficientes de correlación canónica. Se demuestra fácilmente que las sucesivas variables canónicas son incorreladas entre sí.

4.3. Relación con otros contrastes

Diferentes modelos multivariantes pueden verse como casos particulares de análisis de correlación canónica. Mencionamos brevemente la relación con MANOVA de un tratamiento; el mismo argumento puede repetirse en conexión con análisis discriminante (Capítulo 12).

Supongamos que el vector $\boldsymbol{X}_1$ agrupa las variables regresandos, y que como vector $\boldsymbol{X}_2$ tomamos variables indicadoras, en número igual al de niveles del único tratamiento. La muestra tendría la siguiente apariencia:

$$\begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} & 1 & 0 & \dots & 0 \\ X_{21} & X_{22} & \dots & X_{2p} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ X_{n_1,1} & X_{n_1,2} & \dots & X_{n_1,p} & 1 & 0 & \dots & 0 \\ X_{n_1+1,1} & X_{n_1+1,2} & \dots & X_{n_1+1,p} & 0 & 1 & \dots & 0 \\ X_{n_1+2,1} & X_{n_1+2,2} & \dots & X_{n_1+2,p} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ X_{N1} & X_{N2} & \dots & X_{Np} & 0 & 0 & \dots & 1 \end{pmatrix}. \quad (4.12)$$

Es decir, un 1 en posición j-ésima en $\boldsymbol{X}_2$ señala que el caso correspondiente ha recibido el tratamiento j-ésimo.

Es ahora intuitivo que, en el caso de que los diferentes niveles de tratamiento no tengan ninguna influencia, no deberíamos esperar ninguna relación lineal entre las variables en $\boldsymbol{X}_1$ y las variables en $\boldsymbol{X}_2$; y en efecto este

es el caso. Contrastar la hipótesis de efecto nulo en MANOVA y de mayor correlación canónica nula es algo equivalente.

En efecto, salvo en una constante, podríamos identificar las matrices Wishart E y H empleadas en el modelo MANOVA de un tratamiento así:

$$\begin{aligned} E &= S_{11} - S_{12}S_{22}^{-1}S_{21} \\ H &= S_{12}S_{22}^{-1}S_{21} \end{aligned}$$

En MANOVA buscábamos los autovalores definidos por la ecuación característica $|E^{-1}H - \delta I| = 0$. Observemos que,

$$\begin{aligned} |E^{-1}H - \delta I| = 0 &\Leftrightarrow |H - \delta E| = 0 && (4.13)\\ &\Leftrightarrow |S_{12}S_{22}^{-1}S_{21} - \delta(S_{11} - S_{12}S_{22}^{-1}S_{21})| = 0 && (4.14)\\ &\Leftrightarrow |(1+\delta)S_{12}S_{22}^{-1}S_{21} - \delta S_{11}| = 0 && (4.15)\\ &\Leftrightarrow |S_{12}S_{22}^{-1}S_{21} - \frac{\delta}{1+\delta}S_{11}| = 0 && (4.16)\\ &\Leftrightarrow |S_{11}^{-1}S_{12}S_{22}^{-1}S_{21} - \frac{\delta}{1+\delta}I| = 0. && (4.17) \end{aligned}$$

Los autovalores de la matriz $E^{-1}H$ están en relación biunívoca con las correlaciones canónicas al cuadrado:

$$\begin{aligned} r_i^2 &= \lambda_i = \frac{\delta_i}{1+\delta_i} \\ \delta_i &= \frac{\lambda_i}{1-\lambda_i}. \end{aligned}$$

Es equivalente contrastar la hipótesis de nulidad de ρ_1^2 (mayor correlación canónica al cuadrado) o la de δ_1 (mayor autovalor de $E^{-1}H$ "anormalmente grande" bajo $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_K$).

Observación 4.1 Incidentalmente, la relación anterior entre los autovalores de una y otra matriz y (3.5), muestra que bajo la hipótesis "Todos los coeficientes de correlación canónica son nulos", el estadístico

$$\prod_{i}^{J-1}(1 - r_i^2) = \prod_{i=1}^{J-1}\frac{1}{1+\delta_i}$$

se distribuye como una Λ de Wilks.

4.4. Interpretación.

A menudo es difícil, pero cuando resulta posible suele ser iluminante. En ocasiones, cualquier pareja formada por una variable en $\boldsymbol{X}_1$ y otra en $\boldsymbol{X}_2$ tiene débil correlación, y hay sin embargo combinaciones lineales de variables en $\boldsymbol{X}_1$ muy correladas con combinaciones lineales de variables en $\boldsymbol{X}_2$. En

este caso, el examen de dichas combinaciones lineales puede arrojar luz sobre aspectos del problema analizado que de otro modo pasarían desapercibidos.

El empleo de contrastes sobre el primer coeficiente de correlación canónica es también el método adecuado cuando investigamos la existencia de correlación entre características no directamente medibles. Por ejemplo. podríamos estar interesados en la hipótesis de si existe relación entre ideología política de los individuos y su nivel cultural. Ninguna de estas dos cosas es medible de manera unívoca, sino que podemos imaginar múltiples indicadores de cada una de ellas: la ideología política podría venir descrita para cada individuo por un vector $\boldsymbol{X}_1$ de variables conteniendo valoraciones sobre diferentes cuestiones. Análogamente sucedería con el nivel cultural. El investigar pares de variables aisladas sería un procedimiento claramente inadecuado; la utilización de contrastes sobre el primer coeficiente de correlación canónica permite contrastar la hipótesis de interés de modo simple y directo.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

4.1 En R puede realizarse análisis de correlación canónica con comodidad utilizando la función `cancor`.

Capítulo 5

Componentes principales.

5.1. Introducción.

Es frecuente el caso en que se tiene un colectivo cada uno de cuyos integrantes puede ser descrito por un vector $\boldsymbol{X}$, de dimensión p. En tales casos, es también frecuente que entre las diferentes componentes del vector $\boldsymbol{X}$ exista cierta correlación, que, en el caso más extremo, haría que alguna de las variables X_i fuera combinación lineal exacta de otra u otras. En tales casos, surge de modo natural la pregunta de si no sería más útil tomar un subconjunto de las variables originales —o quizá un número reducido de variables compuestas, transformadas de las originales— que describiera el colectivo sin gran pérdida de información.

Naturalmente, el problema así planteado es demasiado vago para admitir una solución precisa. Porque, ¿qué significa "sin gran pérdida de información"? Y, ¿qué nuevas variables, distintas de las primitivas, estamos dispuestos a considerar? Los siguientes ejemplos tratan de ilustrar el problema a resolver y motivar la solución que se ofrece en la Sección 5.2.

Ejemplo 5.1 Consideremos un colectivo de niños sobre cada uno de los cuales se han medido las siguientes tres variables:

Variable	Descripción
X_1	Nota obtenida en Matemáticas
X_2	Nota obtenida en idiomas
X_3	Nota obtenida en Ciencias Naturales

Podemos ver cada niño como descrito por un vector aleatorio $\boldsymbol{X}$, procedente de una distribución cuya matriz de covarianzas es R. Imaginemos

también que, calculada la matriz de correlación entre dichas tres variables (en la práctica, dicha matriz de covarianzas sería normalmente estimada a partir de una muestra de niños), obtenemos el resultado siguiente:

$$R = \begin{pmatrix} 1{,}00 & 0{,}68 & 0{,}92 \\ 0{,}68 & 1{,}00 & 0{,}57 \\ 0{,}92 & 0{,}57 & 1{,}00 \end{pmatrix}. \tag{5.1}$$

El examen de la anterior matriz de correlación sugiere lo siguiente: las notas en Matemáticas (X_1) y en Ciencias Naturales (X_3) están estrechamente correlacionadas. Si un niño tiene nota alta en Matemáticas, con bastante seguridad podemos decir que su nota en Ciencias Naturales es también alta. En cambio, la nota en Idioma Moderno muestra también correlación con las otras dos, pero mucho mas baja (0.57 y 0.68 respectivamente).

En resumen, podríamos decir que, aunque descrito por tres variables, cada niño podría sin gran pérdida de información ser descrito por dos: una reflejando su aptitud/interés por las Matemáticas y Ciencias Naturales (quizá la nota media en ambas disciplinas) y otra reflejando su aptitud/interés por el Idioma Moderno.

Observemos el razonamiento implícito que hemos efectuado: dos variables (X_1 y X_3) presentan elevada correlación, *lo que sugiere que la información que aportan es muy redundante*. En efecto, conocido el valor que toma una podríamos conocer con bastante aproximación el valor que toma la otra.

Ejemplo 5.2 La Tabla **??** en el Apéndice **??** recoge los *records* obtenidos por atletas de diferentes nacionalidades en varias especialidades. El simple examen de los mismos, sugiere que quizá no son precisas todas las variables para obtener una buena descripción del nivel del atletismo en los diferentes países. Parece que hay países que destacan en todas las especialidades, y otros que muestran bajo nivel también en todas. ¿Podemos asignar una única "nota media" a cada país sin gran pérdida de información respecto a la que aporta la totalidad de las variables? ¿Es, quizá, precisa más de una nota? Si éste fuera el caso, ¿cómo decidir cuántas "notas", y de qué manera obtenerlas? La Sección que sigue plantea el problema de modo formal, y ofrece una posible solución al mismo.

5.2. Obtención de las componentes principales.

Podemos suponer $\boldsymbol{X}$ centrado[1]. Por simplicidad, limitaremos nuestra atención a variables que puedan obtenerse como combinación lineal de las variables originales. Si éstas formaban para cada elemento de la muestra el

[1]Esto simplifica la notación, sin pérdida de generalidad: si $\boldsymbol{X}$ no fuera centrado, bastaría restarle su vector de medias y resolver el problema resultante.

vector $\boldsymbol{X}$ de dimensión p, consideraremos entonces (no más de p) variables de la forma:

$$\begin{aligned} U_1 &= \boldsymbol{a_1}'\boldsymbol{X} \\ U_2 &= \boldsymbol{a_2}'\boldsymbol{X} \\ &\vdots \\ U_p &= \boldsymbol{a_p}'\boldsymbol{X} \end{aligned} \tag{5.2}$$

El problema, pues, radica en la elección de los vectores de coeficientes $\boldsymbol{a_1}, \ldots, \boldsymbol{a_p}$ que permitan obtener $U_1, \ldots, U_p$ como combinaciones lineales de las variables originales en $\boldsymbol{X}$.

Puesto que la correlación entre variables implica redundancia en la información que aportan, resulta sensato requerir de las nuevas variables $U_1, \ldots, U_p$ que sean incorreladas. Por otra parte, tenemos interés en que las nuevas variables $U_1, \ldots, U_p$ tengan varianza lo más grande posible: en efecto, una variable que tomara valores muy parecidos para todos los elementos de la población (es decir, que tuviera reducida varianza) sería de escaso valor descriptivo[2]. Podríamos entonces enunciar el problema que nos ocupa así:

> Encontrar variables, $U_1, \ldots, U_p$, combinación lineal de las primitivas en $\boldsymbol{X}$, que sean mutuamente incorreladas, teniendo cada U_i varianza máxima entre todas las posibles combinaciones lineales de $\boldsymbol{X}$ incorreladas con $U_1, \ldots, U_{i-1}$.

Las variables U_i verificando las condiciones anteriores se denominan *componentes principales.*

Resolveremos el problema de su obtención secuencialmente; obtendremos primero el vector de coeficientes $\boldsymbol{a_1}$ proporcionando la variable U_1, combinación lineal de $\boldsymbol{X}$, con máxima varianza. Obtendremos luego $\boldsymbol{a_2}$ proporcionando U_2 de varianza máxima bajo la restricción de que U_2 sea incorrelada con U_1. A continuación, obtendremos $\boldsymbol{a_3}$ proporcionando U_3 bajo las restricciones de incorrelación con U_1 y U_2, y así sucesivamente.

Observemos, sin embargo, que si no acotamos el módulo de $\boldsymbol{a_i}$, el problema carece de solución. En efecto, siempre podríamos incrementar la varianza de U_i multiplicando por una constante mayor que uno el correspondiente vector de coeficientes $\boldsymbol{a_i}$. Debemos por consiguiente establecer una restricción sobre los coeficientes, que puede ser $||\boldsymbol{a_i}||^2 = 1$, para $i = 1, \ldots, p$. Con esta restricción, debemos en primer lugar solucionar el siguiente problema:

$$\max_{\boldsymbol{a_1}} E[U_1^2] \qquad \text{condicionado a} \qquad \boldsymbol{a_1}'\boldsymbol{a_1} = 1 \tag{5.3}$$

Obsérvese que si, como hemos supuesto, $E[\boldsymbol{X}] = \boldsymbol{0}$, entonces $E[U_1] = E[\boldsymbol{a_1}'\boldsymbol{X}] = 0$ y $\mathrm{Var}(U_1) = E[U_1^2] = \boldsymbol{a_1}'R\boldsymbol{a_1}$. Teniendo en cuenta esto y

[2]Naturalmente, la varianza de las diferentes variables es función de las unidades de medida; volveremos sobre esta cuestión algo más adelante.

usando la técnica habitual para resolver (5.3) mediante multiplicadores de Lagrange, tenemos que el problema se reduce a:

$$\max_{\boldsymbol{a_1}} \left\{ \boldsymbol{a_1}' R \boldsymbol{a_1} - \lambda[\boldsymbol{a_1}' \boldsymbol{a_1} - 1] \right\}. \tag{5.4}$$

Derivando respecto a $\boldsymbol{a_1}$ e igualando la derivada a $\mathbf{0}$ obtenemos

$$2R\boldsymbol{a_1} - 2\lambda \boldsymbol{a_1} = \mathbf{0}, \tag{5.5}$$

lo que muestra que $\boldsymbol{a_1}$ es un vector propio de R, cuyo valor propio asociado es λ. Como estamos buscando la variable U_1 de máxima varianza, y

$$\mathrm{Var}(U_1) = \boldsymbol{a_1}' R \boldsymbol{a_1} = \lambda \boldsymbol{a_1}' \boldsymbol{a_1} = \lambda, \tag{5.6}$$

debemos tomar como $\boldsymbol{a_1}$ el vector propio de R asociado a λ_1, el mayor de los valores propios de R.

La obtención de $\boldsymbol{a_2}$ es similar. Debemos maximizar ahora $\mathrm{Var}(U_2)$ sujeto a dos restricciones: la de normalización $||\boldsymbol{a_2}||^2 = 1$ y la de incorrelación con U_1. Como

$$\mathrm{Cov}(U_1, U_2) = E\left[\boldsymbol{a_1}' \boldsymbol{X} \boldsymbol{a_2}' \boldsymbol{X}\right] = E[\boldsymbol{a_1}' \boldsymbol{X} \boldsymbol{X}' \boldsymbol{a_2}] = \boldsymbol{a_1}' R \boldsymbol{a_2}, \tag{5.7}$$

el problema a resolver ahora es

$$\max_{\boldsymbol{a_2}} \left\{ \boldsymbol{a_2}' R \boldsymbol{a_2} - \lambda(\boldsymbol{a_2}' \boldsymbol{a_2} - 1) - \mu(\boldsymbol{a_2}' R \boldsymbol{a_1}) \right\}, \tag{5.8}$$

que tomando derivadas respecto a $\boldsymbol{a_2}$, λ y μ proporciona:

$$2R\boldsymbol{a_2} - 2\lambda \boldsymbol{a_2} - \mu R \boldsymbol{a_1} = \mathbf{0} \tag{5.9}$$

$$\boldsymbol{a_2}' \boldsymbol{a_2} = 1 \tag{5.10}$$

$$\boldsymbol{a_2}' R \boldsymbol{a_1} = 0. \tag{5.11}$$

Premultiplicando (5.9) por $\boldsymbol{a_1}'$ y teniendo en cuenta (5.11) obtenemos que $\mu = 0$ y por tanto (5.9) es equivalente a

$$2R\boldsymbol{a_2} - 2\lambda \boldsymbol{a_2} = \mathbf{0}, \tag{5.12}$$

lo que de nuevo muestra que $\boldsymbol{a_2}$ es un vector propio de R. Un razonamiento similar al efectuado en el caso de $\boldsymbol{a_1}$ muestra que $\boldsymbol{a_2}$ es el vector propio asociado al segundo mayor valor propio de de R, λ_2, y que $\mathrm{Var}(U_2) = \lambda_2$.

La obtención de las restantes variables $U_3, \ldots, U_p$ se efectúa de manera similar, con el resultado de que cada una de ellas es una combinación lineal de variables en $\boldsymbol{X}$ con vector de coeficientes $\boldsymbol{a_i}$ que es vector propio de R.

5.3. Propiedades de las componentes principales.

Dado que los vectores de coeficientes $\boldsymbol{a_i}$ son vectores propios de R, si definimos $A = (\boldsymbol{a_1}\vdots\boldsymbol{a_2}\vdots\ldots\vdots\boldsymbol{a_p})$ y $\boldsymbol{U}' = (U_1, U_2, \ldots, U_p)$ tenemos:

$$\boldsymbol{U} = A'\boldsymbol{X} \tag{5.13}$$

$$E\left[\boldsymbol{U}\boldsymbol{U}'\right] = A'RA = \Lambda \tag{5.14}$$

siendo Λ una matriz diagonal con los valores propios de R en la diagonal principal. La ecuación (5.14) muestra la incorrelación entre las componentes principales, así como el hecho, ya apuntado, de ser sus respectivas varianzas iguales a los valores propios de R. Como A es ortogonal, pre- y postmultiplicando (5.14) por A y A' obtenemos:

$$R = A\Lambda A' = \sum_{i=1}^{p} \lambda_i \boldsymbol{a_i}\boldsymbol{a_i}' \tag{5.15}$$

La ecuación (5.15) muestra R como una suma de matrices de rango uno.

Observación 5.1 De acuerdo con el *teorema de Eckart-Young*, la mejor aproximación R^* de rango k de R, en el sentido de minimizar traza$((R^* - R)(R^* - R)')$ es $\sum_{i=1}^{k} \lambda_i \boldsymbol{a_i}\boldsymbol{a_i}'$.

Las ecuaciones (5.14)–(5.15) muestran también que traza(R) = traza(Λ) = $\sum \lambda_i$, dado que:

$$p = \text{traza}(R) = \text{traza}(A\Lambda A') = \text{traza}(\Lambda A'A) = \text{traza}(\Lambda) = \sum_{i=1}^{p} \lambda_i.$$

En consecuencia, incluso sin calcular todos los valores propios, puede calcularse con facilidad la fracción que representan sobre el total de traza. Esto es de interés porque algunos de los métodos numéricos para cálculo de valores propios los obtienen por orden de magnitud; se puede entonces detener el proceso de obtención cuando $\sum \lambda_i$ representa una fracción "suficiente"sobre el total de la traza.

Ejemplo 5.3 La matriz de correlación estimada R de los datos en el Apéndice **??**, Tabla **??**, es:

	m100	m200	m400	m800	m1500	Km5	Km10	Maratón
m100	1.000	0.922	0.841	0.756	0.700	0.619	0.632	0.519
m200	0.922	1.000	0.850	0.806	0.774	0.695	0.696	0.596
m400	0.841	0.850	1.000	0.870	0.835	0.778	0.787	0.704
m800	0.756	0.806	0.870	1.000	0.918	0.863	0.869	0.806
m1500	0.700	0.774	0.835	0.918	1.000	0.928	0.934	0.865
Km 5	0.619	0.695	0.778	0.863	0.928	1.000	0.974	0.932
Km10	0.632	0.696	0.787	0.869	0.934	0.974	1.000	0.943
Maratón	0.519	0.596	0.704	0.806	0.865	0.932	0.943	1.000

Cuadro 5.1: Valores propios de R

i (1)	λ_i (2)	% s/traza (3)	$\sum_i \lambda_i$ (4)	% (4) s/traza (5)
1	6.622	82.77	6.622	82.77
2	0.877	10.96	7.499	93.73
3	0.159	1.99	7.658	95.72
4	0.124	1.55	7.782	97.27
5	0.080	1.00	7.862	98.27
6	0.068	0.85	7.930	99.12
7	0.046	0.58	7.976	99.70
8	0.023	0.29	7.999	99.99

Puede verse la acusada correlación existente entre casi todas las variables, siendo la más baja 0.519 (entre las marcas de 100 metros y la de Maratón). A la vista de dicha matriz de correlación, cabría imaginar que un número reducido de componentes principales bastaría para describir adecuadamente el colectivo.

Al diagonalizar la matriz de correlación se obtienen los valores propios en la Tabla 5.1. La primera componente principal es la combinación lineal de variables originales *tipificadas* con coeficientes dados por el vector propio

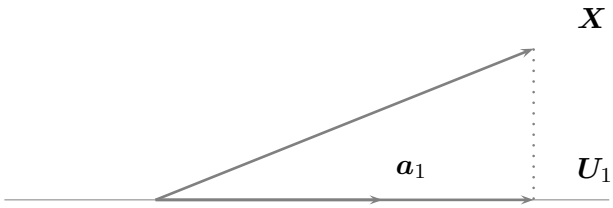
$$\boldsymbol{a_1} = \begin{pmatrix} 0{,}317 \\ 0{,}337 \\ 0{,}355 \\ 0{,}368 \\ 0{,}373 \\ 0{,}364 \\ 0{,}366 \\ 0{,}342 \end{pmatrix}$$

es decir:

$$U_1 = 0{,}317X_1 + 0{,}337X_2 + \ldots + 0{,}342X_8$$

Nótese que si los vectores propios lo son de la matriz de correlación, las variables cuya combinación lineal da las U_i son las de $\boldsymbol{X}$ tipificadas; si los vectores propios lo son de la matriz de covarianzas, las variables a emplear son las originales (centradas, si se quiere que $E[U_i] = 0$). Los vectores propios $\boldsymbol{a_i}$ de la matriz de covarianzas y la matriz de correlación *no* están relacionados de ninguna manera obvia. En la Tabla 5.1 puede verse que, salvo los dos primeros, los valores propios son muy reducidos; parece adecuado describir datos como los exhibidos mediante dos componentes principales. La elección del número de componentes principales a emplear es en principio subjetiva; una regla frecuentemente seguida (cuando las variables han sido tipificadas) es tomar tantas

Figura 5.1: U_i es proyección de $\boldsymbol{X}$ sobre $\boldsymbol{a_i}$



componentes principales como valores propios mayores que la unidad haya, pero esto no es nada absoluto ni que deba realizarse ciegamente.

5.4. Interpretación geométrica.

Si examinamos la ecuación (5.13) podemos interpretar fácilmente los valores que toman las componentes principales $U_1, \ldots, U_p$ como las coordenadas en un cierto sistema de ejes.

De (5.13) se deduce que:

$$U_i = \boldsymbol{a_i}'\boldsymbol{X} \tag{5.16}$$

$$U_i = |\boldsymbol{a_i}||\boldsymbol{X}|\cos(\alpha) = |\boldsymbol{X}|\cos(\alpha), \tag{5.17}$$

en que α es el ángulo formado por el vector $\boldsymbol{X}$ y el vector $\boldsymbol{a_i}$; recuérdese que éste último tiene módulo unitario. En consecuencia, U_i es la coordenada del punto $\boldsymbol{X}$ cuando se representa en un sistema de ejes coordenados en las direcciones (ortogonales) dadas por los vectores $\boldsymbol{a_1}, \ldots, \boldsymbol{a_p}$. La Figura 5.1 ilustra esto.

En general, tal como sugiere la Observación 5.1, las primeras k componentes principales proporcionan la mejor representación k-dimensional de los datos, en el sentido de: i) Dar cuenta del máximo de traza de la matriz de covarianza (o correlación), y ii) Permitir reconstruir aproximaciones de las variables originales que yacen en un subespacio k-dimensional del original con la matriz de covarianzas (o correlación) que mejor aproxima la original, en el sentido que dicha Observación 5.1 especifica.

Por ello, una etapa rutinaria en el análisis de datos multivariantes consiste de ordinario en obtener una representación en pocas dimensiones de los datos. Si con dos o tres componentes principales se obtiene una representación fiel, puede hacerse una gráfica bi- o tridimensional cuya mera observación será instructiva. Cosas como agrupamientos suelen ser fáciles de detectar.

A veces, una determinada componente principal puede ser interpretada. En el caso del Ejemplo 5.3, la primera componente principal podría interpretarse como un índice de la calidad atlética de los respectivos países. Si observamos el segundo vector propio,

$$\boldsymbol{a_2} = \begin{pmatrix} -0{,}566 \\ -0{,}461 \\ -0{,}248 \\ -0{,}012 \\ +0{,}139 \\ +0{,}312 \\ +0{,}306 \\ +0{,}438 \end{pmatrix}$$

podemos ver que pondera con signo negativo las cuatro primeras variables, y con signo positivo las cuatro últimas. La variable U_2 tomará valores grandes para aquellos países en que los tiempos en las pruebas de fondo estén por debajo de la media, y los tiempos en las pruebas de velocidad por encima; es una variable que complementa la información proporcionada por U_1, separando los diversos países según sus respectivas especializaciones en fondo o velocidad.

Ejemplo 5.4 La Figura 5.2 muestra un tal mapa, referido a los datos presentados en el Ejemplo 5.3. Puede verse a algunos países muy separados de la principal concentración, en la esquina inferior. La primera componente principal puede interpretarse como midiendo la "calidad general" atlética de cada país (correspondiendo el lado izquierdo a países "mejores"). La segunda componente principal (vertical) separa países con predominio relativo en distancias cortas (que se sitúan hacia la parte superior del gráfico) y con predominio relativo en distancias largas (que se sitúan hacia la parte inferior).

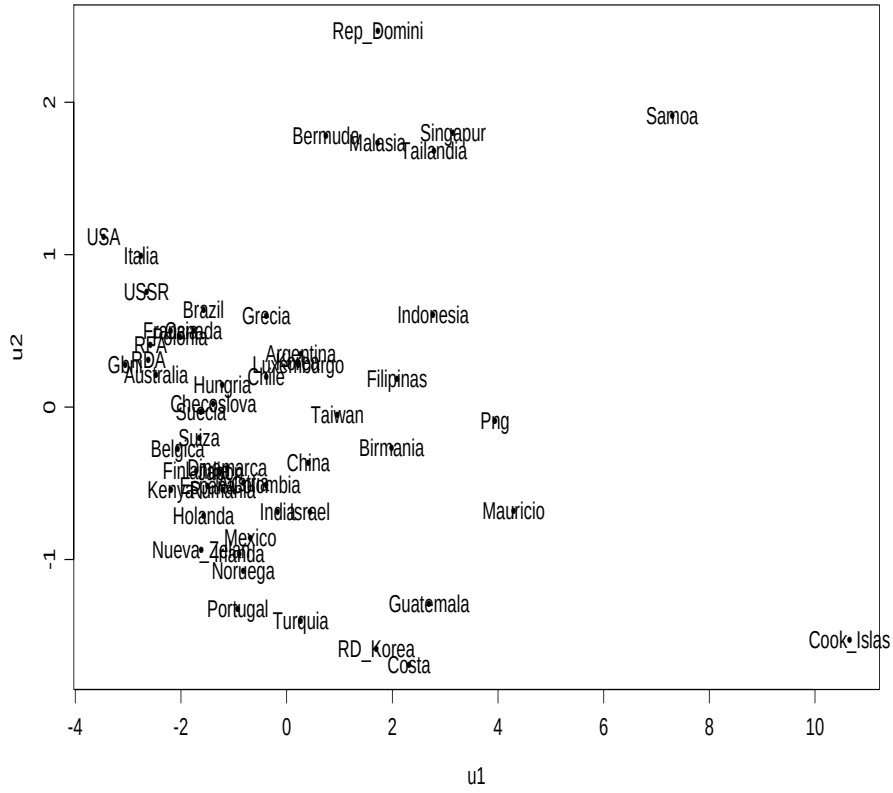
La interpretación de las componentes generales se facilita en ocasiones, como en el caso anterior, atendiendo a los valores que toman los coeficientes a_{ij}. Algunos autores prefieren utilizar como ayuda en la interpretación las correlaciones o covarianzas entre las variables originales y las componentes principales. El argumento es en tales casos que los coeficientes a_{ij} tienen gran varianza. La cuestión está sujeta a controversia: véase por ejemplo el criterio contrario de **?**, p. 361.

5.5. Comentarios adicionales

Es importante reparar en los siguientes aspectos:

1. El empleo de componentes principales no presupone ningún modelo subyacente. Es sólo una técnica, fundamentalmente de naturaleza descriptiva, que obtiene una representación de menor dimensionalidad de un conjunto de puntos en R^p.

Figura 5.2: *Records* representados en el plano generado por U_1 y U_2



2. El método selecciona *un subespacio* de R^p, cuyos ejes vienen dados por las direcciones de $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_k}$, $(k < p)$. Los ejes son ortogonales y en las direcciones de mayor dispersión de los datos. Pero no hay nada que nos fuerce a considerar dichos ejes; lo realmente relevante es la reducción de la dimensionalidad y la fijación de un subespacio adecuado. La base que tomemos del mismo puede escogerse con cualquier criterio conveniente —no tiene por qué estar formada por $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_k}$—.

3. El método se puede emplear tanto con las variables en las escalas originales como con variables tipificadas. Los resultados, en general, son completamente diferentes.

4. Los signos de los $\boldsymbol{a_i}$ son irrelevantes. En efecto, si $\boldsymbol{a_i}$ es vector propio, $-\boldsymbol{a_i}$ también lo es.

En el Capítulo que sigue se introduce el modelo factorial. Por una parte, se hace uso de un modelo explícito, que realiza supuestos acerca del modo de generación de las observaciones. Por otro, en relación a la segunda cuestión mencionada en el apartado anterior, veremos que existen modos alternativos de escoger la base del subespacio de interés, y que ello permite mejorar la interpretabilidad del análisis.

Capítulo 6

Análisis Factorial.

6.1. Introducción.

El Análisis Factorial es un conjunto de técnicas que persiguen identificar factores ocultos. Suponemos que una cierta variable aleatoria multivariante de la que poseemos una muestra se genera así:

$$\boldsymbol{X} = A\boldsymbol{F} + \boldsymbol{L} + \boldsymbol{m} \tag{6.1}$$

En (6.1), $\boldsymbol{F}$ (vector de *factores comunes*) y $\boldsymbol{L}$ (vector de *factores específicos*) son vectores aleatorios, y A es una matríz de constantes. Supondremos en lo que sigue que $\boldsymbol{X}$ ha sido centrado, con lo que prescindiremos del vector de medias $\boldsymbol{m}$. Los respectivos vectores y matrices verifican:

$$\begin{aligned} \boldsymbol{X} &= \text{vector } p \times 1 \\ A &= \text{matriz } p \times k \\ \boldsymbol{F} &= \text{vector } k \times 1 \\ \boldsymbol{L} &= \text{vector } p \times 1 \end{aligned}$$

Se realizan además los siguientes supuestos:

$$E[\boldsymbol{F}] = \mathbf{0}_{(k\times 1)} \tag{6.2}$$
$$E[\boldsymbol{L}] = \mathbf{0}_{(p\times 1)} \tag{6.3}$$
$$E[\boldsymbol{F}\boldsymbol{L}'] = 0_{(k\times p)} \tag{6.4}$$
$$E[\boldsymbol{F}\boldsymbol{F}'] = I_{(k\times k)} \tag{6.5}$$
$$D = E[\boldsymbol{L}\boldsymbol{L}'] = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & d_p \end{pmatrix} \tag{6.6}$$

En (6.1), los factores comunes $\boldsymbol{F}$ influyen en $\boldsymbol{X}$ a traves de los coeficientes en la matriz A; cada uno de los factores específicos en $\boldsymbol{L}$ sólo influye en la variable homóloga. Un modelo como (6.1) parece indicado cuando se investigan fenómenos en que un número grande de variables son concebiblemente causadas por unos pocos factores comunes.

Observación 6.1 Históricamente, la investigación psicométrica proporcionó la motivación inicial para el desarrollo de este tipo de modelos; un vector de items procedente de un *test* sicológico se intentaba poner en correspondencia mediante (6.1) con un número reducido de facetas (inobservables) que supuestamente describen la personalidad.

El problema del Análisis Factorial consiste en estimar A y D. Obsérvese cierta semejanza con el modelo de regresión lineal, pero con la salvedad de que la variable respuesta es multivariante (cada observación es un $\boldsymbol{X}$), los "regresores" $\boldsymbol{F}$ son inobservables, e incluso su número nos es desconocido. Pese a todo ello, las restricciones permiten en general obtener una solución —si bien, como veremos, no única—.

6.2. La igualdad fundamental

De las definiciones se deduce inmediatamente,

Teorema 6.1

$$\Sigma = E[(\boldsymbol{X} - m)(\boldsymbol{X} - \boldsymbol{m})'] = AA' + D \tag{6.7}$$

Demostracion: En efecto,

$$\Sigma = E[(\boldsymbol{X} - \boldsymbol{m})(\boldsymbol{X} - \boldsymbol{m})'] \tag{6.8}$$
$$= E(A\boldsymbol{F} + \boldsymbol{L})(A\boldsymbol{F} + \boldsymbol{L})'] \tag{6.9}$$
$$= E[A\boldsymbol{F}\boldsymbol{F}'A' + A\boldsymbol{F}\boldsymbol{L}' + \boldsymbol{L}\boldsymbol{F}'A' + \boldsymbol{L}\boldsymbol{L}'] \tag{6.10}$$
$$= AA' + D \tag{6.11}$$

La igualdad (6.7), en particular, implica que

$$\begin{aligned} \sigma_{ii} &= \sum_{j=1}^{k} a_{ij}^2 + d_i \qquad (i = 1, \ldots, p) \\ \sigma_{ij} &= \sum_{l=1}^{k} a_{il} a_{jl} \qquad (i \neq j; \quad i, j = 1, \ldots, p) \end{aligned}$$

Se llama *comunalidad* y se denota por h_i^2 a aquélla parte de la varianza de la variable X_i de que dan cuenta los factores comunes, es decir, $h_i^2 = \sum_{j=1}^{k} a_{ij}^2$.

6.3. Análisis Factorial y el objetivo de la parsimonia

Un modelo es una representación estilizada de la realidad, que pretende captar sus rasgos de la manera más simple posible.

> **Observación 6.2** Esto sería una definición si supiéramos qué es la "realidad", qué significa "captar sus rasgos" y qué significa "de la manera más simple posible". Es de temer que no sabemos demasiado bien qué es ninguna de estas cosas, y por tanto la frase anterior sea una tautología o una idiotez. El buscar modelos simples es una regla de economía intelectual, y probablemente no tenga más defensa que la constatación de su enorme eficacia, acreditada desde Guillermo de Ockham hacia acá. Por lo demás, admitiendo una realidad, ¿por qué habría de ser simple y no complicada?

En el contexto en que nos movemos, tomaremos "más simple" por sinónimo de "con el mínimo número de parámetros". Observemos entonces que Σ en el lado izquierdo de (6.7) incluye $\frac{1}{2}p(p+1)$ parámetros diferentes, mientras que, si seleccionamos k como número de factores, el lado derecho requiere $pk + p - \frac{1}{2}k(k-1$ parámetros (pk en la matriz A y otros p adicionales en la diagonal de D, deduciendo $\frac{1}{2}k(k-1)$ porque, como veremos, la solución factorial que obtengamos deja A indeterminada en ese número de parámetros; véase **?**, p. 114, y la Observación 6.3, pág. 70.)

Si k puede hacerse considerablemente menor que p (es decir, si podemos especificar nuestro modelo con muchos menos factores comunes que variables), habremos logrado una reducción considerable en el número de parámetros necesarios, y en este sentido nuestro modelo será más "simple". Llamamos *parsimonia* a esta simplicidad. A título ilustrativo, se recogen los valores de $\frac{1}{2}p(p+1)$ y $pk + p - \frac{1}{2}k(k-1$ para diferentes p y k, y la correspondiente ganancia en parsimonía medida en número de parámetros. Los valores de p y k no son inusuales en problemas como los que se presentan en la práctica.

p	k	**Parámetros** Σ	**Parámetros** $AA' + D$	**Ganancia en parsimonia**
10	3	55	37	18
20	2	210	59	151
20	4	210	94	116
30	3	465	104	349

A la luz de todo lo anterior, podríamos formular el problema a resolver en análisis factorial así:

> "Encontrar matrices A y D verificando (6.7) para una matriz Σ dada, con A teniendo el mínimo número de columnas."

Evidentemente, en la práctica no conocemos Σ y habremos de trabajar con una estimación de la misma. Además, aún cuando el modelo fuera "correcto" (es decir, los datos se generasen realmente tal como especifica (6.1)), la igualdad (6.7) se verificará a lo sumo de modo aproximado. Nuestro objetivo en la práctica será pues obtener una buena reconstrucción de una matriz de covarianzas estimada a partir del producto AA' más una matriz diagonal D.

Ejemplo 6.1 Este ejemplo procede de **?**, quienes a su vez lo toman de un trabajo de Spearman de 1904. Es un caso sumamente simple, pero que ilustra los conceptos anteriores.

Se parte de una matriz de correlación[1], conteniendo las correlaciones entre calificaciones de tres asignaturas (Lenguas Clásicas, Francés e Inglés), estimadas en una muestra de niños. La matriz resulta ser,

$$S = \begin{pmatrix} 1{,}00 & 0{,}83 & 0{,}78 \\ & 1{,}00 & 0{,}67 \\ & & 1{,}00 \end{pmatrix} \tag{6.12}$$

Spearman ajustó un modelo con un sólo factor, es decir,

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} F_1 + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix} \tag{6.13}$$

que implica:

$$\Sigma = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} \end{pmatrix} + \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix} \tag{6.14}$$

de acuerdo con el teorema de Thurstone, (6.7). Sustituyendo Σ en (6.14) por su estimación S tenemos la igualdad matricial

$$\begin{pmatrix} 1{,}00 & 0{,}83 & 0{,}78 \\ & 1{,}00 & 0{,}67 \\ & & 1{,}00 \end{pmatrix} = \begin{pmatrix} \hat{a}_{11} \\ \hat{a}_{21} \\ \hat{a}_{31} \end{pmatrix} \begin{pmatrix} \hat{a}_{11} & \hat{a}_{21} & \hat{a}_{31} \end{pmatrix} + \begin{pmatrix} \hat{d}_1 & 0 & 0 \\ 0 & \hat{d}_2 & 0 \\ 0 & 0 & \hat{d}_3 \end{pmatrix}$$

[1] Sobre el uso de la matriz de covarianzas o correlaciones como punto de partida, valen las observaciones hechas para componentes principales en el Capítulo 5.

de la que obtenemos las ecuaciones:

$$\begin{aligned} 1 &= \hat{a}_{11}^2 + \hat{d}_1 && (6.15)\\ 1 &= \hat{a}_{21}^2 + \hat{d}_2 && (6.16)\\ 1 &= \hat{a}_{31}^2 + \hat{d}_3 && (6.17)\\ 0{,}83 &= \hat{a}_{11}\hat{a}_{21} && (6.18)\\ 0{,}78 &= \hat{a}_{11}\hat{a}_{31} && (6.19)\\ 0{,}67 &= \hat{a}_{21}\hat{a}_{31}. && (6.20) \end{aligned}$$

Tenemos pues seis ecuaciones con seis incógnitas que permiten encontrar una solución "exacta" a partir de la igualdad fundamental (6.7). Tras resolver, el modelo estimado es

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0{,}983 \\ 0{,}844 \\ 0{,}793 \end{pmatrix} F_1 + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix}, \qquad (6.21)$$

y las comunalidades son

$$\begin{aligned} h_1^2 &= 0{,}966\\ h_2^2 &= 0{,}712\\ h_3^2 &= 0{,}629. \end{aligned}$$

Por tanto, el modelo con un único factor da cuenta muy bien de la primera calificación (Lenguas Clásicas), y algo peor de las dos restantes.

6.4. Indeterminación de las soluciones factoriales. Rotaciones

Con el problema planteado como en la Sección anterior, es ahora evidente que la solución no es única. En efecto, si

$$\Sigma = E[(\boldsymbol{X} - \boldsymbol{m})(\boldsymbol{X} - \boldsymbol{m})'] = AA' + D,$$

y G es una matriz ortogonal $(k \times k)$, también será cierto que

$$\Sigma = E[(\boldsymbol{X} - \boldsymbol{m})(\boldsymbol{X} - \boldsymbol{m})'] = AGG'A' + D = BB' + D. \quad (6.22)$$

Por tanto, B será una solución tan válida como A. Obsérvese además de (6.1) se deduce

$$\begin{aligned} \boldsymbol{X} &= AGG'\boldsymbol{F} + \boldsymbol{L} + \boldsymbol{m} && (6.23)\\ &= B\boldsymbol{F}_G + \boldsymbol{L} + \boldsymbol{m} && (6.24) \end{aligned}$$

con $\boldsymbol{F}_G = G'\boldsymbol{F}$ que continúa verificando todas las condiciones impuestas a los factores comunes (6.2)–(6.6), como es fácil comprobar.

Esto tiene enorme trascendencia. Estando las soluciones factoriales indeterminadas hasta el producto por una matriz ortogonal (geométricamente, una rotación, reflexión, o combinación de ambas), somos libres de tomar la solución que más nos convenga. De ordinario, esto permite escoger soluciones con la estructura de A que nos parece más interpretable.

Observación 6.3 Podemos ahora volver al asunto brevemente tocado en la Sección 6.3, acerca del número de grados de libertad consumidos (o parámetros estimados) al encontrar una solución factorial. Si A cuenta con pk parámetros pero está indeterminada, es claro que no hemos consumido de modo efectivo pk grados de libertad, sino menos.

Si reparamos en que las columnas de A deben generar un cierto subespacio de dimensión k, tendremos un modo fácil de persuadirnos de que una solución factorial supone estimar $pk-\frac{1}{2}k(k-1)$ parámetros. En efecto, *cualquier* subespacio de dimensión k de R^p puede generarse mediante una base "escalonada", formada por las columnas de una matriz como

$$\begin{pmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & \dots & 0 \\ a_{31} & a_{32} & a_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{p-1,1} & a_{p-1,2} & a_{p-1,3} & \dots & 0 \\ a_{p1} & a_{p2} & a_{p3} & \dots & a_{pk} \end{pmatrix}; \tag{6.25}$$

y especificar tal matriz requiere precisamente $pk - \frac{1}{2}k(k-1)$ parámetros. Alternativamente, si A está indeterminada hasta el producto por una matriz ortogonal, conservará tantos grados de libertad como existan para fijar una matriz ortogonal $k \times k$. Hay $\frac{1}{2}k(k-1)$ elementos libres en una tal matriz. La primera columna sólo está constreñida a tener módulo unitario ($k-1$ elementos son por tanto libres); la segunda, está además constreñida a ser ortogonal a la primera ($k-2$ elementos libres por tanto); la tercera y sucesivas tienen cada una una restricción adicional. El número total de elementos libres es por tanto $(k-1)+(k-2)+\ldots+1 = \frac{1}{2}k(k-1)$.

Si tenemos cierta margen de maniobra al escoger una solución factorial, desearemos hacerlo de modo que la interpretación resulte favorecida. Idealmente, para poder rotular un factor desearíamos que su influencia alcanzara a algunas de las variables de modo notable, y al resto en absoluto. Por

ejemplo, si tuviéramos una matriz A como,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6.26}$$

recordando que

$$\boldsymbol{X} = A\boldsymbol{F} + \boldsymbol{L} \tag{6.27}$$

razonaríamos así: "El factor F_1 es algo que está relacionado con las variables X_1, X_2 y X_3. Los factores F_2, F_3 y F_4 influyen cada uno en las variables X_4 y X_5, X_6 y X_7 y en X_8 y X_9, respectivamente". El conocimiento de las variables ayudaría así a dotar de interpretación a los factores F_1 a F_4: F_1, por ejemplo, podríamos imaginarlo como lo que quiera que las variables X_1 a X_3 tuvieran en común. Y similarmente con los otros.

Naturalmente, una estructura de ceros y unos, como la del ejemplo anterior, no será muchas veces factible: pero, en la medida de lo posible, desearíamos tender a ella.

Una forma de lograrlo es determinar G de manera que $A_G = AG$ tenga mucho "contraste". Hay varias formas de formalizar esta idea intuitiva hasta convertirla en un problema con solución matemática. En lo que sigue, mencionaremos dos de las propuestas más utilizadas, que ilustran bien el modo de abordar el problema. Más detalles pueden encontrarse en **?**, **?**, **?**, o cualquier texto sobre análisis factorial o multivariante. **?** y **?** son dos de las referencias pioneras. La idea de la rotación *quartimax* es escoger la matriz $A_G = AG$ para la que es máxima la "varianza" *por filas* de los cuadrados de los elementos a_{ij}. La toma del cuadrado obedece a que estamos interesados en lograr términos "grandes" y "pequeños": no nos importa el signo. Maximizamos por ello

$$\frac{1}{k^2}\sum_{i=1}^{p}\left[k\sum_{j=1}^{k}(a_{ij}^2)^2 - \left(\sum_{j=1}^{k}a_{ij}^2\right)^2\right]. \tag{6.28}$$

Esta propuesta logra contraste entre unos términos y otros: pero nada en la forma de la expresion a maximizar impide que los a_{ij} "grandes" se agrupen en la primera columna de la matriz A_G. Ello da lugar a una solución con un factor "general", que parece influir en todas las variables: puede o no ser deseable o fácil de interpretar.

Habitualmente preferimos que cada factor de cuenta del comportamiento de un grupo de variables originales, con las que poder relacionarle. Si es el caso, la rotación *varimax* puede ser más atractiva. Buscamos en ella maximizar

$$\frac{1}{p^2}\sum_{j=1}^{k}\left[p\sum_{i=1}^{p}(a_{ij}^2)^2-\left(\sum_{i=1}^{p}a_{ij}^2\right)^2\right], \tag{6.29}$$

es decir, la "varianza" de los a_{ij}^2 *por columnas*. Ello forzará a que *en cada columna* haya elementos muy grandes y muy pequeños.

Hay algunos detalles adicionales que pueden consultarse en **?**; por ejemplo, en lugar de maximizar las expresiones (6.28) o (6.29) tal cual, frecuentemente se normalizan los elementos de cada fila dividiendo entre la comunalidad: se intenta con ello evitar que las filas de A con elevada comunalidad dominen las expresiones citadas.

6.5. Estimación del modelo

Hemos de hacer frente a dos problemas: determinar el número de factores deseado, y obtener una estimación (inicial, indeterminada) de A. Estimada A, las especificidades y comunalidades quedan también estimadas. Describiremos sólamente dos de los métodos más utilizados.

6.5.1. Método del factor principal

Obsérvese que, si conociéramos las comunalidades (o, equivalentemente, la matriz de especificidades, D), de la igualdad fundamental (6.7) se deduciría que la matriz de covarianzas (o correlación) muestral ha de verificar aproximadamente

$$S - D \approx \hat{A}\hat{A}'; \tag{6.30}$$

ello sugiere emplear alguna estimación de D para computar $S^* = S - \hat{D}$, A continuación, podemos factorizar esta S^* como producto de dos matrices de rango k. Si S^* tiene sus k mayores valores propios positivos, ello no ofrecerá problema: podemos emplear la aproximación

$$S^* \approx \hat{A}\hat{A}', \tag{6.31}$$

en que $\hat{A} = \sum_{i=1}^{k}\sqrt{\lambda_i}\boldsymbol{v}_i$, siendo los λ_i y $\boldsymbol{v}_i$ los valores y vectores propios de S^*.

No es preciso que nos detengamos en la estimación de $\hat{A}$ recién obtenida, sino que podríamos ahora emplearla para obtener una estimación mejor, quizá, de las comunalidades,

$$D_{(2)} = \mathrm{diag}(S - \hat{A}\hat{A}'), \tag{6.32}$$

una estimación actualizada de S^*,

$$S^*_{(2)} = (S - D_{(2)}), \tag{6.33}$$

y consiguientemente una nueva estimación de A por factorización de $S^*_{(2)}$:

$$S^*_{(2)} \approx \hat{A}_{(2)} \hat{A}_{(2)}{}'. \tag{6.34}$$

Con la nueva estimación $\hat{A}_{(2)}$ de A podríamos reiniciar el proceso e iterar hasta convergencia, si se produce (nada garantiza que se produzca, aunque habitualmente se obtiene convergencia cuando k es suficientemente grande).

6.5.2. Método de máxima verosimilitud

Podemos también estimar los parámetros del modelo (6.1) por máxima verosimilitud, si conocemos la distribución de $\boldsymbol{X}$ (en la práctica, ello equivale a suponer normalidad multivariante).

Capítulo 7

Biplots

Estudiaremos en lo que sigue dos técnicas para la representación simultánea de observaciones y variables. La primera —el *biplot*— es un gráfico en el que se representan las observaciones en posiciones dadas por sus dos primeras componentes principales. Sobre el mismo plano se superponen p puntos representando las variables —las columnas de la matriz de datos X en posiciones que hacen interpretables las relaciones entre ellas y las observaciones.

La segunda técnica —el *análisis de correspondencias*— produce de modo similar una representación simultánea de observaciones y variables, y es de aplicación a tablas de contingencia.

A ambas técnicas subyace la *descomposición en valores singulares* de una matriz rectangular, que se presenta a continuación.

7.1. Descomposición en valores singulares.

Sea X una matriz $N \times p$ cualquiera. Mostraremos que puede siempre escribirse como producto de una matriz de columnas ortogonales $N \times p$, una matriz diagonal $p \times p$ con elementos no negativos en la diagonal principal y una matriz ortogonal $p \times p$. La exposición sigue a **?**.

Tanto $X'X$ como XX' son matrices cuadradas simétricas, y por tanto diagonalizables. Para $j = 1, \ldots, p$ hay vectores propios $\boldsymbol{a}_i$ de $X'X$ (y $\boldsymbol{b}_j$ de XX') asociados a valores propios en general no nulos λ_i (para los $\boldsymbol{a}_i$) y ν_j (para los $\boldsymbol{b}_j$).

$$X'\,X\boldsymbol{a}_j = \lambda_j \boldsymbol{a}_j \tag{7.1}$$
$$X\,X'\boldsymbol{b}_j = \nu_j \boldsymbol{b}_j. \tag{7.2}$$

La matriz $X\ X'$ posee además $N-p$ valores propios nulos y correspondientes vectores propios asociados. Los vectores propios $\boldsymbol{a}_j$ y $\boldsymbol{b}_j$ están relacionados. En efecto multiplicando las igualdades anteriores por X y X' respectivamente, obtenemos:

$$X\,X'(X\boldsymbol{a}_j) = \lambda_j\,(X\boldsymbol{a}_j) \tag{7.3}$$
$$X'\,X\,(X'\boldsymbol{b}_j) = \nu_j\,(X'\boldsymbol{b}_j)\,. \tag{7.4}$$

Ello muestra que $X\boldsymbol{a}_j$ es vector propio de $X\,X'$ y $X'\boldsymbol{b}_j$ es vector propio de $X'\,X$.

Es además fácil ver que los valores propios no nulos son idénticos. Supongamos que λ_1 es el mayor valor propio de $X'\,X$ y ν_1 el mayor valor propio de $X\,X'$. Como $X\boldsymbol{a}_1$ es vector propio de $X\,X'$ con valor propio asociado λ_1, se sigue que $\nu_1 = \text{máx}_j\,\nu_j \geq \lambda_1$. Análogamente, si $\boldsymbol{b}_1$ es el vector propio de $X\,X'$ asociado al mayor valor propio ν_1, entonces $X'\boldsymbol{b}_1$ es vector propio de $X'\,X$ con valor propio asociado ν_1, y por tanto $\nu_1 \leq \lambda_1$. De ambas desigualdades se deduce $\nu_1 = \lambda_1$, y el argumento puede reiterarse para los valores propios sucesivos.

En definitiva,

$$\boldsymbol{a}_j \propto X'\boldsymbol{b}_j \tag{7.5}$$
$$\boldsymbol{b}_j \propto X\boldsymbol{a}_j, \tag{7.6}$$

par $j = 1, \ldots, p$. Además, las relaciones de proporcionalidad anteriores pueden convertirse en igualdades si tenemos en cuenta que

$$\|X'\boldsymbol{b}_j\|^2 = \boldsymbol{b}_j{}'X\,X'\boldsymbol{b}_j = \nu_j \tag{7.7}$$
$$\|X\boldsymbol{a}_j\|^2 = \boldsymbol{a}_j{}'X'\,X\boldsymbol{a}_j = \lambda_j, \tag{7.8}$$

lo que permite normalizar los lados derechos de las expresiones (7.5)–(7.6) y convertirlas en igualdades:

$$\boldsymbol{a}_j = \lambda_j^{-\frac{1}{2}} X'\boldsymbol{b}_j \tag{7.9}$$
$$\boldsymbol{b}_j = \lambda_j^{-\frac{1}{2}} X\boldsymbol{a}_j. \tag{7.10}$$

Estas expresiones para $j = 1, \ldots, p$ se resumen en las igualdades matriciales

$$A = X'B\Lambda^{-\frac{1}{2}} \tag{7.11}$$
$$B = XA\Lambda^{-\frac{1}{2}}. \tag{7.12}$$

Si proyectamos las filas y columnas de X sobre los subespacios engendrados por el vector propio $\boldsymbol{a}_j$ y $\boldsymbol{b}_j$ respectivamente, tenemos:

$$\boldsymbol{u}_j = X\boldsymbol{a}_j = \lambda_j^{-\frac{1}{2}} X\, X'\boldsymbol{b}_j = \lambda_j^{\frac{1}{2}}\boldsymbol{b}_j \tag{7.13}$$

$$\boldsymbol{v}_j = X'\boldsymbol{b}_j = \lambda_j^{-\frac{1}{2}} X'\, X\boldsymbol{a}_j = \lambda_j^{\frac{1}{2}}\boldsymbol{a}_j. \tag{7.14}$$

Si tomamos la igualdad (7.9), premultiplicamos por X, postmultiplicamos por $\boldsymbol{a}_j{}'$ y sumamos respecto j, obtenemos:

$$X\left(\sum_{j=1}^{p} \boldsymbol{a}_j\boldsymbol{a}_j{}'\right) = \sum_{j=1}^{p} \lambda_j^{\frac{1}{2}}\boldsymbol{b}_j\boldsymbol{a}_j{}' = B\Lambda^{\frac{1}{2}}A'. \tag{7.15}$$

Como $\sum_{j=1}^{p} \boldsymbol{a}_j\boldsymbol{a}_j{}' = AA' = I$, la igualdad anterior se reduce a:

$$X = \sum_{j=1}^{p} \sqrt{\lambda_j}\boldsymbol{b}_j\boldsymbol{a}_j{}' = B\Lambda^{\frac{1}{2}}A', \tag{7.16}$$

llamada *descomposición en valores singulares* de la matriz X.

7.2. Biplots

En el supuesto de que X sea aproximadamente igual a los $q < p$ primeros sumandos (7.16) obtenemos:

$$X \approx \sum_{j=1}^{q} \sqrt{\lambda_j}\boldsymbol{b}_j\boldsymbol{a}_j{}' = B_q S_q A_q{}'. \tag{7.17}$$

Podemos asociar S a la matriz A, a la matriz B o a ambas a la vez. Por ejemplo, podemos definir $G_q = B_q S^{1-c}$ y $H_q{}' = S^c A_q{}'$. Para cada valor $0 \leq c \leq 1$ que escojamos tenemos

$$X = G_q H_q{}' = B_q S^{1-c} S^c A_q{}' \tag{7.18}$$

El exponente c se puede escoger de diferentes maneras: elecciones habituales son $c = 0$, $c = \frac{1}{2}$ y $c = 1$.

Sea $\boldsymbol{g}_i{}'$ la i-ésima fila de G y $\boldsymbol{h}_j{}'$ la j-ésima fila de H (por tanto, j-ésima columna de H'). Si $q = 2$, los $N+p$ vectores $\boldsymbol{g}_i$ y $\boldsymbol{h}_j$ pueden representarse en el plano dando lugar a la representación conocida como *biplot*. Los puntos $\boldsymbol{g}_i$ representan observaciones, en tanto los puntos $\boldsymbol{h}_j$ representan variables.

7.2.1. Interpretación

Para interpretar un biplot, notemos que si (7.17) se verifica de modo aproximado, entonces

$$X_{ij} \approx \boldsymbol{g}_i{}'\boldsymbol{h}_j = ||\boldsymbol{g}_i||||\boldsymbol{h}_j|| \cos(\alpha_{ij}) \tag{7.19}$$

siendo α_{ij} el ángulo que forman $\boldsymbol{g}_i$ y $\boldsymbol{h}_j$. Por consiguiente, si la variable j tiene gran influencia en la observación i, los vectores representando a ambas tenderán a formar un ángulo pequeño.

Adicionalmente, dependiendo del valor seleccionado para c en (7.18) podemos interpretar las distancias euclídeas entre las representaciones de los puntos fila, de los puntos columna, etc.

Caso $c = 0$. Supongamos $X = GH'$ exactamente (omitimos el subíndice q por simplicidad notacional). Entonces, si tomamos $c = 0$, $H = A$ y es por tanto ortogonal, con lo que $XX' = GH'HG' = GG'$. Por consiguiente, para cualquier fila $\boldsymbol{x_i}$ de X se tiene

$$\begin{aligned} \boldsymbol{x_i}'\boldsymbol{x_i} &= \boldsymbol{g}_i'\boldsymbol{g}_i && (7.20)\\ ||\boldsymbol{x_i}|| &= ||\boldsymbol{g}_i|| && (7.21)\\ ||\boldsymbol{x_i} - \boldsymbol{x_j}|| &= ||\boldsymbol{g}_i - \boldsymbol{g_j}|| && (7.22)\\ \cos(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \cos(\boldsymbol{g_i}, \boldsymbol{g_j}); && (7.23) \end{aligned}$$

es decir, las distancias y ángulos entre los vectores $\boldsymbol{g}_i$ reproducen los existentes entre los vectores $\boldsymbol{x_i}$. Obviamente, esto sólo es posible si la configuración original de puntos fila de X era bidimensional; de otro modo, $X \approx GH'$ y lo anterior sólo tendrá validez como aproximación.

Caso $c = 1$. Razonando de forma exactamente análoga, llegamos a la conclusión de que en este caso las distancias y ángulos entre los vectores fila de H' reproducen los existentes entre los vectores columna de X, dado que con $c = 1$

$$X'X = HG'GH' = HH' \tag{7.24}$$

al ser $G = B$ una matriz ortogonal. (De nuevo la igualdad anterior es sólo aproximada, en la medida en que la matriz original X no sea de rango igual o inferior a 2).

Caso $c = \frac{1}{2}$. Esta elección de c supone un compromiso entre las dos anteriores, tendente a preservar en alguna medida las distancias tanto entre puntos fila como entre puntos columna.

7.2.2. Ejemplo

Consideremos la Tabla 7.1, cuya casilla ij-ésima recoge el total de hogares de la Comunidad Autónoma i-ésima disponiendo del equipamiento a que se refiere la columna j-ésima.

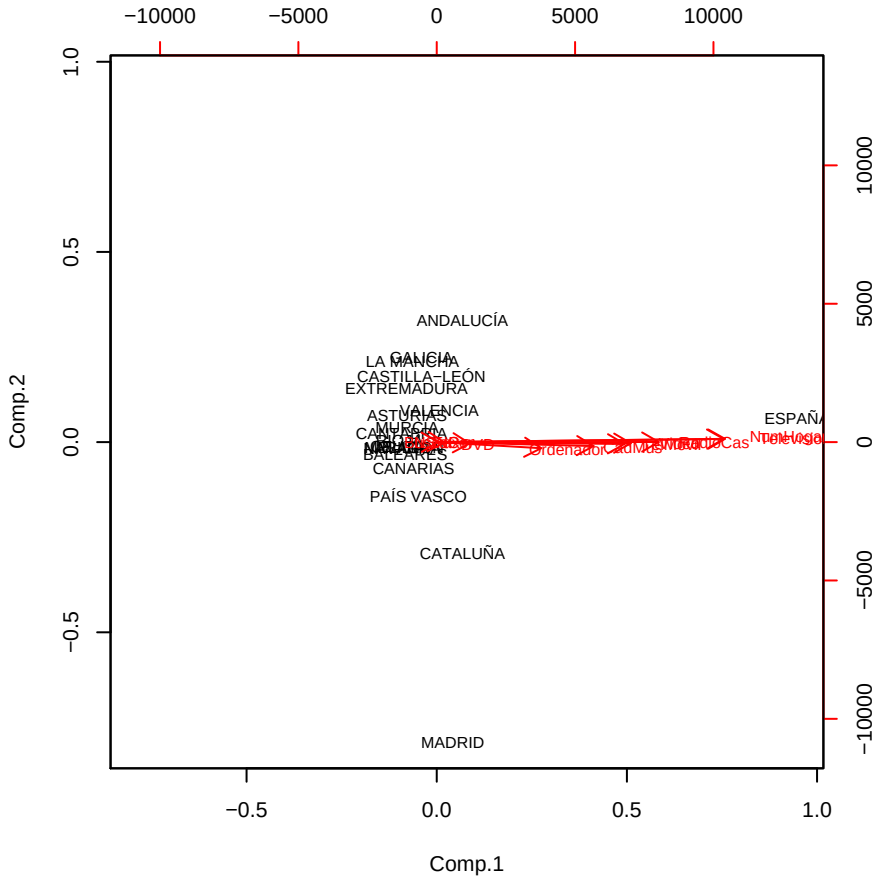
Un análisis de los datos brutos daría lugar a un biplot como el recogido en la Figura 7.1. Es aparente un "efecto tamaño" muy pronunciado: al estar los datos en valores absolutos, todas las columnas son aproximadamente

Cuadro 7.1: Dotación de los hogares por Comunidades Autónomas (miles de hogares que poseen cada uno de los equipamientos indicados). Fuente: INE, *Encuesta de Tecnologías de la información en los hogares, 2002.*

	Número Hogares	Televisión	Ordenador	Fax	Video	DVD	Cadena Música	Radio, cassete	Busca personas	Teléfono móvil	NSNC NSNC
ESPAÑA	13712.90	13650.60	4944.10	371.60	9207.80	1562.30	7451.60	10570.70	75.10	8917.70	5.00
ANDALUCÍA	2306.90	2301.00	717.70	51.30	1553.60	246.60	1151.30	16 49.00	12.60	1482.90	1.20
ARAGÓN	426.30	423.30	158.30	8.40	285.10	45.30	241.30	361.60	2. 40	252.70	0.00
ASTURIAS	364.90	363.70	115.90	7.70	217.70	31.10	173.80	311.80	1.90	221.00	0.00
BALEARES	293.50	290.80	110.50	15.10	200.80	46.50	166.90	212.30	1.50	194.80	0.00
CANARIAS	570.90	569.60	207.20	17.40	403.40	82.70	346.90	410.80	2.90	391.10	0.00
CANTABRIA	170.90	170.50	50.60	5.90	108.20	18.10	87.00	131.60	2 .00	108.20	0.00
CASTILLA-LEÓN	871.10	865.40	263.70	16.90	530.10	72.90	436.70	708 .90	3.20	511.60	0.50
LA MANCHA	580.10	576.50	149.70	11.90	354.10	42.10	249.60	413.40	0.00	326.30	0.00
CATALUÑA	2217.40	2208.60	933.50	75.90	1561.50	277.10	1235.90	174 0.60	17.40	1442.40	1.40
VALENCIA	1461.50	1457.40	473.70	35.40	1021.60	169.20	782.60	1095 .60	5.30	962.30	0.00
EXTREMADURA	358.50	355.00	84.60	3.30	213.50	24.10	155.50	268.60	2.30	204.90	0.00
GALICIA	887.10	878.50	254.90	17.20	485.50	82.80	428.30	670.70	10.50	536.60	2.00
MADRID	1809.30	1802.20	902.80	65.60	1321.50	265.70	1190.40	1452. 20	8.70	1347.70	0.00
MURCIA	362.00	359.00	105.20	7.10	247.30	43.10	188.30	272.30	1. 20	243.80	0.00
NAVARRA	185.20	183.40	72.80	6.00	124.80	13.50	100.90	148.90	0. 50	123.80	0.00
PAÍS VASCO	713.70	712.40	295.50	24.40	485.60	85.70	440.80	615.60	2.00	486.70	0.00
RIOJA	94.80	94.60	31.80	0.60	62.90	9.80	51.10	76.60	0.00	51. 70	0.00
CEUTA	20.50	20.30	7.30	0.70	15.90	2.50	12.90	15.00	0.20	14.9 0	0.00
MELILLA	18.50	18.50	8.60	0.80	14.70	3.40	11.40	15.10	0.40	14 .20	0.00

proporcionales, lo que hace los datos muy "uno-dimensionales": las Comunidades más pobladas, tienen más hogares en posesión de cada uno de los bienes considerados, simplemente por efecto de su tamaño. Puede verse en la figura indicada como "España" aparece en el margen derecho, y el resto de Comunidades ordenadas en el eje de abscisas aproximadamente por su tamaño.

Figura 7.1: Biplot de número de hogares (en valor absoluto) en cada Comunidad Autónoma que poseen diferentes tipos de equipamiento relacionado con la sociedad de la información. Se aprecia el fuerte efecto "tamaño" que oblitera cualquier otro.

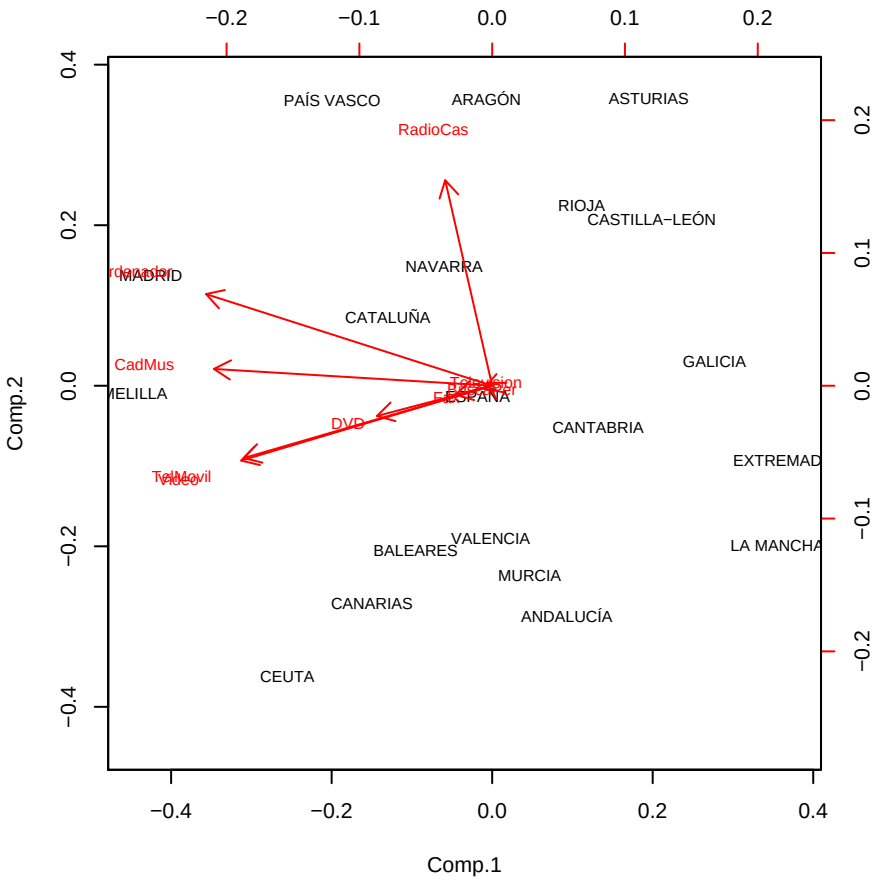


Podemos convertir los datos a porcentajes, evitando así que una dimensión de la representación gráfica sea ocupada por el efecto tamaño, que carece de interés. Así se ha hecho para producir la Figura 7.2, mucho más ilustrativa que la primera. Se aprecia ahora como los puntos que representan variables están todos orientados de manera similar, como corresponde dada su apre-

ciable correlación. Casi superpuesta al punto que representa "Ordenadores" está la Comunidad de Madrid, y bastante a la izquierda también Comunidades como País Vasco y Cataluña, en que los equipamientos considerados han alcanzado una penetración relativamente elevada en los hogares. En el lado derecho del biplot aparecen aquellas comunidades en que dicha penetración es, en términos relativos, menor: Extremadura, Andalucía, Galicia, Castilla-La Mancha.

Algunos otros detalles saltan a la vista en la Figura 7.2; por ejemplo, la ordenada relativamente alta de País Vasco, Aragón y Asturias, que se corresponde con una tenencia también relativamente elevada de radiocassettes, como puede corroborarse observando la tabla.

Figura 7.2: Biplot del porcentaje de hogares en cada Comunidad Autónoma que poseen diferentes tipos de equipamiento relacionado con la sociedad de la información. Al desaparecer el efecto tamaño por ser todas las magnitudes relativas, se aprecian las diferencias entre comunidades en la dotación relativa.



7.3. Lectura recomendada

El biplot e instrumentos de visualización relacionados se describen en **?**, Cap. 4.

Capítulo 8

Datos categóricos multivariantes

8.1. Introducción

En lo que precede, hemos considerado como punto de partida una matriz de datos X de dimensiones $N \times p$ cada una de cuyas filas $\boldsymbol{x_i}'$ era un vector aleatorio en R^p.

En ocasiones, sin embargo, para cada sujeto de la muestra examinamos k *atributos* cualitativos o *caracteres*, cada uno de los cuales con d_i *niveles* $i = 1, \ldots, k$. Por ejemplo, si registráramos el color de pelo y ojos de un colectivo de $N = 5$ sujetos, podríamos presentar la información resultante en una tabla como:

Cuadro 8.1: Color de pelo y ojos medidos para cinco sujetos

Sujeto	**Color pelo**	**Color ojos**
1	Negro	Castaño
2	Rubio	Azul
3	Negro	Azul
4	Negro	Castaño
5	Negro	Castaño

Una forma alternativa de recoger la misma información es efectuando una tabulación cruzada de los dos caracteres (color de pelo y color de ojos) para producir una *tabla de contingencia*como la recogida en el Cuadro 8.2.

De tener una tabla de datos $N \times p$ listando los respectivos niveles de los caracteres para cada uno de los N sujetos, pasamos a tener una tabla de k dimensiones y $\prod_{i=1}^{k} d_i$ celdas relacionando los caracteres entre sí.

Cuadro 8.2: Tabla de contingencia relacionando color de pelo y ojos para cinco sujetos

	Color de pelo	
	Negro	**Rubio**
Ojos azules	1	1
Ojos castaños	3	0

Es fácil ver que la tabla de datos original en el Cuadro 8.1 y la tabla de contingencia en el Cuadro 8.2 proporcionan exactamente la misma información. De la segunda podemos reconstruir la primera (excepto por el orden, normalmente irrelevante).

El análisis de tablas de doble entrada es objeto común de los cursos introductorios de Estadística. Problemas habituales que se resuelven son los de contrastar la independencia de los caracteres, o la homogeneidad de subpoblaciones descritas por las filas o columnas, habitualmente mediante el contraste χ^2 de Pearson (véase por ej. **?**, p. 244–249).

No estamos limitados a considerar tablas de doble entrada, sino que en general trabajaremos con tablas de contingencia con $k > 2$ dimensiones. Cuando lo hagamos, será en general inconveniente examinar los caracteres por parejas: si lo hiciéramos, podriamos tropezar con la *paradoja de Simpson* que ilustramos a continuación.

Notación. Consideremos, por concreción, una tabla de contingencia con $k = 3$ dimensiones (generalizar a cualquier k, no obstante, será inmediato). Denotaremos por A, B y C respectivamente a los tres caracteres, con d_A, d_B y d_C niveles respectivamente.

Sea X la tabla de contingencia, y x_{ijk} el contenido de su celda ijk. Es decir, x_{ijk} sujetos poseen los niveles i, j y k de los tres caracteres considerados y $N = \sum_{i,j,k} x_{ijk}$ el total de sujetos en todas las celdas de la tabla.

8.2. Tipos de muestreo

Una misma tabla de contingencia puede generarse de diferentes modos, y es importante saber cuál ha sido el empleado en cada caso.

Podríamos muestrear durante un periodo de tiempo y clasificar a los sujetos de acuerdo a, por ejemplo, tres caracteres, de modo que cada uno fuera contado en una celda x_{ijk} de una tabla tridimensional. Si hacemos esto, podemos modelizar x_{ijk} como una variable con distribución de Poisson

de parámetro λ_{ijk}. El número total de sujetos tabulados, N, será a su vez una variable aleatoria con distribución de Poisson. Diremos que la tabla se ha generado mediante *muestreo de Poisson*

Alternativamente, podríamos fijar el tamaño muestral N deseado y tabular dichos N sujetos. Entonces, podríamos ver el vector $\boldsymbol{x_{ijk}}$ como variable aleatoria con distribución multinomial,

$$\text{Prob}(\boldsymbol{x_{ijk}}) = \frac{N!}{x_{iii}! \dots x_{ijk}! \dots x_{IJK}!} \cdot p_{111}^{x_{111}} \dots p_{ijk}^{x_{ijk}} \dots p_{IJK}^{x_{IJK}} \tag{8.1}$$

en que I, J, K designan el número de niveles de (respectivamente) los caracteres A, B y C. Decimos en este caso hallarnos ante *muestreo multinomial*

Frecuentemente se toman muestras estratificadas, fijando cuotas para diferentes estratos de la población analizada. Por ejemplo, si examináramos la respuesta a un tratamiento que sólo raramente se administra, porque se emplea para enfermedades infrecuentes, una muestra aleatoria simple proporcionaría muy pocos sujetos tratados: acaso ninguno.

El modo habitual de operar en este caso es tomar una muestra de sujetos tratados y otra de no tratados o controles, de modo que ambas categorías estén adecuadamente representadas. Cada uno de los segmentos de la población, el de los tratados y no tratados, se muestrea así por separado: la muestra obtenida puede verse como la unión de dos muestras para dos subpoblaciones. En este caso, no sólo hemos fijado N, sino también el desglose $N = N_t + N_c$ entre tratados y no tratados o controles. Decimos entonces hallarnos ante *muestreo producto-multinomial* Es importante darse cuenta de que en tales casos las proporciones marginales de la tabla no estiman proporciones en la población: son un mero resultado del diseño muestral. Por ejemplo, N_t/N no estimaría la proporción de sujetos tratados en la población, porque tanto numerador como denominador han sido arbitrariamente fijados.

En situaciones más complejas que la muy simple descrita, podríamos tener, por ejemplo, cuotas por sexo y grupo de edad, y en consecuencia estaríamos fijando el número N_{ij} de sujetos muestreados para cada combinación de sexo y edad.

8.3. La paradoja de Simpson

Consideremos la siguiente tabla de contingencia, relacionando recepción de un tratamiento o un placebo con el hecho de contraer o no una cierta enfermedad. En cursivas, bajo los valores absolutos, aparece entre paréntesis la proporción sobre el total de la fila correspondiente.

	Enferman	**No enferman**	**Total**
Tratamiento	5950 *(0.398)*	9005 *(0.602)*	14955
Placebo	5050 *(0.822)*	1095 *(0.178)*	6145

A la vista de los datos anteriores, estaríamos tentados de concluir que el tratamiento ha tenido realmente un efecto preventivo: menos del 40 % de tratados desarrollan la enfermedad, frente a más del 80 % de quienes tomaron el placebo.

Supongamos, sin embargo, que efectuamos un desglose por en varones y mujeres de la tabla anterior para obtener las dos siguientes:

	Varones		
	Enferman	**No enferman**	**Total**
Tratamiento	5000 *(0.999)*	5 *(0.001)*	5005
Placebo	5000 *(0.981)*	95 *(0.019)*	5095

	Mujeres		
	Enferman	**No enferman**	**Total**
Tratamiento	950 *(0.095)*	9000 *(0.905)*	9950
Placebo	50 *(0.005)*	1000 *(0.995)*	1050

Se da ahora una aparente paradoja: mientras para el total de la población el tratamiento aparentaba ser efectivo, tanto los varones como las mujeres tratados parecen haber enfermado más que los que recibieron el placebo. Esto ocurre por poco margen en el caso de los varones, pero de forma notoria en las mujeres. Resulta así que la tabla para el total de la población

proporciona una información que es contradictoria con la que obtenemos al considerar las tablas desglosadas.

La contradicción entre los resultados que sugieren la tabla conjunta y las dos que forman el desglose se explica cuando notamos que la asignación del tratamiento ha sido muy asimétrica entre hombres y mujeres: las mujeres, que parecen practicamente inmunes a la enfermedad analizada, han recibido mayoritariamente el tratamiento, mientras que los hombres, mucho más vulnerables, no lo han recibido en la misma proporción. Se tiene así una menor incidencia de la enfermedad (en la tabla conjunta) para los receptores del tratamiento, simplemente porque entre ellos hay mayoría de mujeres casi inmunes. Cuando se analizan separadamente las tablas correspondientes a hombres y mujeres apreciamos, sin embargo, que el tratamiento no parece tener ningún efecto positivo.

Si tabuláramos los tres caracteres a la vez, tendríamos una tabla de tres dimensiones (Tratamiento × Enfermedad × Sexo). Sumando sobre la tercera dimensión llegaríamos a la tabla de dos dimensiones (Tratamiento × Enfermedad). Decimos que ésta última resulta de colapsar la primera o que es uno de sus márgenes. Lo que la paradoja de Simpson presentada más arriba muestra es que colapsando una tabla puede llegarse a conclusiones diferentes —incluso radicalmente opuestas— a las que alcanzaríamos al considerar la tabla completa. Nos deberemos por ello abstener de colapsar una tabla si la asociación entre los caracteres correspondientes a las dimensiones que subsisten es diferente para diferentes niveles del carácter o caracteres correspondientes a las dimensiones suprimidas.

> **Observación 8.1** Este efecto es similar al que se presenta al comparar el coeficiente de correlación simple entre dos variables y el coeficiente de correlación parcial controlando el efecto de una tercera. Ambos pueden tener valores completamente diferentes, e incluso signo opuesto, como el Ejemplo 1.2 ponía de manifiesto.

8.4. Modelos logarítmico-lineales

Consideraremos una tabla de tres dimensones, pero de nuevo el planteamiento es fácilmente generalizable.

Denotemos por p_{ijk} la probabilidad de que un sujeto tomado al azar entre los N que componen la tabla esté en la celda (ijk). Denotemos por

$$p_{i++} = \sum_{j=1}^{d_B} \sum_{k=1}^{d_C} p_{ijk} \qquad p_{+j+} = \sum_{i=1}^{d_A} \sum_{k=1}^{d_C} p_{ijk} \qquad p_{++k} = \sum_{i=1}^{d_A} \sum_{j=1}^{d_B} p_{ijk}$$

las probabilidades marginales e imaginemos que hubiera independencia entre los tres caracteres A, B, C examinados. Entonces, tendríamos:

$$p_{ijk} = p_{i++} p_{+j+} p_{++k} \tag{8.2}$$

o, en escala logarítmica,

$$\log(p_{ijk}) = \log(p_{i++}) + \log(p_{+j+}) + \log(p_{++k}); \tag{8.3}$$

en el caso de independencia, $\log(p_{ijk})$ se puede expresar como suma de efectos fila, columna y estrato. Cada nivel de cada caracter contribuye una cantidad fija a $\log(p_{ijk})$, que no depende de cuál sea el nivel observado de ningún otro carácter.

Podríamos considerar modelos más generales para $\log(p_{ijk})$ como suma de diferentes efectos aditivos así:

$$\log(p_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}; \tag{8.4}$$

al objeto de identificar todos los parámetros (y de hacerlos interpretables) necesitamos restricciones del tipo:

$$\sum_i u_i^A = \sum_j u_j^B = \sum_k u_k^C = 0 \tag{8.5}$$

$$\sum_j u_{ij}^{AB} = \sum_i u_{ij}^{AB} = 0 \tag{8.6}$$

$$\sum_i u_{ik}^{AC} = \sum_k u_{ik}^{AC} = 0 \tag{8.7}$$

$$\sum_j u_{jk}^{BC} = \sum_k u_{jk}^{BC} = 0 \tag{8.8}$$

$$\sum_i u_{ijk}^{ABC} = \sum_j u_{ijk}^{ABC} = \sum_k u_{ijk}^{ABC} = 0. \tag{8.9}$$

El modelo (8.4) está saturado:utiliza tantos parámetros libres como celdas. Podemos considerar variedades del mismo, como:

$$\log(p_{ijk}) = u + u_i^A + u_j^B + u_k^C \tag{8.10}$$

$$\log(p_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} \tag{8.11}$$

$$\log(p_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ik}^{AC} \tag{8.12}$$

$$\log(p_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ik}^{AC} + u_{jk}^{BC} \tag{8.13}$$

$$\log(p_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC}. \tag{8.14}$$

El modelo (8.10) corresponde a la independencia entre los tres caracteres, A, B y C. El modelo (8.11) incorpora una interacción entre los caracteres A, B: el efecto de cada nivel i de A no es idéntico para cualquier nivel j de B, sino que combinaciones ij particulares tienen efecto sobre $\log(p_{ijk})$ que difiere de la suma $u_i^A + u_j^B$; analogamente con (8.12) y (8.13). El último de los modelos contiene todas las interacciones de segundo orden y es el más parametrizado antes de llegar al saturado, (8.4).

Los parámetros de un modelo logarítmico-lineal son funciones de $\log(p_{ijk})$; por ejemplo, sumando (8.10) respecto de i, j, k y teniendo en cuenta las restricciones de suma cero, tenemos:

$$u = \frac{1}{d_A d_B d_C} \sum_{i=1}^{d_A} \sum_{j=1}^{d_B} \sum_{k=1}^{d_C} \log(p_{ijk}); \tag{8.15}$$

Si ahora sumamos la misma igualdad sobre j, k llegamos a

$$u_i^A = \frac{1}{d_B d_C} \left(d_B d_C u + \sum_{j=1}^{d_B} \sum_{k=1}^{d_C} \log(p_{ijk}) \right), \tag{8.16}$$

y análogamente para los parámetros u_j^B y u_k^C. Nótese que los resultados son los mismos cuando consideramos cualquiera de los modelos más parametrizados (8.11)–(8.13). Sustituyendo (8.15) en (8.16) llegamos a: Si ahora sumamos la misma igualdad sobre j, k llegamos a

$$u_i^A = \frac{1}{d_B d_C} \sum_{j=1}^{d_B} \sum_{k=1}^{d_C} \log(p_{ijk}) - \frac{1}{d_a d_B d_C} \sum_{i=1}^{d_A} \sum_{j=1}^{d_B} \sum_{k=1}^{d_C} \log(p_{ijk}), \tag{8.17}$$

y análogamente para los términos restantes. Los estimadores máximo verosímiles de los parámetros se pueden obtener así de los de los términos p_{ijk}, y éstos son simplemente $\hat{p}_{ijk} = x_{ijk}/N$.

En la práctica, el *algoritmo de reescalado iterativo* permite la estimación cómoda de cualquier modelo logarítmico lineal.

8.5. Lectura recomendada

Son buenas introducciones **?**, **?**, **?** y **?**.

Capítulo 9

Análisis de Correspondencias

Es una técnica para producir representaciones planas relacionando las observaciones (filas) y variables (columnas) en una tabla de contingencia, es decir, una tabla cada una de cuyas casillas recoge números naturales. Es el caso de la Tabla 7.1, aunque por comodidad el número de hogares se haya expresado en miles.

9.1. Análisis de las filas de X

9.1.1. Notación

El punto de partida será una matriz de datos X de dimensiones $N \times p$ que, como se ha indicado, es una tabla de contingencia. Sea $T = \sum_{i=1}^{N} \sum_{j=1}^{p} x_{ij}$. Emplearemos la siguiente notación:

9.1.2. Distancia entre las filas de la matriz de datos

Si quisiéramos obtener una representación en pocas dimensiones de *las filas* de la matriz X, parecería lo indicado un análisis en componentes principales como el descrito en el Capítulo 5. La condición de tabla de contingencia de los datos de partida sugiere no obstante algunas alteraciones.

Consideremos la matriz F y, dentro de ella, dos filas i, j como las siguientes:

i	0.015	0.02	0.01	0.01	0.02	$f_{i.} = 0.0750$
j	0.0015	0.002	0.001	0.001	0.002	$f_{j.} = 0.0075$

Cuadro 9.1: Notación empleada

Símbolo	**Elemento genérico**	**Descripción**
X	x_{ij}	Tabla de contingencia original $N \times p$.
F	$f_{ij} = T^{-1}x_{ij}$	Matriz de frecuencias relativas $N \times p$.
$f_{i.}$	$f_{i.} = \sum_{j=1}^{p} f_{ij}$	Total marginal fila i-ésima de F.
$f_{.j}$	$f_{.j} = \sum_{i=1}^{N} f_{ij}$	Total marginal columna j-ésima de F.
$\boldsymbol{c}$		$\boldsymbol{c}' = (f_{.1} \dots f_{.p})$, totales marginales columnas.
$\boldsymbol{f}$		$\boldsymbol{f}' = (f_{1.} \dots f_{N.})$, totales marginales filas.
D_f		Matriz diagonal $N \times N$ con $f_{1.} \dots f_{N.}$ en la diagonal principal.
D_c		Matriz diagonal $p \times p$ con $f_{.1} \dots f_{.p}$ en la diagonal principal.

Es aparente que la fila i está mucho más poblada que la fila j (un 7.5 % de los casos totales frente a sólo un 0.75 %). Si prescindimos de este efecto debido al tamaño, vemos no obstante que las frecuencias relativas intrafila de las cinco categorias consideradas en las columnas son idénticas en ambas filas. Por ejemplo, la primera categoría se presenta en i con una frecuencia intrafila de 0.015 / 0.075 = 20 % y de exactamente el mismo valor en la fila j; y así para todas las demás.

En consecuencia, si aspiramos a hacer una análisis que describa las diferencias *relativas* entre las filas, parece que deberíamos corregir el efecto tamaño aludido, lo que se logra sustituyendo cada f_{ij} por $f_{ij}/f_{i.}$, que es lo mismo que reemplazar en nuestro análisis la matriz F por ${D_f}^{-1}F$.

Podríamos pensar que tras hacer esta corrección sólo resta realizar un análisis en componentes principales convencional, pero hay otra peculiaridad a la que debemos enfrentarnos. Imaginemos tres filas de ${D_f}^{-1}F$ tales como las siguientes:

k	0.15	0.02	0.10	0.43	0.30
l	0.15	0.02	0.10	0.44	0.29
m	0.15	0.01	0.10	0.44	0.30

Observemos que, si computamos la distancia euclídea ordinaria $d(k,l)$ entre las filas k,l por un lado y $d(k,m)$ por otro, obtenemos:

$$\begin{aligned} d_e^2(k,l) &= \sum_{j=1}^{p} \left(\frac{f_{kj}}{f_{k.}} - \frac{f_{lj}}{f_{l.}} \right)^2 && (9.1) \\ &= (0{,}43 - 0{,}44)^2 + (0{,}30 - 0{,}29)^2 = 0{,}0002 && (9.2) \\ d_e^2(k,m) &= \sum_{j=1}^{p} \left(\frac{f_{kj}}{f_{k.}} - \frac{f_{mj}}{f_{m.}} \right)^2 && (9.3) \\ &= (0{,}43 - 0{,}44)^2 + (0{,}02 - 0{,}01)^2 = 0{,}0002 && (9.4) \end{aligned}$$

Esto es claramente indeseable en general: no es lo mismo una discrepancia de 0.01 entre 0.29 y 0.30 que entre 0.01 y 0.02. En este último caso, un carácter raro en ambas filas lo es mucho más en una (la m) que en otra (la k), y tenderíamos a atribuir a este hecho mucha mayor significación. Por ejemplo, si las cifras anteriores reflejaran la prevalencia de determinadas enfermedades en distintas comunidades, 0.43 y 0.44 podrían recoger el tanto por uno de personas que han padecido un resfriado común en las comunidades k y m: difícilmente consideraríamos la discrepancia como relevante. En cambio, la segunda columna podría reflejar el tanto por uno de personas atacadas por una enfermedad muy infrecuente, y el hecho de que en la comunidad l este tanto por uno es doble que en la k no dejaría de atraer nuestra atención.

En consecuencia, hay razón para ponderar diferentemente las discrepancias en los diferentes caracteres, y una forma intuitivamente atrayente de hacerlo es sustituir la distancia euclidea ordinaria por:

$$\begin{aligned} d^2(k,l) &= \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{kj}}{f_{k.}} - \frac{f_{lj}}{f_{l.}} \right)^2 && (9.5) \\ &= \sum_{j=1}^{p} \left(\frac{f_{kj}}{f_{k.}\sqrt{f_{.j}}} - \frac{f_{lj}}{f_{l.}\sqrt{f_{.j}}} \right)^2 && (9.6) \end{aligned}$$

Por su semejanza formal con el estadístico χ^2 se denomina a la distancia anterior distancia χ^2.

Observemos, que si sustituimos la matriz ${D_f}^{-1}F$ por $Y = {D_f}^{-1}F{D_c}^{-\frac{1}{2}}$, cuya i-ésima fila es de la forma

$$\left(\frac{f_{i1}}{f_{i.}\sqrt{f_{.1}}}, \frac{f_{i2}}{f_{i.}\sqrt{f_{.2}}}, \ldots, \frac{f_{ip}}{f_{i.}\sqrt{f_{.p}}} \right),$$

un análisis sobre ${D_f}^{-1}F{D_c}^{-\frac{1}{2}}$ haciendo uso de distancias euclídeas equivale al análisis sobre ${D_f}^{-1}F$ haciendo uso de distancias χ^2.

9.1.3. Matriz de covarianzas muestral

El último paso previo al análisis en componentes principales, una vez que hemos decidido hacerlo sobre ${D_f}^{-1}F{D_c}^{-\frac{1}{2}}$, es la estimación de la matriz de covarianzas. El estimador ordinario (y máximo verosímil, en el caso de muestras procedentes de observaciones normales) es:

$$\begin{aligned}\hat{\Sigma} &= N^{-1}\sum_{i=1}^{N}(\boldsymbol{y_i}-\overline{\boldsymbol{y}})(\boldsymbol{y_i}-\overline{\boldsymbol{y}})' && (9.7)\\ &= N^{-1}\sum_{i=1}^{N}\boldsymbol{y_i}\boldsymbol{y_i}'-\overline{\boldsymbol{y}}\overline{\boldsymbol{y}}' && (9.8)\\ &= N^{-1}Y'Y-(N^{-1}Y'\mathbf{1_N})(N^{-1}\mathbf{1_N}'Y); && (9.9)\end{aligned}$$

ello supone dar a cada observación un peso de $1/N$, lo que es razonable en el caso de muestrear de forma aletoria simple una población.

En el caso que nos ocupa, se presenta de nuevo la peculiariedad de que unas observaciones —filas de la matriz X, que tras sucesivas transformaciones se ha convertido en $Y = {D_f}^{-1}F{D_c}^{-\frac{1}{2}}$— son en general más importantes que otras: sus totales $f_{i.}$ marginales difieren. Por ello, es razonable reemplazar el estimador anterior por:

$$\hat{\Sigma} = Y'D_fY-(Y'D_f\mathbf{1_N})(\mathbf{1_N}'D_fY). \qquad (9.10)$$

que supone dar peso $f_{i.}$ en lugar de $1/N$ a la fila i-ésima de Y.

Con las anteriores modificaciones estamos ya en situación de hacer un análisis en componentes principales. Notemos, en primer lugar, que $\boldsymbol{c^{\frac{1}{2}}}$ es vector propio de $\hat{\Sigma}$ asociado a un valor propio nulo. En efecto, como $Y'D_f\mathbf{1_N} = {D_c}^{-\frac{1}{2}}F'{D_f}^{-1}D_f\mathbf{1_N} = \boldsymbol{c^{\frac{1}{2}}}$, tenemos que

$$\begin{aligned}\hat{\Sigma}\boldsymbol{c^{\frac{1}{2}}} &= \left(Y'D_fY-\boldsymbol{c^{\frac{1}{2}}}\boldsymbol{c^{\frac{1}{2}}}'\right)\boldsymbol{c^{\frac{1}{2}}}\\ &= Y'D_fY\boldsymbol{c^{\frac{1}{2}}}-\boldsymbol{c^{\frac{1}{2}}}\\ &= {D_c}^{-\frac{1}{2}}F'{D_f}^{-1}D_f{D_f}^{-1}F{D_c}^{-\frac{1}{2}}\boldsymbol{c^{\frac{1}{2}}}-\boldsymbol{c^{\frac{1}{2}}}\\ &= {D_c}^{-\frac{1}{2}}F'{D_f}^{-1}F\mathbf{1_p}-\boldsymbol{c^{\frac{1}{2}}}\\ &= {D_c}^{-\frac{1}{2}}F'{D_f}^{-1}\boldsymbol{f}-\boldsymbol{c^{\frac{1}{2}}}\\ &= {D_c}^{-\frac{1}{2}}\boldsymbol{c}-\boldsymbol{c^{\frac{1}{2}}}\\ &= \mathbf{0}.\end{aligned}$$

Por tanto, podemos prescindir de una componente principal que no explica ninguna varianza, y utilizar sólo las restantes (ordinariamente, las dos primeras). Además, como los restantes vectores propios $\boldsymbol{a_i}$ ($i = 1,\ldots,p-1$) de $\hat{\Sigma}$ son ortogonales a $\boldsymbol{c^{\frac{1}{2}}}$, tenemos que

$$\hat{\Sigma}\boldsymbol{a_i} = \left(Y'D_fY-\boldsymbol{c^{\frac{1}{2}}}\boldsymbol{c^{\frac{1}{2}}}'\right)\boldsymbol{a_i} = Y'D_fY\boldsymbol{a_i};$$

en consecuencia, los vectores propios correspondientes a valores propios no nulos de $\hat{\Sigma}$ coinciden con los de $Y'D_fY$, y podemos diagonalizar esta última matriz.

Finalmente, observemos que $Y'D_fY = {D_c}^{-\frac{1}{2}}F'{D_f}^{-1}D_f{D_f}^{-1}F{D_c}^{-\frac{1}{2}} = {D_c}^{-\frac{1}{2}}F'{D_f}^{-\frac{1}{2}}{D_f}^{-\frac{1}{2}}F{D_c}^{-\frac{1}{2}}$ y denotando

$$Z = {D_f}^{-\frac{1}{2}}F{D_c}^{-\frac{1}{2}} \tag{9.11}$$

vemos que la matriz que diagonalizamos puede expresarse como $Z'Z$, hecho del que haremos uso en breve.

9.2. Análisis de las columnas de X

Podríamos ahora realizar un análisis en componentes principales de *las columnas* de la matriz X; es decir, buscamos una representación de baja dimensionalidad de los p vectores en R^N constituidos por las columnas de X.

Una discusión del todo paralela a la precedente, intercambiando los papeles de filas y columnas, nos llevaría a diagonalizar la matriz $\tilde{Y}D_c\tilde{Y}'$, en que $\tilde{Y} = {D_f}^{-\frac{1}{2}}F{D_c}^{-1}$. En consecuencia, $\tilde{Y}D_c\tilde{Y}' = {D_f}^{-\frac{1}{2}}F{D_c}^{-1}D_c{D_c}^{-1}F'{D_f}^{-\frac{1}{2}} = ZZ'$ con Z definida como anteriormente.

9.3. Reciprocidad y representación conjunta

Sean A y B las matrices que tienen por columnas los vectores propios de $Z'Z$ y ZZ' respectivamente. La representación de las filas de Y mediante todas las componentes principales viene entonces dada por

$$R = YA = {D_f}^{-1}F{D_c}^{-\frac{1}{2}}A, \tag{9.12}$$

en tanto la representación de las columnas de $\tilde{Y}$ viene dada por

$$C = \tilde{Y}'B = {D_c}^{-1}F'{D_f}^{-\frac{1}{2}}B. \tag{9.13}$$

Notemos sin embargo que las columnas de A y las de B están relacionadas, por ser vectores propios respectivamente de matrices que podemos escribir como $Z'Z$ y ZZ' respectivamente. Haciendo uso de (7.11) y (7.12) tenemos que:

$$R = YA = {D_f}^{-1}F{D_c}^{-\frac{1}{2}}Z'B\Lambda^{-\frac{1}{2}} \tag{9.14}$$

$$C = \tilde{Y}'B = {D_c}^{-1}F'{D_f}^{-\frac{1}{2}}ZA\Lambda^{-\frac{1}{2}}. \tag{9.15}$$

Tomemos la expresión (9.14). Haciendo uso de la definición de Z en (9.11) y de (9.13) tenemos que:

$$\begin{aligned} R &= {D_f}^{-1} F {D_c}^{-\frac{1}{2}} {D_c}^{-\frac{1}{2}} F' {D_f}^{-\frac{1}{2}} B \Lambda^{-\frac{1}{2}} && (9.16)\\ &= {D_f}^{-1} F \underbrace{{D_c}^{-1} F' {D_f}^{-\frac{1}{2}} B}_{C} \Lambda^{-\frac{1}{2}} && (9.17)\\ &= {D_f}^{-1} F C \Lambda^{-\frac{1}{2}} && (9.18) \end{aligned}$$

Análogamente,

$$\begin{aligned} C &= {D_c}^{-1} F' {D_f}^{-\frac{1}{2}} Z A \Lambda^{-\frac{1}{2}} && (9.19)\\ &= {D_c}^{-1} F' {D_f}^{-\frac{1}{2}} {D_f}^{-\frac{1}{2}} F {D_c}^{-\frac{1}{2}} A \Lambda^{-\frac{1}{2}} && (9.20)\\ &= {D_c}^{-1} F' R \Lambda^{-\frac{1}{2}} && (9.21) \end{aligned}$$

Las relaciones (9.18)-(9.21) se conocen como de *reciprocidad baricéntrica* y son las que permiten interpretar las posiciones relativas de filas y columnas. Consideremos, por ejemplo, la i-ésima fila $\boldsymbol{r_i}$ de R. De acuerdo con (9.18), su k-ésima coordenada puede expresarse así:

$$r_{ik} = \lambda_k^{-\frac{1}{2}} \left(\frac{f_{i1}}{f_{i.}} c_{1k} + \ldots + \frac{f_{ip}}{f_{i.}} c_{pk} \right),$$

es decir, como un promedio ponderado de la coordenada homóloga de las columnas, con pesos dados por

$$\frac{f_{i1}}{f_{i.}}, \ldots, \frac{f_{ip}}{f_{i.}};$$

si $f_{ij}/f_{i.}$ es muy grande, la variable j tiene gran relevancia en el perfil fila i, y el punto que representa a dicho perfil fila tendrá sus coordenadas "atraidas" hacia las de $\boldsymbol{c_j}$, las del punto que representa a la variable j. Análogamente para la representación de las columnas.

9.4. Lectura recomendada

Una introducción al Análisis de Correspondencias puede encontrarse tanto en **?** como en **?**; también será de utilidad, entre la bibliografía en español, **?**.

Capítulo 10

Análisis Procrustes

10.1. Introducción.

El análisis Procrustes tiene por objeto examinar en qué medida dos configuraciones de puntos en el espacio euclídeo son similares. Existen generalizaciones a más de dos configuraciones (ver por ej. **?**), pero aquí sólo trataremos el caso más simple. Seguimos en la exposición a **?**.

Consideremos dos configuraciones de N puntos en el espacio euclídeo R^k representadas por sendas matrices X e Y de dimensión $N \times k$. Las filas $\boldsymbol{y_i}$ y $\boldsymbol{x_i}$ de las matrices Y y X respectivamente proporcionan las coordenadas del punto i en las dos configuraciones.

Como medida de ajuste entre ambas tomaremos

$$G(X,Y) = \text{traza}((X-Y)(X-Y)') = \sum_{i=1}^{N} ||\boldsymbol{x_i} - \boldsymbol{y_i}||^2 \tag{10.1}$$

Para examinar si las dos configuraciones son similares, nos fijaremos en si conservan la posición relativa de los puntos excepto por transformaciones "simples" como traslaciones o cambios de escala. Específicamente buscaremos evaluar

$$G(X,Y) = \text{traza}((X-g(Y))(X-g(Y))'). \tag{10.2}$$

para una clase de transformaciones $g(.)$ incluyendo la composición de traslaciones, rotaciones y contracciones/expansiones. Por tanto,

$$g(Y) = \rho(Y - \mathbf{1}'\boldsymbol{a})P \tag{10.3}$$

siendo P una matriz ortogonal, $\boldsymbol{a}$ un vector de constantes y ρ un coeficiente de contracción o expansión de la escala. Llamaremos Γ al conjunto formado por todas las transformaciones $h(.)$ de la forma descrita en (10.3).

Estamos interesados en encontrar

$$G_{\text{mín}}(X, g(Y)) = \min_{\rho,P,\boldsymbol{a}} G(X, \rho(Y - \mathbf{1}'\boldsymbol{a})P) \tag{10.4}$$

y los correspondientes valores $\rho, P, \boldsymbol{a}$ para los que el mínimo se alcanza.

10.2. Obtención de la transformación Procrustes

Lema 10.1 *Sea A una matriz cuadrada y P cualquier matriz ortogonal. Entonces,*

$$\text{traza}(P'A) \leq \text{traza}((A'A)^{\frac{1}{2}}) \tag{10.5}$$

y la igualdad se verifica sólamente si $P'A = (A'A)^{\frac{1}{2}}$.

DEMOSTRACION:

Consideremos la descomposición en valores singulares (fue introducida en la Sección 7.1, pág. 75) $A = USV'$, en que S es la matriz de valores singulares (no negativos) y U, V son matrices ortogonales. Entonces,

$$\text{traza}(P'A) = \text{traza}(P'USV') = \text{traza}(V'P'US). \tag{10.6}$$

Pero $V'P'U$ es una matriz ortogonal que nunca tendrá valores mayores que 1 en la diagonal principal. Por tanto, la traza del término derecho de la ecuación anterior será la suma de los elementos diagonales de S multiplicados por números menores que la unidad. Tendremos:

$$\text{traza}(P'A) \leq \text{traza}(S) \tag{10.7}$$

y se verificará la igualdad sólo cuando $V'P'US = S$; esto último acontece, por ejemplo, para $P' = VU'$. Pero

$$\begin{aligned} \text{traza}(S) &= \text{traza}((S'S)^{\frac{1}{2}}) \\ &= \text{traza}((V'A'UU'AV)^{\frac{1}{2}}) \\ &= \text{traza}((A'A)^{\frac{1}{2}}), \end{aligned}$$

y esto junto con (10.7) establece (10.5). Veamos ahora la segunda aseveración. De

$$V'P'US = S \tag{10.8}$$

se deducen las siguientes desigualdades:

$$\begin{aligned} P'USV' = VSV' &\Rightarrow P'A = VSV' \\ &\Rightarrow P'A = (VS^2V')^{\frac{1}{2}} \\ &\Rightarrow P'A = (VSU'USV')^{\frac{1}{2}} \\ &\Rightarrow P'A = (A'A)^{\frac{1}{2}}, \end{aligned}$$

lo que finaliza la demostración. Podemos ahora resolver el problema de minimización (10.4).

10.2.1. Traslación a

Sean $\overline{x}$, $\overline{y}$ los vectores de medias aritméticas de las columnas de (respectivamente) X e Y. Definamos las matrices

$$\begin{aligned} \overline{X} &= \mathbf{1}\overline{x}' \\ \overline{Y} &= \mathbf{1}\overline{y}'. \end{aligned}$$

y versiones centradas de X e Y así:

$$\begin{aligned} \tilde{X} &= X - \overline{X} \\ \tilde{Y} &= Y - \overline{Y}. \end{aligned}$$

Observemos que

$$\begin{aligned} G(X,Y) &= \text{traza}((X-Y)(X-Y)') \\ &= \text{traza}((\tilde{X}-\tilde{Y})(\tilde{X}-\tilde{Y})') + N\text{traza}((\overline{X}-\overline{Y})(\overline{X}-\overline{Y})') \\ &= G(\tilde{X},\tilde{Y}) + N\text{traza}((\overline{X}-\overline{Y})(\overline{X}-\overline{Y})'); \end{aligned}$$

ello muestra que $G(X,Y)$ se hace mínimo cuando se calcula para configuraciones de puntos cuyos centroides han sido llevados a un origen común.

10.2.2. Rotación P.

Sean $\tilde{X}$ e $\tilde{Y}$ configuraciones centradas. Sean todas las transformaciones $\tilde{Y}P$ en que P es una matriz ortogonal $k \times k$. Tenemos

$$\begin{aligned} G(\tilde{X},\tilde{Y}P) &= \text{traza}((\tilde{X}-\tilde{Y}P)(\tilde{X}-\tilde{Y}P)') \\ &= \text{traza}(\tilde{X}\tilde{X}') + \text{traza}(\tilde{Y}\tilde{Y}') - 2\,\text{traza}(P'\tilde{Y}'\tilde{X}) \\ &\geq \text{traza}(\tilde{X}\tilde{X}') + \text{traza}(\tilde{Y}\tilde{Y}') \\ &\qquad -2\,\text{traza}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{\frac{1}{2}} \end{aligned} \tag{10.9}$$

en que el último paso hace uso del Lema 10.1. De acuerdo con dicho lema, el valor dado por (10.9) es alcanzable haciendo $P = \tilde{Y}'\tilde{X}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{-\frac{1}{2}}$.

10.2.3. Parámetro de escala ρ

El parámetro de escala es ahora muy fácil de obtener. Notemos que dejamos inalterada la escala de las $\tilde{X}$ y cambiamos sólo la de las $\tilde{Y}$. De otro modo, siempre podríamos obtener un valor de $G(\tilde{X},\tilde{Y}P)$ tan pequeño como

deseáramos, sin más que colapsar ambas configuraciones en una región arbitrariamente pequeña en torno al origen. Tenemos entonces que minimizar

$$G(\tilde{X}, \rho\tilde{Y}P) = \text{traza}(\tilde{X}\tilde{X}') + \rho^2\text{traza}(\tilde{Y}\tilde{Y}') - 2\rho\,\text{traza}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{\frac{1}{2}} \quad (10.10)$$

ecuación de segundo grado en ρ cuyo mínimo se alcanza para:

$$\rho = \frac{\text{traza}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{\frac{1}{2}}}{\text{traza}(\tilde{Y}\tilde{Y}')}. \quad (10.11)$$

10.3. Análisis y comentarios adicionales

Si reemplazamos el valor de ρ obtenido de (10.11) en la ecuación (10.10) obtenemos:

$$\begin{aligned} G_{\text{mín}}(\tilde{X}, \rho\tilde{Y}P) &= \text{traza}(\tilde{X}\tilde{X}') + \left[\frac{\text{traza}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{\frac{1}{2}}}{\text{traza}(\tilde{Y}\tilde{Y}')}\right]^2 \text{traza}(\tilde{Y}\tilde{Y}') \\ &\quad -2\left[\frac{\text{traza}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{\frac{1}{2}}}{\text{traza}(\tilde{Y}\tilde{Y}')}\right] \text{traza}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{\frac{1}{2}} \end{aligned}$$

que tras simplificar proporciona:

$$\begin{aligned} G_{\text{mín}}(\tilde{X}, \rho\tilde{Y}P) &= \text{traza}(\tilde{X}\tilde{X}') - \left[\frac{\text{traza}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{\frac{1}{2}}}{\text{traza}(\tilde{Y}\tilde{Y}')}\right] \text{traza}(\tilde{X}'\tilde{Y}\tilde{Y}'\tilde{X})^{\frac{1}{2}} \\ &= \text{traza}(\tilde{X}\tilde{X}') - \rho^2\text{traza}(\tilde{Y}\tilde{Y}') \end{aligned}$$

Reordenando la última igualdad tenemos:

$$G_{\text{mín}}(\tilde{X}, \rho\tilde{Y}P) + \rho^2\text{traza}(\tilde{Y}\tilde{Y}') = \text{traza}(\tilde{X}\tilde{X}'). \quad (10.12)$$

Podemo interpretar la igualdad (10.12) así: la "suma de cuadrados" de las distancias euclídeas de la configuración original $\tilde{X}$ se descompone en $\rho^2\text{traza}(\tilde{Y}\tilde{Y}')$ más una "suma de cuadrados de los errores", $G_{\text{mín}}$, que es lo que hemos minimizado. La igualdad (10.12) es así análoga a la que descompone la suma de cuadrados en el análisis de regresión o ANOVA.

Es de destacar que ρ al ajustar la configuración Y a la X no es en general el mismo (ni el inverso) del que se obtiene al ajustar la configuración X a la Y. Sin embargo, si normalizamos las configuraciones de modo que $\text{traza}(\tilde{X}\tilde{X}') = \text{traza}(\tilde{Y}\tilde{Y}') = 1$, ρ es el mismo en ambos casos, y la igualdad (10.12) se transforma en:

$$G_{\text{mín}}(\tilde{X}, \rho\tilde{Y}P) + \rho^2 = 1. \quad (10.13)$$

En tal caso, ρ^2 es directamente interpretable como la fracción de "suma de cuadrados" de distancias que la configuración adaptada es capaz de reproducir: ρ^2 juega aquí un papel similar al de R^2 en regresión.

Capítulo 11

Reescalado Multidimensional

11.1. Introducción.

Las técnicas conocidas colectivamente como de reescalado multidimensional (RM) *(Multidimensional Scaling, MDS)* tienen por objeto producir representaciones de reducida dimensionalidad de colecciones de objetos. Se diferencian del Análisis en Componentes Principales, Análisis Factorial y AC en el punto de partida. Mientras que en las técnicas citadas cada objeto viene descrito por un vector $\boldsymbol{x}_r$ que proporciona su posición en un espacio p-dimensional, en el caso de del Reescalado Multidimensional el punto de partida es una *matriz de proximidades.* Esta matriz puede contener *disimilaridades*, δ_{ij} en que un mayor valor δ_{ij} corresponde a una mayor desemejanza entre los objetos i y j o *similaridades*, verificando lo contrario.

No se hacen en principio supuestos acerca de la naturaleza de las similaridades o disimilaridades, que pueden obtenerse de muy diversos modos. Típicamente proceden de promediar las percepciones declaradas de un colectivo de sujetos interrogados, pero pueden tener cualquier otro origen.

El objetivo del Reescalado Multidimensional es producir una configuración de puntos, idealmente de muy baja dimensión, cuya distancia euclídea ordinaria reproduzca con la máxima fidelidad las disimilaridades δ_{ij}.

> **Ejemplo 11.1** *(semejanza entre códigos del alfabeto Morse)* En **?**, p. 54 se presenta un experimento realizado por **?**. Un colectivo de individuos escucha parejas de símbolos codificados en el alfabeto Morse, respondiendo si a su juicio son iguales o no. Para la pareja formada por los símbolos i y j se computa la disimilaridad δ_{ij} como el porcentaje de respuestas equivocadas (es decir, en las que el sujeto manifiesta que los dos símbolos no son iguales cuando lo son, o al contrario).

Hay símbolos que son fácilmente reconocibles como diferentes, incluso por un oído no entrenado (por ej., R, `.-.` y Q `-.-`). Otros, en cambio, son fácilmente confundibles. Obsérvese que pueden ser, y de hecho son, diferentes los porcentajes de confusión al escuchar la misma pareja de símbolos en los dos órdenes posibles: por tanto podríamos desear considerar $\delta_{ij} \neq \delta_{ji}$. Obsérvese además que dos símbolos idénticos no siempre son reconocidos como tales, y por tanto $\delta_{ii} \neq 0$ en general.

El empleo de la técnica del Reescalado Multidimensional produce una mapa en dos dimensiones en que la ubicación relativa de los símbolos es la esperable a la vista de su duración y composición de puntos y rayas. Por ejemplo, E (en Morse, `.`) y T (en Morse, `-`) aparecen en posiciones contiguas. Puede verse la configuración bidimensional y una interpretación de la misma en **?**, p. 59.

Ejemplo 11.2 *(reconstrucción de mapas a partir de información sobre distancias)* En ocasiones se emplea una matriz de disimilaridades obtenida de modo objetivo. Por ejemplo, podríamos construir una tabla de doble entrada cuyas filas y columnas se correspondieran con las capitales de provincia en España. En el lugar ij, podemos introducir como disimilaridad la distancia por carretera en kilómetros de una a otra. La configuración de puntos en dos dimensiones proporcionada por las técnicas de Reescalado Multidimensional debería aproximar la ubicación de las respectivas capitales de provincia. La configuración de puntos en dos dimensiones no reproduce con total fidelidad las posiciones de las capitales, porque las distancias consideradas lo son por carretera. La Figura 11.1, pág. 103 muestra el resultado de realizar un tipo de análisis de Reescalado Multidimensional.

11.2. Reescalado multidimensional métrico

La presentación sigue a **?**.

Imaginemos que tenemos las coordenadas de un conjunto de puntos. La distancia euclídea al cuadrado entre los puntos $\boldsymbol{x}_r$ y $\boldsymbol{x}_s$ vendría dada por:

$$d_{rs}^2 = \|\boldsymbol{x}_r - \boldsymbol{x}_s\|^2 = (\boldsymbol{x}_r - \boldsymbol{x}_s)' (\boldsymbol{x}_r - \boldsymbol{x}_s). \tag{11.1}$$

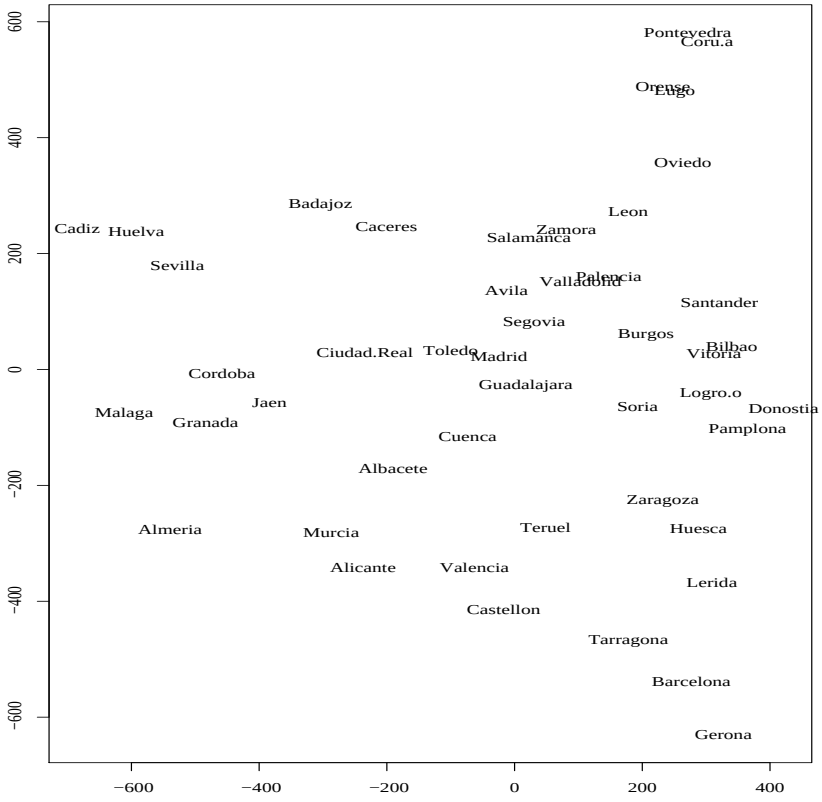
Sea X una matriz $N \times p$ cuya r-ésima fila es $\boldsymbol{x}_r{}'$. Definamos la matriz B cuyo elemento genérico b_{rs} viene dado por $\boldsymbol{x}_r{}'\boldsymbol{x}_s$. Claramente,

$$B = XX' \tag{11.2}$$

es cuadrada, simétrica y puede diagonalizarse:

$$B = V'\Lambda V. \tag{11.3}$$

Figura 11.1: Mapa reconstruido mediante reescalado multidimensional métrico a partir de las distancias por carretera entre capitales de provincia.



A partir de una tal B podríamos encontrar una configuración de puntos $\tilde{X}$ que la reproduce:

$$\tilde{X} = V'\Lambda^{\frac{1}{2}} \tag{11.4}$$

$$\tilde{X}' = \Lambda^{\frac{1}{2}}V. \tag{11.5}$$

El problema de encontrar una configuración de puntos que reproduce una cierta B, por tanto, está resuelto, al menos en tanto en cuanto dicha matriz B sea semidefinida positiva y admita una diagonalización como (11.3). La pregunta es si a partir de las distancias d_{rs}^2 podemos obtener una B para diagonalizarla.

Claramente, no puede haber solución única, porque toda traslación, rotación o reflexión de una configuración de puntos deja sus distancias invariadas. Por tanto, la solución estará indeterminada. No perderemos generalidad si suponemos un origen arbitrario, y por comodidad podemos suponer la nube de puntos centrada, es decir:

$$\frac{1}{N}\sum_{r=1}^{N}\boldsymbol{x}_r = \frac{1}{N}\sum_{s=1}^{N}\boldsymbol{x}_s = \mathbf{0}. \tag{11.6}$$

De (11.1) obtenemos:

$$d_{rs}^2 = \boldsymbol{x}_r{}'\,\boldsymbol{x}_r + \boldsymbol{x}_s{}'\,\boldsymbol{x}_s - 2\boldsymbol{x}_r{}'\boldsymbol{x}_s, \tag{11.7}$$

que sumando respecto de r, s y respecto de ambos índices a la vez proporciona en virtud de (11.6):

$$\frac{1}{N}\sum_{r=1}^{N}d_{rs}^2 = \frac{1}{N}\sum_{r=1}^{N}\boldsymbol{x}_r{}'\,\boldsymbol{x}_r + \boldsymbol{x}_s{}'\,\boldsymbol{x}_s \tag{11.8}$$

$$\frac{1}{N}\sum_{s=1}^{N}d_{rs}^2 = \frac{1}{N}\sum_{s=1}^{N}\boldsymbol{x}_s{}'\,\boldsymbol{x}_s + \boldsymbol{x}_r{}'\,\boldsymbol{x}_r \tag{11.9}$$

$$\frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N}d_{rs}^2 = \frac{2}{N}\sum_{r=1}^{N}\boldsymbol{x}_r{}'\,\boldsymbol{x}_r. \tag{11.10}$$

Por consiguiente, de (11.7) y haciendo uso de (11.8) a (11.10) tenemos que:

$$b_{rs} = \boldsymbol{x}_r{}'\boldsymbol{x}_s \tag{11.11}$$

$$= -\frac{1}{2}\left[d_{rs}^2 - \frac{1}{N}\sum_{r=1}^{N}d_{rs}^2 - \frac{1}{N}\sum_{s=1}^{N}d_{rs}^2 \right. \tag{11.12}$$

$$\left. + \frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N}d_{rs}^2\right]. \tag{11.13}$$

Llamando

$$a_{rs} = -\frac{1}{2}d_{rs}^2, \tag{11.14}$$

tenemos que

$$b_{rs} = a_{rs} - a_{r\cdot} - a_{\cdot s} + a_{\cdot\cdot} \tag{11.15}$$

en que $a_{r\cdot}$ denota el promedio de a_{rs} al sumar sobre el índice s (y análogamente para $a_{\cdot\cdot}$ y $a_{\cdot s}$). y si A es una matriz cuyo elemento genérico es a_{rs}, entonces

$$B = \left(I - \frac{1}{N}\mathbf{1}\,\mathbf{1}'\right) A \left(I - \frac{1}{N}\mathbf{1}\,\mathbf{1}'\right). \tag{11.16}$$

Hemos pues construido a partir de la matriz de distancias una matriz B a la que aplicar la factorización en (11.3). No siempre ocurrirá que B obtenida a partir de una matriz de disimilaridades pueda ser factorizada en la forma (11.3). Ello será imposible cuando B tenga valores propios negativos; en tal caso, es frecuente prescindir de los valores propios negativos, si no son muy grandes, o alterar la matriz de disimilaridades inicial añadiendo una constante c a cada disimilaridad d_{rs} con $r \neq s$. Siempre hay un c que hace que B obtenida a partir de las disimilaridades así transformadas sea semidefinida positiva.

Tenemos pues el siguiente algoritmo:

Algoritmo 1 – Reescalado multidimensional métrico.

1: Obtener una matriz de disimilaridades.
2: $A \leftarrow \left[-\frac{1}{2}d_{rs}^2\right]$.
3: $B \leftarrow \left(I - \frac{1}{N}\mathbf{1}\,\mathbf{1}'\right) A \left(I - \frac{1}{N}\mathbf{1}\,\mathbf{1}'\right)$.
4: Diagonalizar B:
$$B = V'\Lambda V.$$
Si no fuera semidefinida positiva, añadir una constante a las disimilaridades no diagonales, y recalcular; alternativamente, prescindir de los valores propios no positivos de B.
5: Obtener la configuración de puntos $\tilde{X}$:
$$\tilde{X} \leftarrow V'\Lambda^{\frac{1}{2}},$$
y retener el número de columnas deseado (normalmente, 2).

Obsérvese que si realmente existe una configuración de puntos X con matriz B dada por (11.3) y los datos están centrados como hemos supuesto en (11.6), B tiene los mismos valores propios que $X'X$. Es fácil ver entonces que las columnas de $\tilde{X}$ no son otra cosa que las componentes principales. El reescalado multidimensional métrico aplicado a una B procedente de una configuración de puntos en el espacio euclídeo no difiere pues (salvo en traslaciones, rotaciones o reflexiones) de la solución que obtendríamos mediante un análisis en componentes principales de los datos originales.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

11.1 Este es el código empleado en R para construir el mapa en la Figura 11.1. El objeto `spain` es una matriz triangular superior conteniendo las distancias en kilómetros entre capitales de provincia.

```
> distan <- spain + t(spain)
> distan[1:5,1:5]
         Albacete Alicante Almeria Avila Badajoz
Albacete        0      171     369   366     525
Alicante      171        0     294   537     696
Almeria       369      294       0   663     604
Avila         366      537     663     0     318
Badajoz       525      696     604   318       0
> library(mva)
> loc <- cmdscale(distan,k=2)
> x <- loc[,1]
> y <- loc[,2]
> postscript(file="mapa.eps")
> plot(x, y, type="n", xlab="", ylab="")
> text(x, y, names(distan))
```

Capítulo 12

Análisis discriminante

12.1. Introducción.

El problema que nos planteamos es el siguiente: tenemos una muestra de casos clasificados en dos o más grupos. Inicialmente consideraremos sólo dos grupos, para generalizar el análisis a continuación. Además de la clase o grupo a que pertenece cada caso, observamos p variables o características, y estamos interesados en saber si los valores de dichas p variables tienen alguna relación con la pertenencia a un grupo u otro.

La información disponible puede por tanto describirse como en la Tabla 12.1, en que las X son las características observadas y la variable C toma dos valores, C_1 ó C_2, indicativas de la pertenencia del caso correspondiente al primer o segundo grupo.

Un análisis discriminante puede tener objetivo:

- Descriptivo, si estamos sólo interesados en poner en evidencia la capacidad discriminante de un cierto conjunto de variables,
- Decisional, si buscamos un criterio que nos permita decidir sobre la adscripción a uno de los grupos de un caso nuevo, no perteneciente a la muestra de entrenamiento.

Es quizá el segundo objetivo el más usualmente perseguido. Se trata, de emplear la muestra de entrenamiento para buscar relaciones entre las variables X y la variable C_k, $k = 1, 2$, que permitan evaluar lo mejor posible ésta última como función de las primeras. Ello permite clasificar casos no pertenecientes a la muestra de entrenamiento. Los ejemplos siguientes muestran algunas de las muchísimas aplicaciones que se han dado al método.

Ejemplo 12.1 *(recuperación de información perdida)* En ocasiones, la variable C_k se ha perdido irreversiblemente. Por ejemplo, un esqueleto hallado en una necrópolis no contiene atributos que permitan su adscripción directa a un hombre o mujer.

Sin embargo, si contamos con una muestra de entrenamiento formada por esqueletos de los que sabemos si pertenecen a hombres y mujeres (por ejemplo, por la naturaleza de los objetos encontrados en el enterramiento), podemos tratar de ver si existe alguna asociación entre las medidas de los diversos huesos (las X) y el sexo del fallecido (C_k). Esto permite clasificar un nuevo esqueleto del que sólo observamos las X.

Ejemplo 12.2 *(información accesible al hombre, pero no a la máquina)* Hay problemas en los que la adscripción de un caso a un grupo es muy fácil de decidir para un humano, pero no para una máquina. Por ejemplo, reconocemos fácilmente las letras del alfabeto, incluso manuscritas. Sin embargo, el reconocimiento de las mismas por una máquina (a partir, por ejemplo, de una imagen explorada ópticamente), dista de ser trivial.

En un caso como éste, las variables X serían binarias (0=elemento de imagen o *pixel* blanco, 1=negro) o rasgos *(features)* que facilitaran la discriminación (por ejemplo, ratio altura/anchura de la letra, existencia de descendentes, ...).

Ejemplo 12.3 *(predicción)* En ocasiones, la adscripción a grupo es todavía incierta o inexistente, y el tratar de anticiparla es del mayor interés. Por ejemplo, sobre la base de análisis clínicos (cuyos resultados serían las X) un médico puede tratar de clasificar sus pacientes en aquéllos que presentan grave riesgo de padecer un infarto y aquéllos que no.

Análogamente, sobre la base de información sobre un cliente podemos intentar decidir si comprará o no un producto, o si entrará o no en morosidad si se le concede un crédito.

En ambos casos, la variable C_k todavía no ha tomado un valor, pero con ayuda de una muestra de casos en que si lo ha hecho, tratamos de anticipar el valor probable a la vista de las variables X observables.

Es importante notar que estamos ante un problema genuinamente estadístico, y no podemos habitualmente esperar un discriminación perfecta. Los grupos pueden tener cierto solapamiento (por ejemplo, de dos pacientes con exactamente los mismos valores de X, uno puede padecer un infarto y otro no).

Es también de interés señalar que es específico al análisis discriminante el contar con una muestra de entrenamiento: sabemos de partida a qué grupos pertenecen los componentes de la misma. Otro grupo de técnicas relacionadas (análisis de agrupamientos o análisis *cluster*) aborda el problema en que sólo conocemos las X, y queremos decidir sobre la existencia o no de grupos, cuantos, y cuáles. En la literatura sobre Inteligencia Artificial, técnicas como las del análisis discriminante se engloban en la denominación *aprendizaje*

Cuadro 12.1: Muestra de entrenamiento en análisis discriminante con dos grupos

X_{11}	$\dots$	X_{1p}	C_1
X_{21}	$\dots$	X_{2p}	C_1
$\vdots$		$\vdots$	$\vdots$
$X_{N_1 1}$	$\dots$	$X_{N_1 p}$	C_1
$X_{N_1+1,1}$	$\dots$	$X_{N_1+1,p}$	C_2
$X_{N_1+2,1}$	$\dots$	$X_{N_1+2,p}$	C_2
$\vdots$		$\vdots$	$\vdots$
$X_{N_1+N_2,1}$	$\dots$	$X_{N_1+N_2,p}$	C_2

supervisado, en tanto las del análisis de agrupamientos se describen como *aprendizaje no supervisado.*

12.2. Discriminación máximo-verosímil

Una manera conceptualmente simple e intuitiva de resolver el problema es abordarlo con criterio máximo verosímil. Asignaremos una observación con $\boldsymbol{X} = \boldsymbol{x}$ a la clase C_k si ésta tiene óptima capacidad generadora de la misma, es decir, si

$$f(\boldsymbol{x}|C_k) = \max_j f(\boldsymbol{x}|C_j). \tag{12.1}$$

Al margen de su carácter intuitivamente atrayente, es fácil demostrar que asignar a C_k cuando se verifica (12.1) minimiza la probabilidad total de error de asignación. En efecto, cualquier regla discriminante puede verse como una partición $\{R_1, R_2\}$ del dominio de definición $\mathcal{X}$ de las X, de forma que $\boldsymbol{x} \in R_1$ suponga asignar a C_1 y $\boldsymbol{x} \in R_2$ suponga asignar a C_2. La probabilidad total de error, $P(e)$, es entonces

$$P(e) = \int_{R_1} f(\boldsymbol{x}|C_2)d\boldsymbol{x} + \int_{R_2} f(\boldsymbol{x}|C_1)d\boldsymbol{x} \tag{12.2}$$

$$= \int_{R_1} f(\boldsymbol{x}|C_2)d\boldsymbol{x} + \int_{\mathcal{X}-R_1} f(\boldsymbol{x}|C_1)d\boldsymbol{x} \tag{12.3}$$

La primera integral en (12.2) es la probabilidad de que un caso perteneciente a la clase C_2 (con densidad por tanto $f(\boldsymbol{x}|C_2)$) esté en R_1. El valor de la integral es por tanto la probabilidad de uno de los tipos posibles de error: el de clasificar en C_1 (por ser $\boldsymbol{x} \in R_1$) un caso que en realidad pertenece a C_2. Análogamente, la segunda integral es la probabilidad de clasificar en C_2 un caso perteneciente a C_1.

En (12.3), $P(e)$ ha de minimizarse sobre R_1. Es claro entonces que, siendo los integrandos necesariamente no negativos, convendrá incluir en R_1 todos aquellos puntos de $\mathcal{X}$ tales que $f(\boldsymbol{x}|C_2) < f(\boldsymbol{x}|C_1)$ y en R_2 los que verifiquen lo contrario[1]. Esta es precisamente la regla (12.1).

Formalmente, de (12.3) obtenemos:

$$P(e) = \int_{R_1} f(\boldsymbol{x}|C_2)d\boldsymbol{x} + \int_{\mathcal{X}} f(\boldsymbol{x}|C_1)d\boldsymbol{x} - \int_{R_1} f(\boldsymbol{x}|C_1)d\boldsymbol{x} \quad (12.4)$$

$$= \int_{R_1} (f(\boldsymbol{x}|C_2) - f(\boldsymbol{x}|C_1))d\boldsymbol{x} + 1 \quad (12.5)$$

expresión que claramente queda minimizada si tomamos como R_1 la región de $\mathcal{X}$ definida así:

$$R_1 = \{\boldsymbol{x} : f(\boldsymbol{x}|C_2) - f(\boldsymbol{x}|C_1) \leq 0\} \quad (12.6)$$

La regla de asignación indicada puede además con gran facilidad modificarse de modo que tenga en cuenta información a priori y/o diferentes costos de error en la clasificación. Esta cuestión se detalla en la Sección que sigue, que generaliza y amplía la regla de asignación máximo verosímil dando entrada a información a priori.

Ejemplo 12.4 Las situaciones de fuerte asimetría en los costes de deficiente clasificación son la regla antes que la excepción. Por ejemplo, puede pensarse en las muy diferentes consecuencias que tiene el clasificar a una persona sana como enferma y a una persona enferma como sana. En el primer caso, el coste será quizá el de un tratamiento innecesario; el el segundo, el (normalmente mucho mayor) de permitir que un paciente desarrolle una enfermedad que quizá hubiera podido atajarse con un diagnóstico precoz.

Las situaciones con información a priori son también muy frecuentes. Un caso frecuente es aquél en que la abundancia relativa de los grupos es diferente, situación en la que tiene sentido adoptar probabilidades a priori diferentes para cada grupo (Sección 12.3).

12.3. Discriminación con información a priori

Es lo habitual que contemos con información a priori, distinta de la proporcionada por las X, acerca de la probabilidad de pertenencia a cada uno de los grupos considerados. Por ejemplo, si sabemos que la clase C_1 es nueve veces más numerosa que la clase C_2 en la población que analizamos, tendría sentido fijar a priori las probabilidades de pertenencia $P(C_1) = 0{,}9$ y $P(C_2) = 0{,}1$. La intuición sugiere, y el análisis que sigue confirma, que en tal situación la evidencia proporcionada por las X debería ser mucho más

[1] A efectos de probabilidad de error, los puntos verificando $f(\boldsymbol{x}|C_2) = f(\boldsymbol{x}|C_1)$ pueden arbitrariamente asignarse a cualquiera de las dos clases.

favorable a C_2 para lograr la asignación a dicha clase que cuando ambas clases son igual de numerosas.

El teorema de Bayes es cuanto necesitamos para incorporar información a priori a nuestra regla de decisión. En efecto, si consideramos la densidad conjunta $f(\boldsymbol{x}, C_k)$ tenemos que:

$$P(C_k|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|C_k)P(C_k)}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|C_k)P(C_k)}{\sum_j f(\boldsymbol{x}|C_j)P(C_j)} \tag{12.7}$$

La regla ahora será asignar $\boldsymbol{x}$ a aquella clase cuya probabilidad a posteriori $P(C_k|\boldsymbol{x})$ sea máxima. Por lo tanto, podemos particionar $\mathcal{X}$ en dos regiones, $\{R_1, R_2\}$ definidas así:

$$R_1 = \{\boldsymbol{x} : f(\boldsymbol{x}|C_1)P(C_1) > f(\boldsymbol{x}|C_2)P(C_2)\} \tag{12.8}$$

$$R_2 = \mathcal{X} - R_1 \tag{12.9}$$

Un argumento idéntico al empleado en la sección anterior muestra, en efecto, que actuando así minimizamos la probabilidad total de error. Obsérvese que, siendo el denominador de (12.7) el mismo en todos los casos, maximizar respecto a C_k el producto $f(\boldsymbol{x}|C_k)P(C_k)$ es equivalente a maximizar $P(C_k|\boldsymbol{x})$.

Por otra parte, al ser en (12.7) el denominador siempre el mismo,

$$P(C_k|\boldsymbol{x}) \propto f(\boldsymbol{x}|C_k)P(C_k). \tag{12.10}$$

Si todas las probabilidades a priori $P(C_k)$ son iguales, $P(C_k|\boldsymbol{x}) \propto f(\boldsymbol{x}|C_k)$ y la regla bayesiana coincide con la máximo verosímil, pues (12.1) y (12.10) alcanzan el máximo para la misma clase C_k. Cuando hay información a priori, los resultados pueden en cambio variar sustancialmente. El ejemplo siguiente, una situación artificialmente simple de control de calidad presentada como un problema de análisis discriminante, lo muestra.

Ejemplo 12.5 Una prensa moldea piezas en lotes de 100 a la vez. La experiencia muestra que con probabilidad 0.99 se obtienen lotes casi perfectos, con un 2 % de fallos. Con probabilidad 0.01, sin embargo, se obtienen lotes de muy mala calidad, con un 30 % de piezas defectuosas.

Supongamos que para decidir si un lote es "bueno" (B) o "malo" (M) tenemos la posibilidad de extraer una pieza al azar del lote, que examinada puede ser "correcta" (c) ó "defectuosa" (d). Podemos ver este problema de decisión como un problema de análisis discriminante, en que observamos una única variable X —el estado de la pieza examinada— y hemos de decidir la clase a la que pertenece el lote muestreado (B ó M).

Supongamos que examinamos una pieza extraída de un lote y resulta ser defectuosa. Si nos limitamos a seguir el criterio máximo verosímil sin considerar la información a priori, tendríamos,

$$P(X = d|B) = 0{,}02 \tag{12.11}$$

$$P(X = d|M) = 0{,}30, \tag{12.12}$$

a la vista de lo cual concluiríamos que el lote es M. La situación es completamente diferente si consideramos la información a priori que tenemos, pues entonces hemos de comparar:

$$P(B|X=d) = \frac{P(X=d|B)P(B)}{P(X=d)} = \frac{0{,}02\times 0{,}99}{0{,}02\times 0{,}99+0{,}3\times 0{,}01} = 0{,}8684 \quad (12.13)$$

$$P(M|X=d) = \frac{P(X=d|M)P(M)}{P(X=d)} = \frac{0{,}30\times 0{,}01}{0{,}02\times 0{,}99+0{,}3\times 0{,}01} = 0{,}1316 \quad (12.14)$$

Pese a ser la pieza examinada defectuosa, la probabilidad a posteriori de que el lote examinado sea bueno sigue siendo superior. En otras palabras, es tan grande el "prejuicio" a favor de que el lote examinado sea bueno que no basta encontrar una sola pieza defectuosa para derrotarlo.

Obsérvese que, como ya ha sido hecho notar, los denominadores en (12.13) y (12.14) son idénticos, por lo que a efectos de decidir cuál es la clase con mayor probabilidad a posteriori bastaba con calcular los numeradores. Estos numeradores, o cualquier transformación monótona de los mismos, se denominan *funciones discriminantes*. En la práctica, se estiman las funciones discriminantes con ayuda de la muestra de entrenamiento, y luego basta evaluar cada una de ellas para los nuevos casos a clasificar.

El caso de diferentes costes de error, arriba mencionado, puede ser tratado de forma simple. Si en lugar de la probabilidad de error minimizamos el coste medio total de error, la expresión a minimizar se transforma en

$$C(e) = \ell_2 \int_{R_1} f(\boldsymbol{x}|C_2)P(C_2)d\boldsymbol{x} + \ell_1 \int_{\mathcal{X}-R_1} f(\boldsymbol{x}|C_1)P(C_1)d\boldsymbol{x} \quad (12.15)$$

en que ℓ_i $(i=1,2)$ es el coste asociado a clasificar mal un caso del grupo i-ésimo. Las integrales en (12.15) son las probabilidades a posteriori de que un caso en el grupo C_2 (o C_1) quede clasificado en el grupo C_1 (respectivamente C_2). Un desarrollo idéntico al efectuado más arriba lleva a ver que la regla de clasificación minimizadora consiste en tomar R_1 la región del espacio $\mathcal{X}$ definida así:

$$R_1 = \{\boldsymbol{x} : \ell_2 f(\boldsymbol{x}|C_2)P(C_2) - \ell_1 f(\boldsymbol{x}|C_1)P(C_1) \leq 0\} \quad (12.16)$$

Hemos razonado para el caso de dos grupos, pero la generalización a K grupos es inmediata. Para cada caso $\boldsymbol{x}$ a clasificar y grupo C_j, $(j=1,\ldots,K)$, evaluaremos las funciones discriminantes $y_i(\boldsymbol{x})$, $i=1,\ldots,K$. Asignaremos al grupo k si $y_k(\boldsymbol{x}) = \text{máx}_j\, y_j(\boldsymbol{x})$. Las funciones discriminantes serán

$$y_j(\boldsymbol{x}) = f(\boldsymbol{x}|C_j)P(C_j). \quad (12.17)$$

En el caso de que tengamos una matriz de costes asociados a deficiente clasificación, $L = \{\ell_{ij}\}$, en que ℓ_{ij} es el coste de clasificar en C_j un caso que pertenece a C_i, asignaríamos a C_j si

$$j \quad = \quad \arg\min_j \sum_i \ell_{ij} f(\boldsymbol{x}|C_i) P(C_i). \tag{12.18}$$

Como funciones discriminantes $y_j(\boldsymbol{x})$ podríamos emplear cualesquiera que fueran transformaciones monótonas de las que aparecen en el lado derecho de (12.18).

12.4. Variables normales

El desarrollo anterior presupone conocidas las funciones de densidad o probabilidad $f(\boldsymbol{x}|C_k)$, y, en su caso, las probabilidades a priori de pertenencia a cada grupo. En ocasiones (como en el Ejemplo 12.5 anterior) puede admitirse que dichas funciones son conocidas. Pero en el caso más habitual, tenemos que estimar $f(\boldsymbol{x}|C_k)$ y el modelo más frecuentemente utilizado es el normal multivariante.

Al margen de su interés y aplicabilidad en sí mismo, por ser adecuado a multitud de situaciones, sucede que los resultados a que da lugar son muy simples (variables discriminantes lineales, en el caso más habitual) y pueden ser justificados de modos alternativos (empleando el enfoque de Fisher, como veremos más abajo). Esto hace que las reglas discriminantes que describimos a continuación sean las más empleadas en la práctica. Si las observaciones obedecen aproximadamente un modelo normal multivariante, los resultados son óptimos en el sentido en que la discriminación bayesiana lo es. Si la aproximación normal no es buena, la discriminación lineal todavía es justificable desde perspectivas alternativas. En algunos casos, que mencionaremos, el problema simplemente no se presta a una discriminación lineal y hay que emplear procedimientos diferentes.

12.4.1. Matriz de covarianzas Σ común y dos grupos

Cuando $f(\boldsymbol{x}|C_k) \sim N(\boldsymbol{\mu}_k, \Sigma)$, $k = 1, 2$, la regla de decisión consiste en asignar al grupo C_1 si:

$$\ell_2 f(\boldsymbol{x}|C_2) P(C_2) - \ell_1 f(\boldsymbol{x}|C_1) P(C_1) \leq 0 \tag{12.19}$$

equivalente, tras sencillas manipulaciones, a:

$$\frac{(2\pi)^{-p/2}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right\}}{(2\pi)^{-p/2}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\}} \geq \frac{\ell_2 P(C_2)}{\ell_1 P(C_1)}. \tag{12.20}$$

Simplificando y tomando logaritmos, la expresión anterior es equivalente a

$$-(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+(\boldsymbol{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\geq 2\log_e\left(\frac{\ell_2 P(C_2)}{\ell_1 P(C_1)}\right).$$

Tras realizar los productos en las formas cuadráticas del lado izquierdo y cancelar términos iguales, obtenemos la regla:

"Asignar a C_1 si:

$$\boldsymbol{x}'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\geq\frac{1}{2}\boldsymbol{\mu}_1'\Sigma^{-1}\boldsymbol{\mu}_1-\frac{1}{2}\boldsymbol{\mu}_2'\Sigma^{-1}\boldsymbol{\mu}_2+\log_e\left(\frac{\ell_2 P(C_2)}{\ell_1 P(C_1)}\right) \quad (12.21)$$

y a C_2 en caso contrario."

Vemos que el lado derecho de (12.21) es constante, y su valor c puede ser estimado una sola vez. El lado izquierdo es una forma lineal $\boldsymbol{a}'\boldsymbol{x}$ en que los coeficientes $\boldsymbol{a}$ también pueden ser estimados una sola vez. Hecho esto, la regla discriminante es tan simple como evaluar para cada nuevo caso una función lineal $\boldsymbol{a}'\boldsymbol{x}$ y comparar el valor obtenido con el umbral c:

"Asignar $\boldsymbol{x}$ a C_1 si $\boldsymbol{a}'\boldsymbol{x}\geq c$, y a C_2 en caso contrario."

Las estimaciones tanto de $\boldsymbol{a}$ como de c se obtienen sustituyendo $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ y Σ por sus respectivos estimadores.

Aunque en la forma expresada la regla discriminante es de utilización muy simple, podemos definir también funciones discriminantes

$$y_1(\boldsymbol{x}) = \boldsymbol{a}'\boldsymbol{x}-c \quad (12.22)$$

$$y_2(\boldsymbol{x}) = c-\boldsymbol{a}'\boldsymbol{x} \quad (12.23)$$

asignando $\boldsymbol{x}$ al grupo k si $y_k(\boldsymbol{x})$ es máximo.

Obsérvese que $\ell_1, \ell_2, P(C_1)$ y $P(C_2)$ sólo intervienen en la regla discriminante modificando el umbral que $\boldsymbol{a}'\boldsymbol{x}$ debe superar para dar lugar a asignación al grupo C_1. La influencia sobre dicho umbral es la esperable: mayores valores de ℓ_2 (coste de clasificar en C_1 un caso que realmente pertenece a C_2) y $P(C_2)$ incrementan el umbral, en tanto mayores valores de ℓ_1 y $P(C_1)$ lo disminuyen.

12.4.2. Diferentes covarianzas: $\Sigma_1 \neq \Sigma_2$, y dos grupos

El análisis es enteramente similar, pero el resultado menos simple. En efecto, en lugar de la expresión (12.20) tenemos ahora

$$\frac{(2\pi)^{-p/2}|\Sigma_1|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right\}}{(2\pi)^{-p/2}|\Sigma_2|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\}}\geq\frac{\ell_2 P(C_2)}{\ell_1 P(C_1)},$$

que tomando logaritmos, proporciona:

$$-(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+(\boldsymbol{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \;\geq\; 2\log_e\left(\frac{\ell_2 P(C_2)|\Sigma_2|^{-\frac{1}{2}}}{\ell_1 P(C_1)|\Sigma_1|^{-\frac{1}{2}}}\right).$$

Simplificando y llevando constantes al lado derecho, obtenemos:

$$\begin{aligned}-\boldsymbol{x}'(\Sigma_1^{-1}-\Sigma_2^{-1})\boldsymbol{x}+2\boldsymbol{x}'(\Sigma_1^{-1}\boldsymbol{\mu}_1-\Sigma_2^{-1}\boldsymbol{\mu}_2) \;\geq\; & 2\log_e\left(\frac{\ell_2 P(C_2)|\Sigma_2|^{-\frac{1}{2}}}{\ell_1 P(C_1)|\Sigma_1|^{-\frac{1}{2}}}\right)\\ & +\boldsymbol{\mu}_1{}'\Sigma_1^{-1}\boldsymbol{\mu}_1\\ & -\boldsymbol{\mu}_2{}'\Sigma_2^{-1}\boldsymbol{\mu}_2. \qquad (12.24)\end{aligned}$$

No ha habido en (12.24) cancelación del término cuadrático en $\boldsymbol{x}$ como ocurre cuando $\Sigma_1=\Sigma_2$. La regla discriminante es ahora

"Asignar $\boldsymbol{x}$ a C_1 si $\boldsymbol{x}'A\boldsymbol{x}+\boldsymbol{a}'\boldsymbol{x}\geq c$, y a C_2 en caso contrario."

en que:

$$\begin{aligned}A &= -(\Sigma_1^{-1}-\Sigma_2^{-1})\\ \boldsymbol{a} &= 2(\Sigma_1^{-1}\boldsymbol{\mu}_1-\Sigma_2^{-1}\boldsymbol{\mu}_2)\\ c &= 2\log_e\left(\frac{\ell_2 P(C_2)|\Sigma_2|^{-\frac{1}{2}}}{\ell_1 P(C_1)|\Sigma_1|^{-\frac{1}{2}}}\right)+\boldsymbol{\mu}_1{}'\Sigma_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2{}'\Sigma_2^{-1}\boldsymbol{\mu}_2.\end{aligned}$$

La frontera entre las dos regiones en que queda dividido el espacio $\mathcal{X}$ es ahora una hiper-superficie de ecuación cuadrática, mientras que cuando $\Sigma_1=\Sigma_2$ dicha hiper-superficie es un hiper-plano.

12.4.3. Caso de varios grupos

El desarrollo al final de la Sección 12.3 es ahora de aplicación, sustituyendo en (12.18) las densidades por sus expresiones correspondientes. Algunos casos particulares son de interés. Si $\ell_{ij}=1$ para $i\neq j$ y $\ell_{ii}=0$ para todo i, entonces la regla será asignar al grupo C_i cuando

$$i \;=\; \arg\max_j\left\{\frac{1}{(\sqrt{2\pi})^p|\Sigma_j|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)}P(C_j)\right\},$$

o, tomando logaritmos y prescindiendo de constantes, cuando:

$$i \;=\; \arg\max_j\left\{-\log_e|\Sigma_j|^{\frac{1}{2}}-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)+\log_e P(C_j)\right\}.$$

En el caso aún más particular de matrices de covarianzas idénticas, la regla anterior se reduce a asignar a C_i cuando

$$i \;=\; \arg\max_j\left\{\log_e P(C_j)+(\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}_j)'\Sigma^{-1}\boldsymbol{\mu}_j\right\}.$$

12.5. La regla lineal de Fisher

Fisher propuso en 1936 un procedimiento de discriminación lineal que coincide con la regla derivada para dos poblaciones normales con matriz de covarianzas común. En la aproximación de Fisher, la normalidad no es un supuesto. En cambio, la linealidad sí que lo es, en lugar de aparecer como un resultado.

12.5.1. Dos grupos con matriz de covarianzas Σ común

El razonamiento es el siguiente: buscamos una función lineal $\boldsymbol{a}'\boldsymbol{x}$ que separe óptimamente dos grupos, en un sentido que veremos. Ello requiere que $\boldsymbol{a}'\boldsymbol{x}$ tome valores "altos" en promedio para valores en un grupo, y "bajos" en otro. Una manera de requerir esto, es buscar un $\boldsymbol{a}$ que maximice

$$\left[\boldsymbol{a}'\boldsymbol{\mu}_1 - \boldsymbol{a}'\boldsymbol{\mu}_2\right]^2 = \left[\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^2, \tag{12.25}$$

es decir, que separe bien los vectores de medias de ambos grupos. El cuadrado tiene por objeto eliminar el signo, pues nos importa la diferencia de $\boldsymbol{a}'\boldsymbol{x}$ evaluada en $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$, y no su signo.

Maximizar (12.25) es un problema mal especificado: basta multiplicar $\boldsymbol{a}$ por $\alpha > 1$ para incrementar (12.25). Esto carece de interés: no estamos interesados en maximizar el valor numérico de (12.25) *per se*, sino en lograr que tome valores lo más claramente diferenciados posibles para casos en cada uno de los dos grupos.

Un modo de obtener una solución única es fijando la escala de $\boldsymbol{a}$. Podríamos fijar $||\boldsymbol{a}||^2 = 1$, pero, como veremos en lo que sigue, tiene mayor atractivo hacer $\boldsymbol{a}'\Sigma\boldsymbol{a} = 1$; o, alternativamente, resolver

$$\max_{\boldsymbol{a}} \left(\frac{\left[\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^2}{\boldsymbol{a}'\Sigma\boldsymbol{a}} \right), \tag{12.26}$$

que es de nuevo un problema indeterminado hasta un factor de escala[2], y normalizar una solución cualquiera de modo que $\boldsymbol{a}'\Sigma\boldsymbol{a} = 1$.

Adoptemos esta última vía. Derivando (12.26) respecto de $\boldsymbol{a}$ e igualando el numerador a cero, obtenemos (véase Apéndice **??**)

$$2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{a}'[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2](\boldsymbol{a}'\Sigma\boldsymbol{a}) - 2\left[\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^2 \Sigma\boldsymbol{a} = \boldsymbol{0}. \tag{12.27}$$

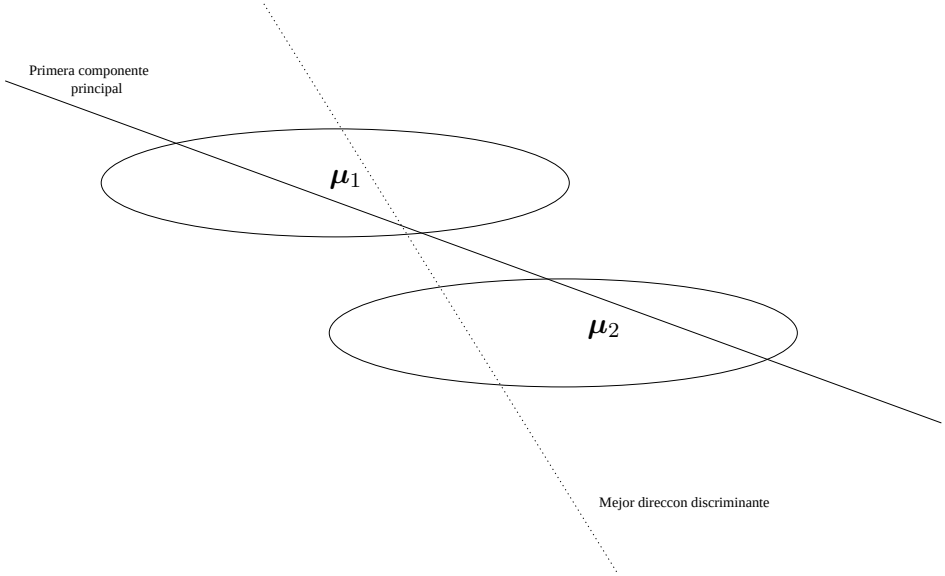
Si prescindimos de las constantes, vemos que (12.27) proporciona

$$\Sigma\boldsymbol{a} \propto (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \Rightarrow \boldsymbol{a} \propto \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{12.28}$$

que es la solución que ya teníamos para $\boldsymbol{a}$ en la Sección 12.4.1.

[2]Pues (12.26) es invariante al multiplicar $\boldsymbol{a}$ por una constante cualquiera.

Figura 12.1: La mejor dirección discriminante puede no ser aquélla en que más dispersión presentan las observaciones



La expresión (12.26) cuya maximización proporciona $\boldsymbol{a}$ (hasta una constante de proporcionalidad, como se ha visto) es de interés. Obsérvese que el denominador es la varianza de $\boldsymbol{a}'\boldsymbol{X}$. El numerador es el cuadrado de la diferencia entre los valores que toma $\boldsymbol{a}'\boldsymbol{X}$ en $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$. Lo que se maximiza, pues, es la razón de esta diferencia al cuadrado de valores de $\boldsymbol{a}'\boldsymbol{X}$ *en términos de su propia varianza, var*$(\boldsymbol{a}'\boldsymbol{X})$.

Podemos ver (12.26) como una relación señal/ruido: el numerador es la "señal" y el denominador el "ruido." Buscamos pues una función $\boldsymbol{a}'\boldsymbol{X}$ que maximice la relación señal/ruido.

Es importante observar que la dirección en la que las observaciones presenta máxima dispersión (que corresponde a la primera componente principal) *no necesariamente* es la mejor dirección discriminante, incluso aunque a lo largo de la misma los vectores de medias de los grupos resultasen máximamente separados. La Figura 12.1 es ilustrativa: se muestran contornos de igual densidad de dos grupos, y una línea sólida en la dirección de la primera componente principal. En esta dirección se presenta la máxima varianza de las observaciones. Sin embargo, es fácil ver que en la dirección de la línea discontinua se obtiene una separación mucho mejor de los dos grupos: es la dirección de $\boldsymbol{a}$ en (12.28).

12.5.2. Más de dos grupos con matriz de covarianzas Σ común

Conceptualmente el planteamiento es idéntico, pero los resultados son más complejos. Si hay K grupos, hay en general no una sino hasta $K-1$ variables discriminantes, combinaciones lineales de las $\boldsymbol{X}$ originales.

Sean pues K grupos, y consideremos una muestra de entrenamiento con n_i casos $(i = 1, \ldots, K)$ en cada grupo. El tamaño total de la muestra es así $n = \sum_{i=1}^{K} n_i$. Denotamos por $\boldsymbol{X}_{i(j)}$ la observación i-ésima en el grupo j-ésimo. Definamos:

$$\overline{X} = n^{-1} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \boldsymbol{X}_{i(j)} \tag{12.29}$$

$$\overline{X}_i = n_i^{-1} \sum_{j=1}^{n_i} \boldsymbol{X}_{i(j)} \tag{12.30}$$

$$T = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{X}_{i(j)} - \overline{X})(\boldsymbol{X}_{i(j)} - \overline{X})' \tag{12.31}$$

$$W_i = \sum_{j=1}^{n_i} (\boldsymbol{X}_{i(j)} - \overline{X}_i)(\boldsymbol{X}_{i(j)} - \overline{X}_i)' \tag{12.32}$$

$$W = W_1 + \ldots + W_K \tag{12.33}$$

$$B = T - W. \tag{12.34}$$

Es entonces fácil demostrar (véase Ejercicio 12.1) que $B = \sum_{i=1}^{K} n_i (\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})'$ y $\overline{X} = n^{-1} \sum_{i=1}^{K} n_i \overline{X}_i$. Un razonamiento similar al empleado al obtener el discriminante lineal en el caso de dos grupos, sugeriría ahora maximizar

$$\frac{\sum_{i=1}^{K} \left[\boldsymbol{a}' \sqrt{n_i} (\overline{X}_i - \overline{X}) \right]^2}{\sum_{i=1}^{K} \left[\boldsymbol{a}' \sum_{j=1}^{n_i} (\boldsymbol{X_{i(j)}} - \overline{X}_i) \right]^2} = \frac{\boldsymbol{a}' B \boldsymbol{a}}{\boldsymbol{a}' W \boldsymbol{a}} \stackrel{\text{def}}{=} \lambda. \tag{12.35}$$

Derivando respecto a $\boldsymbol{a}$ obtenemos la igualdad matricial

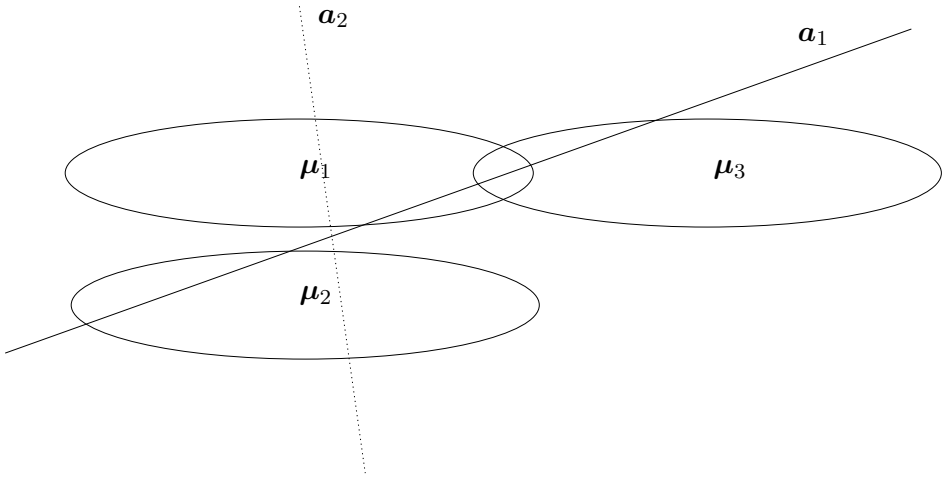
$$(B - \lambda W)\boldsymbol{a} = \boldsymbol{0}. \tag{12.36}$$

Bajo el supuesto de que W tiene inversa, la igualdad anterior es equivalente a

$$(W^{-1}B - \lambda I)\boldsymbol{a} = \boldsymbol{0}. \tag{12.37}$$

Esta tiene solución no trivial para valores λ y vectores $\boldsymbol{a}$ que son respectivamente valores y vectores propios de la matriz cuadrada $W^{-1}B$. Hay a lo sumo $q = \text{mín}(p, K-1)$ valores propios no nulos (por ser este el rango de B y por tanto de $W^{-1}B$; Ejercicio 12.2).

Figura 12.2: Con $p = 3$ grupos hay hasta $p - 1$ direcciones discriminantes. Puede haber direcciones discriminantes asociadas a un λ bajo, y no obstante muy útiles para discriminar en algún subconjunto. Por ejemplo, la dirección asociada a $\boldsymbol{a}_2$ discrimina bien entre los grupos C_1 y C_2 por un lado y C_3 por otro.



Es interesante observar lo que proporciona el método. Si hubiéramos de retener una sola dirección discriminante —como hacíamos en el caso de dos grupos—, tomaríamos la determinada por $\boldsymbol{a}_1$, siendo $(\lambda_1, \boldsymbol{a}_1)$ el par formado por el mayor valor propio y su vector propio asociado. En efecto, tal elección de $\boldsymbol{a}$ maximiza el cociente

$$\lambda = \frac{\boldsymbol{a}'B\boldsymbol{a}}{\boldsymbol{a}'W\boldsymbol{a}}$$

(véase Ejercicio 12.3). Pero puede haber otras direcciones (como la asociada a $\boldsymbol{a}_2$ en la Figura 12.2) "especializadas" en separar algún subconjunto de los grupos (C_1 y C_2 por un lado y C_3 por otro, en la Figura 12.2). Obsérvese que los vectores propios de $W^{-1}B$, y por tanto las direcciones discriminantes, no son en general ortogonales, pues $W^{-1}B$ no es simétrica.

Observación 12.1 Hay una interesante relación entre la solución anterior y los resultados que derivarían de análisis de correlación canónica y MANOVA equivalentes. Si completamos los datos de la muestra de entrenamiento con K columnas con valores 0 y 1 tal como en la ecuación (4.12), pág. 52, obtendríamos pares de variables canónicas incorreladas y con correlación entre ellas respectivamente máxima. Los vectores $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{K-1}$ coincidirían con los obtenidos al hacer análisis discriminante lineal de los K grupos. Los vectores de coeficientes $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{K-1}$ de las variables canónicas "parejas", aportarían una información interesante: son combinaciones de variables 0-1 que resultan

máximamente correladas con las $\boldsymbol{a}_1{}'\boldsymbol{X}, \ldots, \boldsymbol{a}_{K-1}{}'\boldsymbol{X}$, e indican entre qué grupos discriminan dichas variables.

12.6. Evaluación de funciones discriminantes

Estimadas la o las funciones discriminantes con ayuda de la muestra de entrenamiento, hay interés en tener un modo de medir su eficacia en la separación de grupos. Conceptualmente, no hay mucha diferencia entre evaluar una función discriminante y un modelo de regresión. En el caso de una función discriminante el problema es más arduo, por causa de la (habitualmente) elevada dimensionalidad. Nos limitaremos a algunas ideas básicas: un tratamiento más completo puede encontrarse en **?**.

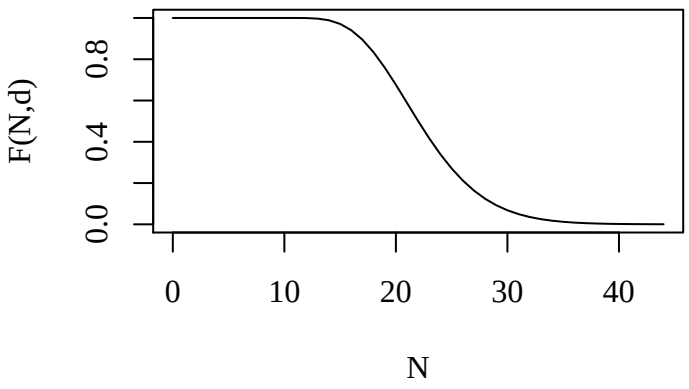
La idea que primero acude a nuestra mente es la de examinar el comportamiento de la función discriminante sobre la muestra de entrenamiento. ¿Clasifica bien los casos en dicha muestra? Esto es similar a examinar el ajuste —quizá mediante el R^2— de un modelo de regresión lineal. Alternativamente, podríamos llevar a cabo un análisis MANOVA para contrastar la hipótesis de igualdad de grupos: esto sería similar a contrastar la nulidad de todos los parámetros en un modelo de regresión lineal.

Sin embargo, a poco grande que sea el número de variables empleadas en la discriminación, la *tasa de error aparente* (la tasa de error al reclasificar la muestra de entrenamiento) será una estimación muy optimista. Al emplear la función discriminante sobre datos diferentes a los de la muestra de entrenamiento, obtendremos tasas de error, por lo general, sensiblemente mayores.

> **Observación 12.2** En esencia, la razón por la que la tasa de error aparente es un estimador optimista de la tasa de error real esperable es la misma que hace que $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ sea un estimador optimista de la varianza poblacional: hemos reemplazado $E(X)$ por $\overline{X}$, el estimador de la media que mejor se adapta a la muestra (en términos de suma de cuadrados residual). No es extraño que $\hat{\sigma}^2$ sea sesgado por defecto. Este sesgo es el que se corrige sustrayendo del denominador n el número de grados de libertad consumidos (en este caso, uno), lo que proporciona el estimador insesgado habitual $(n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
>
> En el análisis discriminante, la probabilidad de obtener una separación espúrea cuando podemos fijar la posición del hiperplano separador en un espacio elevadamente dimensional, es sorprendentemente alta, como el Teorema 12.1 más abajo pone de manifiesto.

Una percepción intuitiva de lo extremadamente optimista que puede resultar una función discriminante lineal en un espacio de elevada dimensionalidad puede obtenerse así: consideremos N puntos procedentes todos de una misma distribución d-dimensional, etiquetados al azar como proviniendo la mitad de ellos del grupo G1 y la otra mitad del G2. La probabilidad teórica

Figura 12.3: Probabilidad $F(N, d)$ de separar perfectamente N puntos en posición general en un espacio de $d = 10$ dimensiones



de que un procedimiento cualquiera asigne bien un punto sería de $p = 0{,}5$: los puntos provienen en realidad de la misma distribución, y no podemos obtener mejor tasa de error que la que resultaría de asignar puntos a uno u otro grupo lanzando una moneda al aire.

La probabilidad de encontrar un hiperplano que separa *perfectamente* los puntos aleatoriamente asignados a un grupo de los asignados al otro, es sin embargo bastante apreciable, como se deduce del siguiente teorema debido a Cover (ver **?**, pág. 86-87).

Teorema 12.1 *La probabilidad $F(N, d)$ de perfecta separación de N puntos en posición general en un espacio d dimensional viene dada por*

$$F(N,d) \quad = \quad \begin{cases} 1 & si\ N \leq d+1 \\ 2^{-N+1} \sum_{i=0}^{d} \binom{N-1}{i} & cuando\ N \geq d+1. \end{cases} \tag{12.38}$$

Si representamos gráficamente $F(N, d)$ frente a N (para $d = 10$), obtenemos una gráfica como la de la Figura 12.3. Hasta que el número de puntos N duplica el de dimensiones d, la probabilidad de perfecta separabilidad es superior a $\frac{1}{2}$. Separaciones no perfectas se obtienen con probabilidad aún mayor, pese a que los puntos son indistinguibles.

Hay varias opciones para combatir el sesgo en la tasa de error aparente. Podemos evaluar la función discriminante sobre una muestra de validación,

distinta de la que ha servido para estimar la función: ello dará una estimación insesgada de la tasa de error.

Si no disponemos de una muestra de validación, podemos recurrir a hacer validación cruzada, consistente en subdividir la muestra en K partes, estimar la función discriminante con $(K-1)$ de ellas y evaluar sobre la restante. Si hacemos que cada una de las K partes sea por turno la muestra de validación, tenemos la técnica de *validación cruzada*: obtenemos K diferentes estimadores de la tasa de error —cada uno de ellos, dejando fuera a efectos de validación una de las K partes en que se ha subdividido la muestra—, y podemos promediarlos para obtener un estimador final. En el caso extremo *(leave one out)*, podemos dividir la muestra en N partes consistentes en una única observación, estimar N funciones discriminantes con $(N-1)$ observaciones y asignar la restante tomando nota del acierto o error. El total de errores dividido entre N estimaría la tasa de error.

12.7. Bibliografía comentada

Casi todos los manuales de Análisis Multivariante contienen una introducción al análisis discriminante. Ejemplos son **?**, **?**, y **?**.

Una monografía algo antigua pero todavía de valor es **?**, que contiene mucha bibliografía. **?** es otro libro que continua manteniendo su interés. Más actual, con una buena bibliografía, es **?**.

Una monografía moderna es **?**; no tiene estructura de texto, ni es quizá la fuente más adecuada para una primera aproximación al tema, pero es útil para profundizar en el mismo. **?** es un libro sobre redes neuronales, especialmente aplicadas a reconocimiento de pautas y desde una perspectiva estadística; el Capítulo 3 compara la versión más simple de perceptrón con el método clásico de Fisher. El resto del libro es también de interés.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

12.1 En la Sección 12.5.2 se ha definido $B = T - W$. Demuéstrese que

$$B = \sum_{i=1}^{K} n_i (\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})'. \tag{12.39}$$

Ayuda: puede sumarse y restarse $\overline{X}_i$ en cada uno de los paréntesis de la definición (12.31) de T.

12.2 ($\uparrow$ 12.1) Demuéstrese que B tiene rango no mayor que $K-1$.

12.3 Demostrar que si λ y $\boldsymbol{a}$ son respectivamente un valor propio de $W^{-1}B$ y el correspondiente vector propio asociado, entonces

$$\lambda = \frac{\boldsymbol{a}'B\boldsymbol{a}}{\boldsymbol{a}'W\boldsymbol{a}}.$$

12.4 Compruébese que en el caso de diferentes costes de mala clasificación y distribución normal, las funciones discriminantes son en general no lineales, incluso aunque las matrices de covarianzas intra-grupos sean idénticas.

12.5 Sea un problema de discriminación entre dos grupos con n_1 y n_2 observaciones en la muestra de entrenamiento. Muéstrese que si estimamos el modelo de regresión lineal,

$$y_i = \boldsymbol{x}_i{}'\boldsymbol{\beta} + \epsilon_i$$

con

$$y_i = \begin{cases} \frac{n_2}{n_1+n_2} & \text{si } i = 1, \ldots, n_1, \\ -\frac{n_1}{n_1+n_2} & \text{si } i = n_1+1, \ldots, n_1+n_2. \end{cases}$$

y $\boldsymbol{x}_i$ = vector de variables correspondiente al caso i-ésimo, entonces el $\hat{\beta}$ obtenido por MCO coincide con el $\boldsymbol{a}$ obtenido por Fisher, y la T^2 de Hotelling puede obtenerse como transformación monótona de la R^2.

12.6 Demuéstrese que los valores propios de $W^{-1}B$ cuyos vectores propios asociados definen las direcciones discriminantes, son: no negativos.

12.7 Llamamos distancia en un espacio R^p a toda aplicación $d : R^p \times R^p \longrightarrow R$ verificando $\forall x, y \in R^p$ lo siguiente:

1. $d(x, y) > 0$ si $x \neq y$ y $d(x, y) = 0$ si $x = y$.
2. $d(x, y) = d(y, x)$.
3. $d(x, z) \leq d(x, y) + d(y, z)$ para todo $x, y, z \in R^p$.

Muéstrese que si Σ es de rango completo la expresión

$$d(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{y})$$

define una distancia (distancia de Mahalanobis[3])

12.8 (↑ 12.7) Compruébese que la distancia de Mahalanobis es invariante frente a transformaciones lineales de las variables.

12.9 Como primera aproximación al problema de discriminar entre dos grupos podríamos concebir la siguiente regla: Asignar $\boldsymbol{x}$ al grupo de cuyo vector de medias, $\boldsymbol{\mu}_1$ ó $\boldsymbol{\mu}_2$, esté más próximo en términos de distancia euclídea ordinaria: $d(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})'I(\boldsymbol{x} - \boldsymbol{y}) =$

[3]Hay alguna ambigüedad en la denominación, en cuanto que algunos autores llaman *distancia de Mahalanobis* a la expresión anterior con Σ reemplazada por su análogo muestral.

$\sum_{i=1}^{p}(x_i - y_i)^2$. Esta regla podría dar lugar a clasificar un caso en un grupo cuando en realidad es más plausible que proceda de otro, si las matrices de covarianzas en ambos grupos no fueran escalares (diagonales y con idénticos elementos a lo largo de la diagonal) e iguales. Ilústrese con un ejemplo de dos grupos con distribución normal bivariante y matrices de covarianzas no escalares.

12.10 (↑ 12.7) Consideremos la distancia de Mahalanobis definida entre observaciones procedentes de una misma población con matriz de covarianzas Σ. Muéstrese que siempre es posible hacer una transformación lineal de las variables originales de modo que las transformadas verifican:

1. Su matriz de covarianzas es I.
2. La distancia euclídea ordinaria entre ellas coincide con la distancia de Mahalanobis entre las originales.

12.11 (↑ 12.9) (↑ 12.7) Dado que el problema puesto de manifiesto en el Ejercicio 12.9 se presenta con matrices de covarianzas no escalares, podría pensarse en transformar el problema original en otro con matriz de covarianzas escalar y resolver éste último. Muéstrese que la regla que se obtiene es idéntica a la obtenida por Fisher, y da lugar a un discriminador lineal entre los dos grupos.

Capítulo 13

Arboles de regresión y clasificación

13.1. Arboles binarios

Llamamos *árbol binario* a un grafo formado por nodos y arcos verificando lo siguiente:

1. Hay un sólo nodo (la *raíz*) que no tiene padre.
2. Cada nodo distinto de la raíz tiene un único padre.
3. Cada nodo tiene exactamente dos o ningún hijo. En el caso de nodos sin hijos (o *nodos terminales*) hablamos también de "hojas".

Gráficamente representaremos los árboles con la raíz arriba, como en la Figura 13.1.

Podemos ver un árbol binario como una representación esquemática de un proceso de partición recursiva, en que en cada nodo no terminal tomamos la decisión de particionar una muestra de una cierta manera. Por ejemplo, el árbol de la Figura 13.1 designaría una sucesión de operaciones de partición recursiva de una muestra. Primeramente separamos, en r, una clase, que denominamos C. El resto se lleva al nodo n en el que tomamos una decisión ulterior, separándolo en las clases A y B.

En un árbol binario, cada nodo no terminal designa una decisión para particionar la fracción de muestra que llega a él en dos partes. Cada nodo terminal u hoja designa una de las clases a las que finalmente van a parar los elementos que dejamos caer desde la raíz.

Figura 13.1: Árbol binario con tres hojas, A, B, C y raíz r.

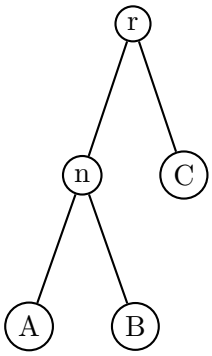
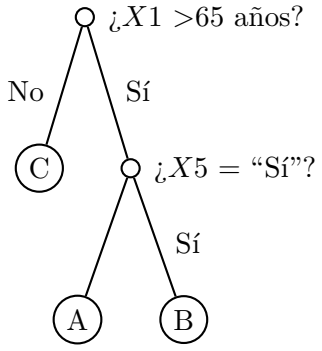


Figura 13.2: Árbol binario para clasificar pacientes en grupos de supervivencia homogénea



Ejemplo 13.1 Imaginemos una situación en que la muestra de entrenamiento consiste en N sujetos de cada uno de los cuales tenemos p variables, $x_1, \ldots, x_p$, recogiendo diferentes características clínicas. Tenemos también los valores que ha tomado una variable de interés —como por ejemplo, si han sobrevivido o no a una cierta operación—. Un árbol binario de clasificación describiría las operaciones de partición a realizar y el orden en que se efectúan las mismas, para acabar clasificando la muestra en clases relativamente homogéneas en lo que se refiere a la variable respuesta. Supongamos, por ejemplo, que X_1 es "edad" y X_5 es "Ha sufrido un infarto previo". Entonces, un árbol como el de la Figura 13.2 realizaría una clasificación de los sujetos en la muestra de entrenamiento en tres hojas A, B y C. Si resultara que el desglose de los casos que caen en las mismas es:

Hoja	Supervivientes	Fallecidos
A	40 %	60 %
B	20 %	80 %
C	80 %	20 %

estaríamos justificados en rotular la clase B como de alto riesgo, la C como de bajo riesgo y la A como de riesgo intermedio.

Un nuevo sujeto del que sólo conociéramos los valores de las X podría ser "dejado caer" desde la raíz y clasificado en uno de los grupos de riesgo de acuerdo con la hoja en que cayera.

Ejemplo 13.2 *(un árbol de regresión)* En el ejemplo anterior, la variable respuesta Y era cualitativa: podía tomar uno de dos estados, Podemos imaginar una respuesta Y continua en una situación similar: por ejemplo, el tiempo de supervivencia a partir del tiempo de una intervención quirúrgica.

En este caso, podríamos tener un árbol quizá exactamente igual al presentado en la Figura 13.2, pero su uso e interpretación sería diferente. Los casos que acabaran en las hojas A, B y C sería, si el árbol está bien construido, homogéneos en cuanto a sus valores de Y. El árbol serviría para, dados los valores de las X de un nuevo sujeto, asignarlo a una de las hojas y efectuar una predicción del valor de su Y: típicamente, la media aritmética de los valores en la hoja en que ha caído.

Este uso del árbol es completamente análogo al que se hace de una ecuación de regresión estimada. De hecho, si regresáramos las Y sobre tres columnas cada una de las cuales tuviera unos para los sujetos en una de las tres clases, A, B y C, las estimaciones de los parámetros β de la regresión coincidirían con las medias aritméticas de las clases. Nótese, sin embargo, que al construir el árbol *especificamos los "regresores"*, en cierto modo. Por ejemplo, la variable $X1$ (Edad) en el Ejemplo 13.1 se recodifica a "Sí" y No" (ó 0 y 1) a partir de un cierto umbral: podíamos haber tomado cualquier otro, y si tomamos ése es porque la división que logra es la "mejor", en un sentido que habremos de especificar más abajo.

Nótese también que, a diferencia de lo que ocurre en un modelo de regresión, las variables continuas se discretizan: la edad $X1$ queda reducida a dos grupos: mayores de 65 años o no. Un árbol sustituye una superficie de respuesta continua por una superficie de respuesta a escalones.

13.2. Construcción de árboles binarios

La metodología a seguir para construir un árbol binario resulta de conjugar varios elementos:

1. Un criterio para evaluar la ventaja derivada de la división de un nodo. ¿Qué nodo procede dividir en cada etapa?

2. Una especificación del espacio de búsqueda: ¿que tipos de particiones estamos dispuestos a considerar?

3. ¿Cómo estimar la tasa de mala clasificación (o varianza de predicción en el caso de árboles de regresión)?

4. Un criterio para decidir cuándo detener el crecimiento del árbol, o, como veremos, sobre la conveniencia de podar un árbol que ha crecido en exceso.

5. Un criterio para asignar un valor (o etiqueta de clase) a cada hoja.

Examinaremos cada cuestión por separado, describiendo a continuación el algoritmo de construcción de árboles.

13.2.1. Medidas de "impureza" de nodos y árboles.

Siguiendo la notación de **?** denotaremos la impureza del nodo t por $i(t)$.

En el caso de árboles de regresión, la $i(t)$ se toma habitualmente igual a la varianza muestral intranodo: nodos muy homogéneos son aquéllos con escasa varianza interna.

En el caso de árboles de clasificación, en que la respuesta es cualitativa, la impureza de un nodo debería estar en relación con las proporciones en que se presentan los elementos de las diferentes clases. Imaginemos que la variable respuesta cualitativa Y puede tomar J valores. Sea $p(j|t)$ la proporción de elementos de clase j en la muestra de entrenamiento que han ido a parar al nodo t. Claramente desearíamos que $i(t)$ fuera mínima si

$$\begin{aligned} p(\ell|t) &= 1 \\ p(j|t) &= 0 \qquad\qquad \forall j \neq \ell. \end{aligned}$$

Ello, en efecto, correspondería a un nodo "puro": todos los elementos que van a parar a él son de la clase ℓ. Por el contrario, desearíamos que la función $i(t)$ fuera máxima cuando

$$p(j|t) = J^{-1} \qquad\qquad \forall j,$$

pues un nodo en que todas las clases aparecen equi-representadas es en cierto sentido máximamente impuro.

Hay varias elecciones de $i(t)$ de uso común que verifican las propiedades anteriores, más otras deseables —como simetría en sus argumentos—. Tenemos así la función *entropía*

$$i(t) = -\sum_{i=1}^{J} p(j|t) \log_e p(j|t),$$

y el índice de Gini,

$$i(t) \quad = \quad \sum_{i \neq j} p(i|t)p(j|t).$$

En realidad, no nos interesa de ordinario la $i(t)$ de un nodo *per se*, sino en relación a la de sus posibles descendientes. Queremos valorar la ganancia en términos de impureza de una división del nodo t. Una posibilidad intuitivamente atractiva es

$$\Delta(s,t) \quad = \quad i(t) - p_L i(t_L) - p_R i(t_R),$$

en que la mejora en términos de impureza resultante de elegir la división s del nodo t se evalúa como la diferencia entre la impureza de dicho nodo y las de sus dos hijos, t_L y t_R, ponderadas por las respectivas proporciones p_L y p_R de elementos de la muestra que la división s hace ir a cada uno de ellos.

Una posibilidad adicional que evalúa la ganancia de la división s sin evaluar explícitamente una función de impureza en el padre y cada uno de los hijos, es:

$$\Delta(s,t) \quad = \quad \frac{p_L p_R}{4} \sum_j \left| p(j|t_L) - p(j|t_R) \right|^2 . \tag{13.1}$$

Observemos que la expresión (13.1) crece, por un lado, con la simetría de la división en cuanto al número de elementos de la muestra enviados a cada hijo, y por otro con la separación lograda entre las proporciones de cada clase en los dos hijos; lo que es intuitivamente atrayente.

La impureza total $I(T)$ de un árbol T se define como la suma ponderada de impurezas de sus hojas. Si $\tilde{T}$ es el conjunto formado por las hojas de T, entonces

$$I(T) \quad = \quad \sum_{t \in \tilde{T}} p(t)i(t) \tag{13.2}$$

Podríamos también evaluar la calidad de un árbol atendiendo a su *tasa de error*, $R(T)$. En el caso de un árbol de clasificación, típicamente es la probabilidad de obtener una mala clasificación al dejar caer un caso por él. Nótese que $R(T)$ es relativa al criterio de asignación de clase a los casos que caen en cada nodo terminal. Normalmente, el criterio es el de mayoría —se asigna el caso a la clase más representada en el nodo— o de máxima probabilidad *a posteriori*. Hablaremos también de la tasa de error en un nodo, $R(t)$, o en el subárbol T_t que crece desde el nodo t, $R(T_t)$. Un nodo terminal puede verse como un árbol degenerado con un sólo nodo terminal, y por lo tanto tendremos como notaciones equivalentes $R(\{t\})$ y $R(t)$.

En el caso de árboles de regresión, la tasa de error es alguna medida conveniente —normalmente, valor medio de suma de cuadrados intra-nodo de las desviaciones respecto a la media—.

13.2.2. Espacio de búsqueda

Hay una infinidad de formas posibles de efectuar divisiones en función de los valores que tomen las variables predictoras, X, y no podemos en general considerar todas ellas. Distinguiremos varias situaciones.

Variable X nominal. En este caso, X toma K valores distintos, como "rojo", "verde", "azul" o "Nacionalidad A", "Nacionalidad B", y Nacionalidad C", entre los que no cabe establecer un orden natural. Si tenemos que discriminar con ayuda de una variable nominal los elementos que van a los hijos izquierdo y derecho en la división del nodo t, podemos formar todos los subgrupos de los K valores que puede tomar X y enviar a la izquierda los casos con X tomando valores en un subgrupo y a la derecha los restantes.

> **Observación 13.1** Si $i(t)$ es estrictamente cóncava y estamos ante un árbol de clasificación en dos clases, etiquetadas $Y = 1$ e $Y = 0$, el cálculo se simplifica. Ordenemos los K valores que toma el predictor X en el nodo t de modo que
>
> $$p(1|X = x_1) \leq p(1|X = x_2) \leq \cdots \leq p(1|X = x_K).$$
>
> Se puede mostrar que no es preciso considerar todas las $2^{K-1} - 1$ posibilidades de agrupar las K categorías de X en dos grupos; basta considerar los $K - 1$ divisiones agrupando las categorías así
>
> $$\{x_1, \ldots, x_\ell\}\,\{x_{\ell+1}, \ldots, x_K\}\,,$$
>
> $(1 \leq \ell \leq K - 1)$ y enviando un grupo al hijo derecho del nodo t y el otro al hijo izquierdo. Véase **?**, pág. 218 ó **?**, pág. 101.

Variable X ordinal. En este caso, si la variable X toma n valores, se consideran como posibles cortes los $(n - 1)$ valores intermedios. En cada nodo nos formulamos una pregunta tal como: "¿Es $X_i < c$?", cuya respuesta afirmativa o negativa decidirá si el elemento que examinamos es enviado al hijo izquierdo o al hijo derecho del nodo en que estamos.

Variable X continua. Operaremos como con las variables ordinarias, si bien aquí será frecuente que el número de valores de corte a ensayar sea mucho mayor —si no hay repeticiones, como habitualmente acontecerá para una variable continua, el número de cortes a ensayar será de $N - 1$, siendo N el tamaño de la muestra de entrenamiento—.

> **Observación 13.2** En el caso de árboles de clasificación, el cálculo puede reducirse algo respecto de lo que sugiere el párrafo anterior. Si ordenamos los N elementos en un nodo t de acuerdo con el valor que que toma para ellos una variable continua X, podemos obtener hasta N valores diferentes: pero no necesitan ser considerados aquellos elementos flanqueados por otros de su misma clase, Véase **?**, pág. 237 y **?**.

Adicionalmente, al coste de un esfuerzo de cálculo superior, podemos formular en cada nodo una pregunta del tipo "¿Es $\boldsymbol{a}'\boldsymbol{X} < c$?", en que tanto $\boldsymbol{a}$ como c han de optimizarse para lograr divisiones con la máxima pureza en los nodos hijos. Divisiones así dan lugar a hiper-planos de separación que ya no han de ser paralelos a los ejes.

13.2.3. Estimación de la tasa de error

La elección de un árbol con preferencia a otro dependerá en general de sus respectivas $R(T)$. Se presenta el problema de estimarlas: según como lo hagamos, podríamos tener una imagen excesivamente optimista del ajuste del árbol a los datos, que nos desviaría notablemente de la construcción de un árbol óptimo; es útil por consiguiente prestar alguna atención al modo de estimar $R(T)$.

> **Observación 13.3** El problema no es muy diferente del que se presenta al evaluar la tasa de error en la clasificación de una función discriminante. Si lo hacemos reclasificando la muestra de entrenamiento, encontraremos, como vimos, una tasa de error sesgada por defecto.
>
> El problema se reproduce aquí, incluso agravado; porque, a igualdad de dimensionalidad de los datos, un árbol de clasificación tiene mucha más flexibilidad que un discriminante lineal para adaptarse a las peculiaridades de una muestra particular, y en consecuencia de dar una imagen excesivamente optimista al emplearlos para reclasificar dicha muestra.

Estimador por resustitución. El estimador más simple, pero también el potencialmente más sesgado a la baja, es el *estimador por resustitución*. Consiste simplemente en dejar caer por el árbol *la misma* muestra que ha servido para construirlo. Como se deduce de la Observación 13.3, tal estimador puede estar severamente sesgado a la baja, al permitir los árboles binarios una gran flexibilidad para adaptarse a una muestra dada.

No obstante, $\hat{R}(T)$ es de fácil y rápido cálculo, y puede ser útil para comparar árboles con igual o muy similar número de nodos.

Estimador por muestra de validación. La idea es similar a la del apartado anterior, pero lo que se deja caer ahora por el árbol es una muestra distinta a la de entrenamiento, formada por tanto por casos que no han sido vistos por el árbol y a los cuáles no se ha podido adaptar. Tenemos así un estimador $R^{ts}(T)$ que cabe suponer insesgado por lo menos aproximadamente, pero que tiene el inconveniente de forzarnos a reservar para su uso en validación una parte de la muestra, que de otro modo habríamos podido emplear en el entrenamiento.

Estimación por validación cruzada La idea de validación cruzada , tan presente en multitud de contextos, es de aplicación también aquí. Para estimar $R(T)$ parecería que podemos proceder reiteradamente como en el apartado anterior, dejando cada vez fuera de la muestra de entrenamiento (para validación) una fracción de k^{-1} del tamaño muestral total. Obtendríamos así k estimaciones $R^{(1)}(T), \ldots, R^{(k)}(T)$ y, promediándolas,

$$R^{cv}(T) \quad = \quad \frac{R^{(1)}(T) + \cdots + R^{(k)}(T)}{k}. \tag{13.3}$$

Obsérvese, sin embargo, que el árbol que hiciéramos crecer con cada una de las submuestras podría quizá ser distinto a los demás: la expresión anterior sólo tendría sentido tal cual está escrita en el (improbable) caso de que obtuviéramos exactamente el mismo árbol con las k submuestras empleadas.

No podemos, por ello, emplear validación cruzada para obtener una estimación de la tasa de error *asociada a un árbol concreto*. Si podremos hacerlo para seleccionar un árbol, del modo que se verá en 13.2.6.

Estimadores bootstrap. Se ha propuesto también hacer uso de estimadores basados en técnicas de *bootstrap*. Véase **?**, pág. 238.

13.2.4. Tasa de error penalizada

Para la selección de un árbol entre los muchos que podemos construir sobre una muestra, podemos pensar en el empleo de criterios análogos a la C_p de Mallows o AIC de Akaike. En el contexto actual, podríamos penalizar la tasa de error así:

$$R_\alpha(T) \quad = \quad \hat{R}(T) + \alpha|\tilde{T}|, \tag{13.4}$$

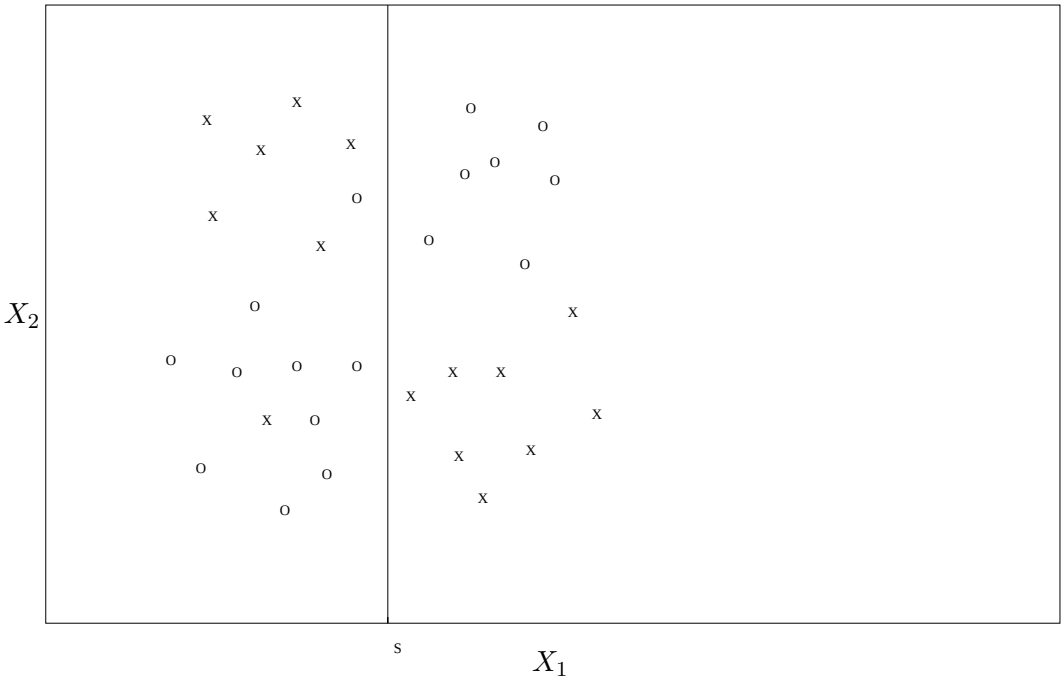
siendo $|\tilde{T}|$ el número de hojas del árbol T y α un parámetro de coste de cada hoja. La complejidad del árbol queda medida así por el número de hojas; la expresión (13.4) pondera tanto la bondad de ajuste del árbol (medida por $\hat{R}(T)$) como su complejidad.

No obstante, no tenemos idea de cuál haya de ser un valor adecuado de α. No tenemos tampoco claro que $|\tilde{T}|$ sea una medida adecuada de la complejidad: no es el número de parámetros, porque incluso en el caso más simple de un árbol de regresión, no nos limitamos a ajustar un parámetro (la media) en cada hoja. Hacemos más cosas: seleccionamos las variables con arreglo a las que particionamos, y los umbrales. El Ejemplo 13.2, pág. 127, ilustra ésto con claridad: dividir un nodo *no* es igual que reemplazar un regresor por otros dos.

13.2.5. Criterios de parada y/o poda

Una de las ideas más fecundas en la metodología propuesta por **?** es la de "mirar hacia adelante". Inicialmente se ensayaron estrategias consistentes

Figura 13.3: Una división en $X_1 = S$ es inútil por si misma, pero abre la vía a otras sumamente provechosas



en subdividir nodos (escogiendo en cada momento la división que produjera la máxima disminución de impureza $i(t)$) mientras un estimador adecuado de $R(T)$ disminuyera. Dado que en cada paso se examinan árboles con un número de nodos muy similar, basta a efectos de dictaminar la procedencia de una nueva división con estimar $R(T)$ por $\hat{R}(T)$.

Se observó, sin embargo, que esta estrategia daba resultados muy pobres y esto es debido a que, en ocasiones, subdivisiones que por sí mismas no serían justificables, abren el camino a otras muy provechosas. La Figura 13.3 lo ilustra en un caso artificialmente simple, con dos variables y dos clases. Puede verse, en efecto, que particionar el espacio a lo largo de $X_1 = S$ no logra prácticamente ninguna reducción de la impureza: ambas mitades tienen aproximadamente un 50 % de elementos 'O' y 'X'. No obstante, cada una de dichas mitades puede ahora ser subdividida en dos regiones prácticamente puras.

Esto sugiere que conviene construir árboles muy frondosos, porque no sabemos lo que hay "más allá" de la división de un nodo hasta que lo vemos. Si lo que se encuentra no justifica la frondosidad añadida al árbol siempre estamos a tiempo de podarlo. La cuestión clave no es por tanto *dónde parar*

el crecimiento del árbol, sino *cuánto podar* un árbol que deliberadamente hemos dejado crecer hasta tamaños mayores de lo concebiblemente necesario.

El procedimiento de poda propuesto en **?** es muy simple. Consideremos la oportunidad de podar la rama T_t que brota del nodo t en un cierto árbol. La tasa de error penalizada de dicho nodo y de la rama que brota de él, serían respectivamente:

$$\begin{aligned} R_\alpha(t) &= \hat{R}(t) + \alpha && (13.5)\\ R_\alpha(T_t) &= \hat{R}(T_t) + \alpha|\tilde{T}_t| && (13.6)\\ &= \sum_{s\in\tilde{T}_t} \hat{R}(s) + \alpha|\tilde{T}_t|. && (13.7) \end{aligned}$$

Es fácil ver que para $\alpha = 0$,

$$R_\alpha(t) = \hat{R}(t) > \hat{R}(T_t) = R_\alpha(T_t), \tag{13.8}$$

en tanto que para α lo suficientemente grande se verifica la desigualdad contraria, $R_\alpha(t) < R_\alpha(T_t)$. Por tanto habrá un valor de α, llamémosle $g(t,T)$, verificando $R_\alpha(t) = R_\alpha(T_t)$. Podemos obtener fácilmente este valor despejando α de la igualdad

$$\hat{R}(t) + \alpha \;=\; \hat{R}(T_t) + \alpha|\tilde{T}_t|,$$

lo que nos proporciona

$$g(t,T) \;=\; \frac{\hat{R}(t) - \hat{R}(T_t)}{|\tilde{T}_t| - 1}.$$

Un valor α igual a $g(t,T)$ hace que nos sintamos indiferentes entre la poda o no de la rama T_t. Valores superiores de α (= mayor coste de la complejidad) nos impulsarían a podar la rama, en tanto que valores menores nos impulsarían a conservarla.

La estrategia de poda propuesta por **?** es muy simple: para cada nodo no terminal (en que no ha lugar a podar nada) se evalúa $g(t,T)$, Se poda a continuación la rama T_{t_*} brotando del nodo t_* verificando $\alpha_1 \stackrel{\text{def}}{=} g(t_*,T) = \text{mín}_t\, g(t,T)$.

Tras la poda de la rama T_{t_*} obtenemos el árbol $T(\alpha_1)$; sobre el repetiremos el cálculo de los valores $g(t,T(\alpha_1))$ para todos los nodos no terminales, y podaremos la rama que brote del nodo con menor $g(t,T(\alpha_1))$ (valor que denominaremos α_2). El árbol así podado lo denominamos $T(\alpha_2)$. Proseguiremos del mismo modo hasta haber reducido el árbol inicial T al árbol degenerado que consiste sólo en el nodo raíz.

Se puede demostrar que con el modo de proceder anterior se obtiene una sucesión de árboles con la misma raíz, anidados. Es decir, una sucesión

$$T \succ T(\alpha_1) \succ T(\alpha_2) \succ \ldots \succ \{\text{raíz}\}.$$

13.2.6. El algoritmo de construcción de árboles

(por escribir)

13.3. Antecedentes y refinamientos

Se han propuesto metodologías alternativas a la descrita (CART). Por ejemplo, **?** propone un método llamado FIRM y **?** una simbiosis de construcción de árboles y análisis discriminante (que no da lugar a árboles binarios sino n-arios). Otra generalización se conoce como MARS (Multivariate Adaptive Regression Splines). Toma la idea de particionar recursivamente el espacio de las variables predictores, pero en lugar de ajustar una constante en cada hoja —al igual que un árbol de regresión como los descritos— ajusta *splines*. El resultado es una superficie sin discontinuidades, y con el grado de suavidad que se desee (fijando el orden de los *splines* en el valor que se desee). La referencia seminal es **?**. Una aproximación similar, orientada a la clasificación, es la seguida por **?**.

13.4. Bibliografía comentada

La monografía **?** continúa siendo una referencia básica. Fue el libro que otorgó carta de ciudadanía a métodos que habían sido propuestos previamente desde perspectivas menos generales. El Capítulo 4 de **?** es un resumen útil, desde el punto de vista de los problemas de clasificación. El libro **?** da una panorámica de lo que hay disponible en `S-Plus` *standard*; pueden utilizarse también las rutinas de **?**, que añaden alguna funcionalidad como particiones suplentes *(surrogate splitting)*. **?** dedica el Cap. 7 a árboles de clasificación, y proporciona bibliografía actualizada. Otros manuales que tratan sobre árboles de regresión y clasificación son **?** y **?**, que se refieren también a cuestiones no tratadas aquí (*boosting*, MARS, etc.). **?** en su Cap. 20 habla de árboles desde una perspectiva marcadamente más matemática.

Capítulo 14

Redes Neuronales Artificiales

14.1. Introducción

Los primeros intentos de construir una *red neuronal artificial (RNA)* buscaban replicar la estructura del cerebro de los animales superiores, tal y como se percibía en la época; el precedente más antiguo, **?**, se remonta a los años cuarenta.

Aunque la neurobiología ha sido de modo continuado una fuente de inspiración y una metáfora adecuada del trabajo en RNA, la investigación en este campo ha seguido un camino propio. Una descripción del curso entrelazado de ambos campos —neurobiología y RNA— y sus respectivas influencias puede verse en **?**, Cap. 2, y **?**, Cap. 1.

14.2. Neuronas biológicas y neuronas artificiales

14.2.1. Morfología y funcionamiento de una neurona humana

Ciñéndonos sólo a los aspectos esenciales, una neurona humana es una célula que consta de las siguientes partes: el *soma* o cuerpo celular del que emanan *dendritas* y el *axon*; unas y otro poseen terminaciones sinápticas con las que se unen a otras neuronas. El axon puede tener del orden de 10^3 terminaciones sinápticas. Un esquema simplificado puede verse en la Figura 14.1, tomada de **?**, pág. 6.

Una neurona recibe estímulos de otras neuronas a traves de las terminaciones sinápticas. A su vez, produce señales que a través del axon estimulan a otras neuronas. Hay del orden de 10^{11} neuronas en un cerebro humano, cada

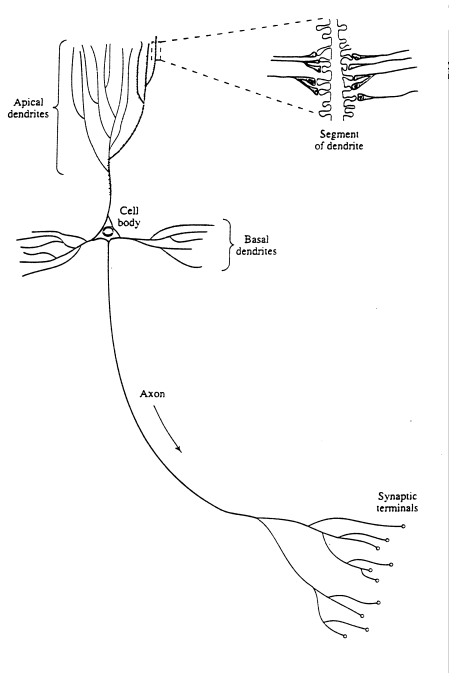


Figura 14.1: Esquema describiendo las partes principales de una neurona humana. Tomado de **?**, p. 8.

una con un elevado número de entradas y salidas sinápticas conectadas con otras neuronas, lo que da un sistema masivamente paralelo de complejidad casi inimaginable.

En el trabajo pionero **?** se suponía que cada neurona "computa" su salida o respuesta de modo muy simple: suma los inputs, quizá afectados de ponderaciones, y si la suma sobrepasa un cierto nivel crítico de excitación, "dispara", es decir, produce una salida en su axon. Se trataría así de un dispositivo de activación de tipo umbral: todo o nada, dependiendo de si se traspasa dicho umbral.

Hoy se sabe (cf. por ejemplo **?**, Sec. 2.2) que la naturaleza de las interacciones entre neuronas es más compleja de lo que la simple descripción anterior haría pensar. Dicha descripción, sin embargo, proporciona un punto de arranque e inspiración para el desarrollo de neuronas artificiales, como se describe a continuación.

14.2.2. Neuronas artificiales

La descripción anterior, transcrita a notación matemática, equivale a que una neurona toma todos sus entradas, las pondera mediante coeficientes

$w_1, \ldots, w_p$, y proporciona a la salida:

$$Y = \frac{1}{2} + \frac{1}{2}\text{sgn}\left(\sum_{i=1}^{p} w_i x_i + w_0\right), \tag{14.1}$$

en que "sgn" es la función definida por

$$\text{sgn}(u) = \begin{cases} +1 & \text{si } u > 0 \\ -1 & \text{en caso contrario.} \end{cases} \tag{14.2}$$

Podemos considerar neuronas que realizan un cómputo más general, relacionando las entradas con la salida de acuerdo con una expresión como

$$Y = f(\varphi(\boldsymbol{x}, \boldsymbol{w})). \tag{14.3}$$

En la expresión anterior, $\boldsymbol{x}$ es el vector de entradas o estímulos que recibe la neurona, y $\varphi()$ una función de excitación dependiente de los parámetros en $\boldsymbol{w}$; habitualmente, $\varphi(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{p}(w_i x_i + w_0)$, pero podría tomar cualquier otra forma. Por simplicidad notacional consideraremos la existencia de una componente x_0 de $\boldsymbol{x}$ con valor fijo igual a 1 (el "sesgo" u *offset* en la jerga del área, sin ninguna relación con la noción estadística de sesgo). Escribiremos entonces $\sum_{i=0}^{p} w_i x_i$ como función de excitación de la neurona, sin tener que recoger separadamente el coeficiente w_0.

La función $f()$ *activación* es habitualmente no lineal. Las siguientes son posibilidades utilizadas para $f()$:

Nombre	**Descripción**	**Valores**
Escalón (o signo)	$\text{sgn}(u)$	± 1
Heaviside (o umbral)	$\frac{1}{2} + \frac{1}{2}\text{sgn}(u)$	0 ó 1
Logística	$(1 + e^{-u})^{-1}$	(0,1)
Identidad	u	$(-\infty, +\infty)$

Cuadro 14.1: Funciones de activación $f(u)$ usuales

Tenemos así que una neurona artificial realiza el cómputo esquematizado en la Figura **??**.

Observación 14.1 Una neurona como la descrita en la Figura **??** con función de activación no lineal $\varphi(u) = sgn(u)$ fue propuesta por Rosenblatt con el nombre de *perceptrón*, con el propósito de aproximar una respuesta binaria.

Observación 14.2 Una neurona con la función de excitación lineal $f(\boldsymbol{x}) = \sum_{i=0}^{p} w_i x_i$ y con función de activación $\varphi(u) = u$ (identidad), realiza un cómputo análogo al de un modelo de regresión lineal.

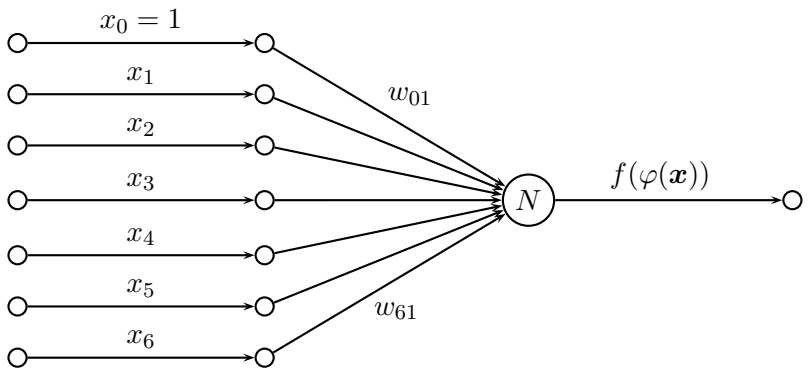


Figura 14.2: Esquema de una neurona artificial N. Recibe la entrada $\boldsymbol{x} = (x_0, \ldots, x_6)$ computando la función de excitación $\varphi(\boldsymbol{x}) = \sum_{i=0}^{6} w_{i1} x_i$ y entregado $f(\varphi(\boldsymbol{x}))$ a la salida.

Seleccionando la función de activación $f(u)$ de modo diferente, podríamos lograr que la neurona realizara el mismo cómputo que un modelo lineal generalizado. Por ejemplo, mediante $f(u) = (1 + e^{-u})^{-1}$ tendríamos un modelo de regresión logística. Si la salida deseada fuera un variable cualitativa, la neurona podría realizar el cómputo análogo a una función discriminante (lineal o no lineal, dependiendo de las funciones $f()$ y $\varphi()$ escogidas).

14.2.3. Redes neuronales artificiales (RNA)

A imagen de como acontece en el cerebro humano, podemos conectar varias neuronas entre sí para formar una RNA. Por ejemplo, una RNA con una única capa oculta de tres neuronas, una entrada $\boldsymbol{x} = (x_0, x_1, \ldots, x_6)$ y una salida $\boldsymbol{y} = (y_1, y_2)$ tendría una disposición como la de la Figura **??**.

14.3. Entrenamiento de una RNA

El *entrenamiento* o *aprendizaje* de una red neuronal es el proceso por el cual, mediante la presentación de ejemplos de parejas de vectores $(\boldsymbol{x}, \boldsymbol{d})$ (entradas y salidas observadas), se fijan los valores de los coeficientes (o *pesos*) w_{ij}.

Los pesos juegan un papel similar al de los parámetros en un modelo estadístico convencional, y el proceso de entrenamiento es equivalente al de estimación en los términos estadísticos habituales. Con más frecuencia que en la estimación estadística ordinaria, sin embargo, el entrenamiento se

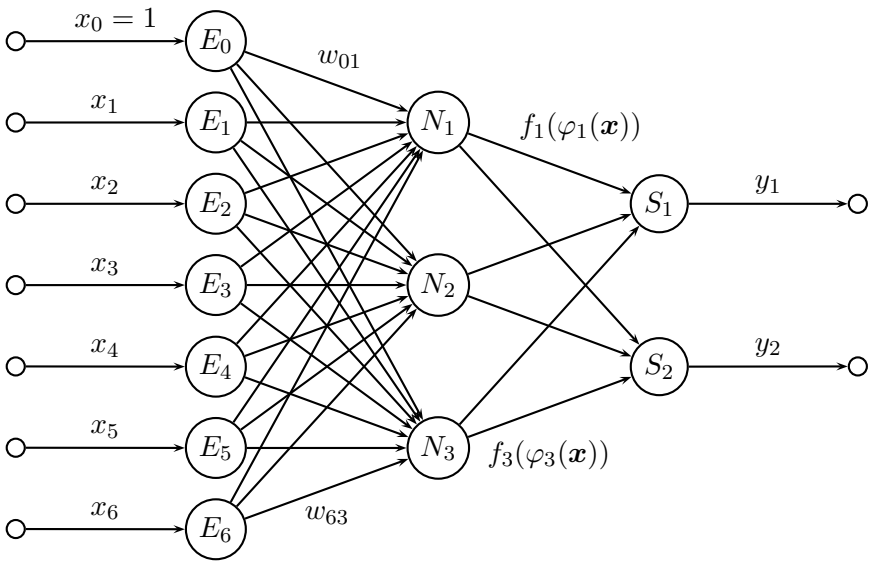


Figura 14.3: RNA con tres neuronas. Las unidades de entrada, E_0 a E_6, reparten el input $\boldsymbol{x} = (x_0, \ldots, x_6)$ a las tres neuronas que forman la capa oculta, N_j $(j = 1, 3)$. Cada una de estas neuronas computa $\varphi_j(\boldsymbol{x}) = \sum_{i=0}^{6} w_{ij} x_i$ y entrega $f_j(\varphi_j(\boldsymbol{x}))$ a cada unidad de salida. S_1 y S_2 suman sus inputs y producen $\boldsymbol{y} = (y_1, y_2)$.

lleva a cabo de forma adaptativa, presentando a la red instancias o ejemplos (pares $(\boldsymbol{x}, \boldsymbol{d})$) de uno en uno. Examinaremos primero un ejemplo con interés histórico —el del *perceptrón*— y el modo de entrenarlo, para luego considerar ejemplos más elaborados de redes y diferentes medios de entrenarlas.

14.3.1. Entrenamiento de un perceptrón

El perceptrón ha sido ya introducido en la Observación 14.1. Se trata de una red neuronal muy simple compuesta por una única neurona cuyo objetivo es distinguir entre objetos de dos clases, convencionalmente rotuladas como $+1$ y -1.

Consideremos el problema de su entrenamiento *en el caso simple de que los objetos de las dos clases sean linealmente separables;* es decir, supongamos que existe un vector de pesos $\boldsymbol{w}$ tal que $\boldsymbol{w}'\boldsymbol{x} > 0$ para todos los objetos de una clase y $\boldsymbol{w}'\boldsymbol{x} < 0$ para todos los de la otra. Cuando esto sucede, hay un algoritmo muy simple (Algoritmo **??**) con convergencia asegurada, que produce un vector $\boldsymbol{w}$ separando correctamente los casos.

La idea es muy sencilla: se presentan los casos $(\boldsymbol{x}, g)$ al perceptrón y se computa $\boldsymbol{w}'\boldsymbol{x}$. Si el resultado es "correcto" ($\boldsymbol{w}'\boldsymbol{x} > 0$ para objetos en el grupo $\mathcal{G}_1$ y $\boldsymbol{w}'\boldsymbol{x} \leq 0$ para objetos en el grupo $\mathcal{G}_2$; la asignación de las etiquetas -1 y $+1$ a los grupos $\mathcal{G}_1$ y $\mathcal{G}_2$ es arbitraria), los pesos se dejan

Algoritmo 2 – Entrenamiento de perceptrón por corrección de error.

1: $N \leftarrow$ Número de ejemplos en la muestra de entrenamiento
2: $\boldsymbol{w} \leftarrow \mathbf{0}$; $\eta \leftarrow$ Parámetro aprendizaje;
3: **repeat**
4: $\quad E \leftarrow 0$
5: $\quad$ **for** $i = 1$ to N **do**
6: $\quad\quad$ **if** $(\boldsymbol{w}'\boldsymbol{x}_i > 0) \wedge (\boldsymbol{x}_i \in \mathcal{G}_2)$ **then**
7: $\quad\quad\quad \boldsymbol{w} \leftarrow \boldsymbol{w} - \eta\boldsymbol{x}_i$
8: $\quad\quad\quad E \leftarrow E + 1$
9: $\quad\quad$ **else if** $(\boldsymbol{w}'\boldsymbol{x}_i \leq 0) \wedge (\boldsymbol{x}_i \in \mathcal{G}_1)$ **then**
10: $\quad\quad\quad \boldsymbol{w} \leftarrow \boldsymbol{w} + \eta\boldsymbol{x}_i$
11: $\quad\quad\quad E \leftarrow E + 1$
12: $\quad\quad$ **end if**
13: $\quad$ **end for**
14: **until** $(E = 0)$
15: $\boldsymbol{w}_{\text{final}} \leftarrow \boldsymbol{w}$

en los valores preexistentes en la iteración anterior. No es preciso ningún cambio.

Si, por el contrario, se produce un error de clasificación, se modifican los pesos tal como recogen las asignaciones 7 y 10 en el algoritmo. El parámetro η o *parámetro de aprendizaje* puede tomar cualquier valor, con tal de que sea positivo. Diferentes valores afectan sólo a la velocidad a la que converge el algoritmo.

> **Observación 14.3** El parámetro η no necesariamente ha de permanecer constante. Frecuentemente se reemplaza por una sucesión de parámetros $\eta(n)$, con n contando el número de "pasadas" sobre los datos *(epochs)*, de modo que $\eta(n)$ disminuye en valor absoluto conforme el aprendizaje avanza.

Cuando se comete un error que requiere la modificación del vector de pesos $\boldsymbol{w}$, se incrementa la variable contadora de errores, E. El algoritmo finaliza cuando en una pasada sobre todos los N casos no se produce ningún error, circunstancia que se comprueba en la línea 17; esto puede requerir varias pasadas sobre la muestra de entrenamiento. Obsérvese que el algoritmo se presta al aprendizaje *on line*, en que los ejemplos se muestran a medida que van apareciendo.

La demostración de la convergencia es simple y puede consultarse en **?**, p. 100 ó **?**, p. 139, por ejemplo. Sin entrar a detallarla aquí, es fácil ver que la actualización que se hace en las líneas 7 y 10 del Algoritmo **??** es "lógica". Si el nuevo caso es correctamente clasificado por el perceptrón, $\boldsymbol{w}$ no se toca. Si $\boldsymbol{w}'\boldsymbol{x}_i > 0$ y hubiéramos deseado que $\boldsymbol{w}'\boldsymbol{x}_i \leq 0$ (línea 6), la actualización

que se realiza es:

$$\boldsymbol{w}_* \leftarrow \boldsymbol{w} - \eta \boldsymbol{x}_i$$

con lo que

$$\begin{aligned} \boldsymbol{w}_*{}' \boldsymbol{x}_i &= \boldsymbol{w}' \boldsymbol{x}_i - \eta \, ||\boldsymbol{x}_i||^2 \\ &\leq \boldsymbol{w}' \boldsymbol{x}_i; \end{aligned}$$

es decir, nos movemos en la dirección deseada ($\boldsymbol{w}_*{}' \boldsymbol{x}_i$ se hace "menos positivo"), a tanta mayor velocidad cuanto mayor sea η. (Obsérvese que una actualización de este género puede introducir errores en ejemplos previamente bien clasificados, por lo que de ordinario serán necesarias varias pasadas sobre los datos.) De modo análogo sucede con la corrección en la línea 10 del algoritmo, cuando $\boldsymbol{w}' \boldsymbol{x}_i \leq 0$ indebidamente en la línea 9.

En definitiva, el algoritmo consiste en ir perturbando secuencialmente un hiperplano de modo que consigamos separar todos los casos. Claramente, sólo podremos tener éxito cuando los casos sean linealmente separables. Cuando esto ocurre, el algoritmo suministra un método de discriminación alternativo a los estudiados en el Capítulo 12 para el caso de dos grupos.

14.3.2. El método de corrección de error.

El procedimiento anterior puede ser generalizado al caso en que la respuesta no es binaria. Dicha generalización puede por otra parte verse como un caso particular del método de aproximación estocástica de Robbins-Monro (véase **?** y **?**, pág. 46–48) que describimos a continuación.

Teorema 14.1 *Consideremos dos variables correladas, g y θ verificando que $f(\theta) = E[g|\theta]$ (es decir, $f()$ es una función de regresión de $g()$ sobre θ). Supongamos que*

$$E[(g(\theta) - f(\theta))^2] < \infty \tag{14.4}$$

y, sin pérdida de generalidad, que $f(\theta)$ es monónota decreciente. Sea una sucesión de números reales η_n verificando:

$$\lim_{n\to\infty} \eta_n = 0 \tag{14.5}$$

$$\sum_{n=1}^{\infty} \eta_n = \infty \tag{14.6}$$

$$\sum_{n=1}^{\infty} \eta_n^2 < \infty; \tag{14.7}$$

entonces, si podemos evaluar la función $g(\theta)$ en una sucesión de valores $\theta_1, \ldots, \theta_n, \ldots$ generados así:

$$\theta_{n+1} = \theta_n + \eta_n g(\theta_n), \tag{14.8}$$

se tiene que θ_n converge con probabilidad 1 a θ_0, una raíz de $f(\theta) = E[g|\theta] = 0$.

El teorema anterior sugiere un procedimiento para entrenar secuencialmente una red neuronal. Estamos interesados en optimizar una función de error $\mathcal{E}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{w})$ continua y suficientemente derivable, como por ejemplo

$$\mathcal{E}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{m} (y_i^{(n)} - f_i(\boldsymbol{x^{(n)}}, \boldsymbol{w}))^2 \tag{14.9}$$

Las condiciones de primer orden estipulan

$$\frac{\partial}{\partial \boldsymbol{w}} \mathcal{E}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{w}) = \sum_{n=1}^{N} \left[\sum_{i=1}^{m} (y_i^{(n)} - f_i(\boldsymbol{x^{(n)}}, \boldsymbol{w})) \frac{\partial}{\partial \boldsymbol{w}} f_i(\boldsymbol{x^{(n)}}, \boldsymbol{w}) \right] = \mathbf{0} \tag{14.10}$$

Es equivalente resolver la ecuación anterior o:

$$\frac{1}{N} \sum_{n=1}^{N} \left[\sum_{i=1}^{m} (y_i^{(n)} - f_i(\boldsymbol{x^{(n)}}, \boldsymbol{w})) \frac{\partial}{\partial \boldsymbol{w}} f_i(\boldsymbol{x^{(n)}}, \boldsymbol{w}) \right] = \mathbf{0}, \tag{14.11}$$

y para N grande, el lado izquierdo de la igualdad anterior es aproximadamente igual al valor medio

$$E \left(\sum_{i=1}^{m} (y_i - f_i(\boldsymbol{x}, \boldsymbol{w})) \frac{\partial}{\partial \boldsymbol{w}} f_i(\boldsymbol{x}, \boldsymbol{w}) \right); \tag{14.12}$$

si identificamos la función cuyo valor medio se computa en (**??**) con $f(\theta)$ y θ con $\boldsymbol{w}$, vemos que es de aplicación el Teorema **??**. Podemos pensar pues en aplicar el procedimiento de Robbins-Monro, que converge casi seguramente a una raíz de (**??**) y por tanto, aproximadamente, a una raíz de (**??**). Esto conduce a:

$$\boldsymbol{w^{(n+1)}} = \boldsymbol{w^{(n)}} + \eta \sum_{i=1}^{m} \left[y_i^{(n)} - f_i(\boldsymbol{x^{(n+1)}}, \boldsymbol{w^{(n)}}) \right] \frac{\partial}{\partial \boldsymbol{w}} f_i(\boldsymbol{x^{(n+1)}}, \boldsymbol{w^{(n)}}) \tag{14.13}$$

Si consideramos el caso de una red neuronal similar al perceptrón de la Sección 14.1 pero con activación lineal y respuesta continua, vemos que la expresión (**??**) se particulariza a:

$$\begin{aligned} \boldsymbol{w^{(n+1)}} &= \boldsymbol{w^{(n)}} + \eta \left(y_i^{(n)} - f(\boldsymbol{x^{(n+1)}}, \boldsymbol{w^{(n)}}) \right) \boldsymbol{x^{(n)}} && (14.14) \\ &= \boldsymbol{w^{(n)}} + \eta e^{(n+1)} \boldsymbol{x^{(n)}} && (14.15) \end{aligned}$$

en que $e^{(n+1)}$ designa el error de ajuste de la $n+1$ observación con los pesos existentes tras procesar la n-ésima observación y $\boldsymbol{x^{(n)}}$ es el vector de derivadas parcial de la activación respecto del vector de pesos $\boldsymbol{w}$. La fórmula de corrección de error (**??**) generaliza la que se presentó en la Sección 14.1; η ocupa el lugar de η.

Si la activación no fuera lineal, la expresión (**??**) se convertiría en

$$\boldsymbol{w}^{(n+1)} = \boldsymbol{w}^{(n)} + \eta e^{(n+1)} f'(a^{(n+1)})\boldsymbol{x}^{(n)} \tag{14.16}$$

en que $a^{(n+1)} = (\boldsymbol{w}^{(n)})'\boldsymbol{x}^{(n+1)}$ es la *excitación* de la neurona. Denominaremos *gradiente local* de la neurona a:

$$\delta^{(n+1)} \stackrel{\text{def}}{=} \frac{\partial \mathcal{E}^{(n+1)}}{\partial a^{(n+1)}} \tag{14.17}$$

$$= e^{(n+1)} f'(a^{(n+1)}). \tag{14.18}$$

Con esta notación, (**??**) se reescribe así:

$$\boldsymbol{w}^{(n+1)} = \boldsymbol{w}^{(n)} + \eta \delta^{(n+1)} \boldsymbol{x}^{(n)}; \tag{14.19}$$

en redes con más de una neurona, utilizaremos $\delta_k^{(n+1)}$ para designar el gradiente local de la neurona k-ésima.

Observación 14.4 Si observamos la última expresión, veremos que se trata de simplemente de aplicar un método gradiente observación a observación. En lugar de calcular las derivadas de la función objetivo haciendo uso de toda la muestra y llevar a cabo una optimización por el método del gradiente ordinario, tomamos las derivadas de la contribución a la función objetivo de *cada observación*. Como es lógico, debemos entonces ir amortiguando las contribuciones sucesivas, de modo que el influjo de la observación $n+1$ sobre el vector de pesos calculado con ayuda de las n precedentes, sea convenientemente pequeño: esta es la función del coeficiente de aprendizaje η.

Observación 14.5 Observemos también que la regla de actualización es muy sencilla porque sabemos lo que deseamos obtener, $y^{(n)}$, y lo que obtenemos, $f(a^{(n)})$; podemos "responsabilizar" del error a los pesos de la única neurona que interviene. La situación se complica cuando hay más de una neurona, quizá en cascada, en que no es obvio qué pesos hay que modificar para reducir la discrepancia entre lo computado y lo deseado. Sucede, sin embargo, que hay un algoritmo que permite hacer esta tarea no trivial de modo eficaz: es el algoritmo de *back-propagation* de que se ocupa la siguiente Sección.

14.3.3. El algoritmo de propagación hacia atrás

El algoritmo de propagación hacia atrás o *back-propagation* es, en esencia, una generalización a redes con más de una neurona del algoritmo de corrección de error presentado en la sección anterior. fue popularizado por **?** aunque la idea parece preexistente (ver **?**, p. 141).

La Sección anterior, en particular la ecuación (**??**), muestra el modo de actualizar los pesos a la entrada de una neurona en la primera capa cuando

se presenta el caso $\boldsymbol{x}^{(n)}$: basta multiplicar el gradiente local de la neurona por $\boldsymbol{x}^{(n)}$ y un parámetro de aprendizaje η.

Exactamente la misma regla es de aplicación a una neurona k en una capa intermedia, con la salvedad de que lo que se presenta a la entrada de la misma ya no es $\boldsymbol{x}^{(n)}$ sino el vector $\boldsymbol{z}^{(n)}$ de salidas de todas las neuronas en la capa precedente conectadas directamente a la k. El único problema, pues, es calcular el gradiente local para una tal neurona.

Puesto que podemos calcular δ_k para una neurona en la última capa, porque podemos hacer uso de (**??**) en que $e^{(n+1)}$ y $a^{(n+1)}$ son ambos calculables, haciendo uso de la regla de la cadena:

$$\delta_j = \frac{\partial \mathcal{E}^{(n+1)}}{\partial a_j^{(n+1)}} = \sum_k \frac{\partial \mathcal{E}^{(n+1)}}{\partial a_k^{(n+1)}} \frac{\partial a_k^{(n+1)}}{\partial a_j^{(n+1)}} = \sum_k \delta_k f'(a_j) w_{kj}, \tag{14.20}$$

en que la suma se toma sobre todas las neuronas k que reciben como entrada la salida de la neurona j. Efectivamente: la excitación de la neurona k depende linealmente (a traves del peso w_{kj}) de la salida z_j de la neurona j, y dicha salida depende de a_j a través de la función de activación f.

Tenemos pues un método simple que permite calcular las derivadas de la función de error respecto de las activaciones (y respecto de los pesos en consecuencia), para utilizarlas en algoritmo de tipo gradiente.

Algoritmo 3 – Entrenamiento de una RNA por *back-propagation.*

1: $N \leftarrow$ Número de ejemplos en la muestra de entrenamiento
2: $\eta \leftarrow$ Parámetro aprendizaje ; $\boldsymbol{w} \leftarrow \mathbf{0}$
3: $c \leftarrow$ Número de capas ; $S \leftarrow$ Número de épocas
4: **for** $s = 1$ to S **do**
5: **for** $n = 1$ to N **do**
6: Presentar el caso $x^{(n)}$ y calcular todas las activaciones a_i.
7: Evaluar $\delta_k^{(n)}$ para todas las neuronas conectadas a la salida.
8: **for** $\ell \in \{c-1, \ldots, 1\}$ **do**
9: **for** $j \in \{\text{Capa } \ell\}$ **do**
10: $\delta_j^{(n)} \leftarrow f'(a_j) \sum_k w_{kj} \delta_k^{(n)}$ $k \in$ Capa $(\ell + 1)$
11: $\delta_i^{(n)} \leftarrow \partial \mathcal{E}^{(n)} / \partial w_{ji} \leftarrow \delta_j^{(n)} z_i$ $z_i =$ Salida neurona i
12: **end for**
13: **end for**
14: $\boldsymbol{\nabla}(\boldsymbol{\mathcal{E}}^{(n)}) \leftarrow [\boldsymbol{\partial \mathcal{E}}^{(n)} / \boldsymbol{\partial w}]$
15: Actualizar los pesos mediante $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \boldsymbol{\nabla}(\boldsymbol{\mathcal{E}}^{(n)})$
16: **end for**
17: **end for**
18: Devolver solucion en $\boldsymbol{w}$.

14.4. Mapas auto-organizados (SOM)

Los mapas auto-organizados (*self-organizing maps, SOM* son un tipo de redes neuronales directamente inspiradas como los perceptrones en lo que parece ser un modo de funcionar del cerebro. Se aprecia en el mismo una organización espacial: las neuronas tienden a estimular a, y ser estimuladas por, aquéllas que les quedan más próximas, lo que produce que se especialicen en una función grupos de neuronas próximas.

? propuso un tipo de red neuronal artificial que imita dicho comportamiento. Básicamente opera así:

1. Se adopta para las neuronas una disposición espacial predeterminada: típicamente se disponen en filas y columnas. A cada neurona se le asigna un vector de pesos $\boldsymbol{w_{ij}}$ (los dos índices hacen referencia a la fila y columna en que esta ubicada la neurona).

2. Se inicializan los vectores $\boldsymbol{w_{ij}}$ de cualquier modo conveniente.

3. Se presenta a la red cada uno de las observaciones $\boldsymbol{x_k}$ de la muestra de entrenamiento $\{\boldsymbol{x_k}\}$, $k = 1, \ldots, n$.

4. Para cada neurona y cada observación en la muestra de entrenamiento se computa $R_{ij,k} = ||\boldsymbol{x_k} - \boldsymbol{w_{ij}}||^2$. Si

$$(i_{opt}, j_{opt}) = \arg\min_{i,j} R_{ij,k}$$

   se dice que la neurona en la posición (i_{opt}, j_{opt}) "gana" la competición. Entonces, su vector de pesos (y, aunque en menor medida, *los de todas las neuronas vecinas)*, se alteran en orden a realzar su ventaja competitiva al responder a la observación $\boldsymbol{x_k}$.

La descripción anterior, para hacerse más precisa, requiere especificar como es alteran los vectores de las neuronas "triunfantes" y sus vecinas, y quienes consideramos vecinas.

Respecto de la última cuestión, debemos definir en la red una distancia entre neuronas. Si las tenemos dispuestas en filas y comunas podríamos recurrir a una distancia entre las neuronas (i, j) y (k, l) como:

$$d^2_{ij,kl} = |i - k|^2 + |j - l|^2; \tag{14.21}$$

las neuronas vecinas de la (i, j) serían aquéllas (k, l) verificando $d^2_{ij,kl} < d$ para un cierto umbral d que debemos determinar. Este umbral no necesita ser fijo durante toda la duración del entrenamiento de la red, sino que, como veremos, ira por lo general disminuyendo.

Por lo que hace a la modificación de pesos de la neurona triunfante (i, j) y sus vecinas, la haremos del modo que sigue. Definamos $h_{ij,kl}$ como

una función decreciente de $d^2_{ij,kl}$. Entonces, cuando la neurona (i, j) triunfa al presentarle la observación $\boldsymbol{x}^{(n+1)}$, modificamos los vectores de pesos de todas las demás así:

$$\boldsymbol{w}_{kl}^{(n+1)} = \boldsymbol{w}_{kl}^{(n)} + \eta h_{ij,kl}(\boldsymbol{x}^{(n+1)} - \boldsymbol{w}_{kl}^{(n)}). \tag{14.22}$$

En la expresión anterior, η es un parámetro de aprendizaje, típicamente mucho menor que 1. La actualización de $\boldsymbol{w}_{kl}^{(n)}$ tiene lugar sumándole una fraccióon de su discrepancia con la observación $\boldsymbol{x}^{(n+1)}$, con lo que el vector actualizado está más cerca de ésta. Además de η, el parámetro $h_{ij,kl}$ hace que la actualización sea más intensa cuanto más cerca está la neurona $k, l)$ de la vencedora (i, j) (puesto que $h_{ij,kl}$ decrece con $d^2_{ij,kl}$).

La regla de entrenamiento (**??**) garantiza que neuronas próximas tendrán vectores de pesos parecidos.

14.5. Maquinas de vectores soporte (SVM)

Por escribir

Capítulo 15

Análisis de agrupamientos

15.1. Introducción

Consideramos un colectivo de N objetos, el i-ésimo de los cuales viene descrito por un vector $\boldsymbol{x_i}$. La información de partida es pues, como de costumbre, una tabla X de dimensiones $N \times p$. En principio, las componentes de dicho vector pueden ser reales, cualitativas o cualitativas ordenadas, e incluso cualquier combinación de dichos tipos.

El objetivo es, sobre la base de los vectores observados, agruparlos en k grupos, de tal modo que los que se incluyen en cada grupo tengan más parecido entre sí que con los de otros grupos.

Naturalmente, el problema así formulado es muy vago y requiere formalización adicional para poder ser abordado de manera algorítmica. Hemos de precisar qué significa "parecerse" dos objetos —lo que nos llevará a definir nociones de similaridad (o alternativamente disimilaridad) entre objetos: esta cuestión se aborda en la Sección **??**. Adicionalmente, dado que en el proceso de examinar agrupamientos habremos de considerar la posibilidad de unir o separar grupos ya formados, necesitaremos extender las nociones de similaridad o disimilaridad anteriores a grupos, lo que haremos en la Sección **??**. Finalmente, en la Sección **??** examinaremos las estrategias de construcción de grupos.

15.2. Medidas de similaridad y disimilaridad entre objetos

En lo que sigue se consideran diferentes medidas de similaridad o disimilaridad, adecuadas a situaciones diversas. En ocasiones resulta más natural pensar en términos de similaridad, en otras en términos de disimilaridad.

15.2.1. Variables reales

Consideremos en primer lugar el caso en que $\boldsymbol{x_i}$ está integramente compuesto por variables reales. La definición más inmediata de disimilaridad entre $\boldsymbol{x_i}$ y $\boldsymbol{x_j}$ vendría proporcionada por la distancia euclídea ordinaria entre ambos, vistos como puntos en R^p:

$$d^2(i,j) = ||\boldsymbol{x_i} - \boldsymbol{x_j}||^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2. \tag{15.1}$$

Obsérvese que esta noción de disimilaridad es dependiente de las escalas de medida: un cambio de unidades de medida en alguna o algunas de las variables altera las distancias entre objetos. Puede recurrirse a normalizar las variables antes de calcular la distancia euclídea entre objetos, o, lo que es equivalente, a calcular una distancia euclídea generalizada así:

$$d_D^2(i,j) = ||\boldsymbol{x_i} - \boldsymbol{x_j}||_D^2 = (\boldsymbol{x_i} - \boldsymbol{x_j})'D\,(\boldsymbol{x_i} - \boldsymbol{x_j}) \tag{15.2}$$

en que D es una matriz diagonal cuyo elemento k,k contiene el inverso de la norma (euclídea) de la k-ésima columna de X.

Si las p variables consideradas tienen correlación entre ellos, un refinamiento inmediato de la idea anterior consistiría en considera la distancia de Mahalanobis,

$$d_\Sigma^2(i,j) = ||\boldsymbol{x_i} - \boldsymbol{x_j}||_\Sigma^2 = (\boldsymbol{x_i} - \boldsymbol{x_j})'\,\Sigma^{-1}\,(\boldsymbol{x_i} - \boldsymbol{x_j}), \tag{15.3}$$

con Σ igual a la matriz de covarianzas de las p variables (si fuera conocida) o una estimación de ella en el caso habitual de que no lo sea.

Una vía diferente de generalización de la distancia euclídea ordinaria deriva de observar que $d(i,j)$ es realmente un caso particular, con $m = 2$, de la definición más general:

$$d_m(i,j) = \left(\sum_{k=1}^{p}|x_{ik} - x_{jk}|^m\right)^{1/m}. \tag{15.4}$$

Además de identificarse con la distancia auclídea ordinaria cuando $m = 2$, la expresión anterior da lugar a otras distancias de interés. Cuando $m = 1$ tenemos la distancia "bloque de casas" o "Manhattan". Cuando $m \to \infty$,

Cuadro 15.1: Tabulación cruzada de valores de p variables dicotómicas en $\boldsymbol{x_i}$, $\boldsymbol{x_j}$.

	0	**1**
0	a	b
1	c	d

tenemos que $d_m(i,j) \to \sup_{1\leq k\leq p} |x_{ik} - x_{jk}|$, y de entre todas las discrepancias entre los objetos i, j, sólo la mayor se toma en consideración. Cualquier valor $0 < m \leq \infty$ puede utilizarse, dando lugar a la *distancia de Minkowskye* parámetro m.

15.2.2. Variables cualitativas nominales

Consideremos el caso, más simple, de variables cualitativas dicotómicas, pudiendo tomar únicamente dos valores que convencionalmente designaremos por 0 y 1. Podríamos hacer uso con estas variables de cualquiera de las definiciones en el apartado precedente, pero con frecuencia tiene sentido hacer uso de definiciones alternativas.

Cuando los vectores $\boldsymbol{x_i}$ y $\boldsymbol{x_j}$ describiendo a los sujetos i, j, están compuestos en su integridad por variables dicotómicas, podemos construir una tabla de contingencia como la recogida en el Cuadro **??**. Vemos que, por ejemplo, para a variables hubo una concidencia en los valores que toman en $\boldsymbol{x_i}$ y $\boldsymbol{x_j}$, siendo ambas 0. Para d variables se verificó una coincidencia en el valor 1, y para $b + c$ variables hubo una discrepancia. (Obviamente, $a + b + c + d = p$ si todas las variables han sido registradas, es decir, no hay valores faltantes.)

A partir de los números tabulados en las cuatro casillas del Cuadro **??** podemos definir similaridad de muy diversas formas. Podemos por ejemplo considerar

$$s(i,j) = \frac{a+d}{a+b+c+d} \tag{15.5}$$

$$s(i,j) = \frac{2d}{a+b+c+d} \tag{15.6}$$

$$s(i,j) = \frac{d}{a+b+c+d}. \tag{15.7}$$

15.3. Medidas de similaridad y disimilaridad entre grupos

No basta definir similaridad o disimilaridad entre objetos. En algunos algoritmos para la obtención de agrupamientos se requiere en algunas fases decidir qué dos grupos ya formados se amalgaman, por ser los más similares. Es preciso por tanto extender la noción de similaridad (o dismilaridad) entre objetos de manera que proporciona una noción homóloga para grupos. Son muchas las posibilidades, entre las que citaremos tres.

Ligadura simple

Cuando utilizamos *ligadura simple*(single linkage) definimos como disimilaridad entre dos grupos la disimilaridad entre los dos objetos, uno en cada grupo, menos disimilares entre sí. Todo lo que se requiere para que dos grupos estén próximos es una pareja de puntos, uno en cada grupo, próximos.

Ligadura completa

La ligadura completa *ligadura completa*(complete linkage) es el criterio diametralmwente opuesto. Definimos como disimilaridad entre dos grupos la disimilaridad entre los dos objetos, uno en cada grupo, *más* disimilares entre sí. Para que dos grupos estén próximos, es preciso que los representantes de ambos más disimilares estén próximos —lo que supone que *todos* los objetos de un grupo han de estar en la vecindad de *todos* los del otro.

15.4. Estrategias de construcción de grupos

15.4.1. Procedimientos jerárquicos

Estrategias aglomerativas o divisivas

Examinaremos una estrategia aglomerativa; su homóloga divisiva es similar con los cambios obvios.

Inicialmente, en la etapa $t = 0$ del proceso de agrupamiento, todos los N objetos a agrupar se consideran separados. Los designaremos $O_1, \ldots, O_N$. A lo largo del proceso de aglomerado, los objetos se irán integrando en grupos. Emplearemos la notación $G_k = \{O_{i_1}, \ldots, O_{i_k}\}$ para indicar el grupo G_k contiene los objetos $O_{i_1}, \ldots, O_{i_k}$.

Comenzamos computando la matriz de disimilaridad entre todos los objetos:

	O_1	O_2	O_3	...	O_N
O_1	–	d_{12}	d_{13}	...	d_{1N}
O_2		–	d_{23}	...	d_{2N}
O_3			–	...	d_{3N}
⋮					
O_N					–

Recorreremos dicha matriz en busca de la disimilaridad d_{ij} menor. Supongamos que es la que corresponde a la pareja formada por O_2 y O_3. Tomaremos nota de dicha distancia y amalgamaremos ambos puntos para formar el grupo $G_1 = \{O_2, O_3\}$. A continuación eliminaremos las distancias en la fila y columna correspondientes a O_2 y O_3 y añadiremos una fila y columna correspondientes al grupo recién formado:

	O_1	O_2	O_3	...	O_N	G_1
O_1	–	–	–	...	d_{1N}	d_{1,G_1}
O_2		–	–	...	–	–
O_3			–	...	–	–
⋮						
O_N					–	d_{N,G_1}
G_1						–

Obsérvese que han desaparecido de la matriz de disimilaridades todas aquéllas que involucraban directamente a los objetos $=_2$ y O_3, y ha aparecido en cambio una nueva columna con las disimilaridades entre el grupo G_1 —que engloba a los dos objetos citados— y todos los demás. Las distancias en la nueva columna lo son de un grupo a objetos, y se calculan, por ejemplo, de acuerdo con uno de los criterios relacionados en la Sección **??**.

La nueva matriz de disimilaridades es de nuevo rastreada en busca de la menor. Si ésta corresponde a dos objetos, se amalgamarán en un nuevo grupo. Si corresponde a una distancia entre un objeto aislado y un grupo ya formado, se amalgamará el objeto a dicho grupo. En todos los casos, tomamos nota de la distancia de amalgamado y actualizamos la matriz de disimilarirdades en aquéllos elementos que lo requieren y se continúa el proceso. Nótes que cada vex el número de columnas se reduce en uno. El proceso finaliza cuando se amalgaman los objetos o grupos que asociados a las dos últimas columnas que subsistan, en cuyo momento hemos creado un único agrupamiento que engloba a la totalidad de los objetos iniciales.

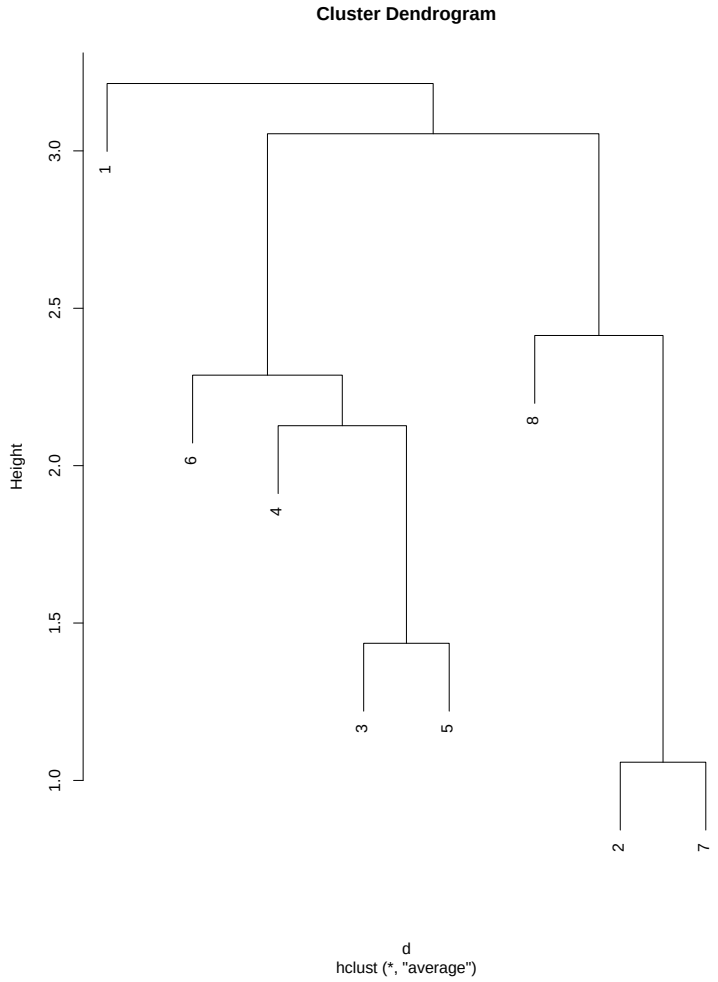
El procedimiento anterior se dice que es *jerárquico*. En efecto, en cada etapa del proceso la relación entre dos grupos cualesquiera sólo puede ser de inclusión (uno totalmente contenido en otro) o de exclusión (ambos completamente disjuntos).

Dendrograma

El proceso de amalgamado en una estrategia jerárquica puede representarse convenientemente mediante un *dengrograma*.

R: Ejemplo 15.1

Figura 15.1: Agrupamiento jerárquico con distancia promedio de 10 puntos tomados al azar en R^4



Apéndice A

Cálculo diferencial. Notación matricial.

Hay aquí sólo una breve recopilación de resultados útiles. Más detalles y demostraciones en **?** y **?**.

A.0.2. Notación

Haremos uso de las siguientes definiciones y notación.

Definición A.1 *Sea* $\boldsymbol{X}$ *un vector* $m \times 1$ *e* Y *una función escalar de* $\boldsymbol{X}$*:* $Y = f(X_1, \ldots, X_m) = f(\boldsymbol{X})$*. Entonces:*

$$\left(\frac{\partial Y}{\partial X}\right) \quad \stackrel{\text{def}}{=} \quad \begin{pmatrix} \frac{\partial Y}{\partial X_1} \\ \frac{\partial Y}{\partial X_2} \\ \vdots \\ \frac{\partial Y}{\partial X_m} \end{pmatrix}$$

Si $Y = \boldsymbol{X}'A\boldsymbol{X}$ siendo A una matriz cuadrada cualquiera, es inmediato comprobar que:

$$\left(\frac{\partial Y}{\partial \boldsymbol{X}}\right) = (A + A')\boldsymbol{X}.$$

En el caso, frecuente, de que A sea simétrica, tenemos que:

$$\left(\frac{\partial Y}{\partial \boldsymbol{X}}\right) = 2A'\boldsymbol{X}$$

Definición A.2 *Sea $\vec{Y}$ una función vectorial $n \times 1$–valorada de $\boldsymbol{X}$, vector $m \times 1$. Entonces:*

$$\left(\frac{\partial \vec{Y}}{\partial \boldsymbol{X}}\right) \quad \stackrel{\text{def}}{=} \quad \begin{pmatrix} \frac{\partial Y_1}{\partial X_1} & \frac{\partial Y_2}{\partial X_1} & \cdots & \frac{\partial Y_n}{\partial X_1} \\ \vdots & \vdots & & \vdots \\ \frac{\partial Y_1}{\partial X_m} & \frac{\partial Y_2}{\partial X_m} & \cdots & \frac{\partial Y_n}{\partial X_m} \end{pmatrix}$$

Hay algunos casos particulares de interés. Si $Y = \boldsymbol{a}'\boldsymbol{X} = a_1X_1 + \ldots + a_mX_m$, siendo $\boldsymbol{a}$ un vector de constantes,

$$\frac{\partial Y}{\partial \boldsymbol{X}} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = \boldsymbol{a};$$

si $\vec{Y} = A\boldsymbol{X}$, siendo A una matriz $(n \times m)$ de constantes,

$$\left(\frac{\partial \vec{Y}}{\partial \boldsymbol{X}}\right) = A'.$$

A.0.3. Algunos resultados útiles

$$\frac{\partial \boldsymbol{X}' A \boldsymbol{X}}{\partial \boldsymbol{X}} = 2A\boldsymbol{X} \tag{A.1}$$

$$\frac{\partial \log_e |A|}{\partial A} = \left[A'\right]^{-1} \tag{A.2}$$

$$\frac{\partial \text{tr}(BA^{-1}C)}{\partial A} = -(A^{-1}CBA^{-1}) \tag{A.3}$$

Apéndice B

Datos

B.1. Records atléticos de diversos países.

País	100m	200m	400m	800m	1500m	5Km	10Km	Maratón
Argentina	10.39	20.81	46.84	1.81	3.70	14.04	29.39	137.72
Australia	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30
Austria	10.44	20.81	46.82	1.79	3.60	13.26	27.72	135.90
Bélgica	10.34	20.68	45.04	1.73	3.60	13.22	27.45	129.95
Bermuda	10.28	20.58	45.91	1.80	3.75	14.68	30.55	146.62
Brazil	10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
Birmania	10.64	21.52	48.30	1.80	3.85	14.45	30.28	139.95
Canada	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
Chile	10.34	20.80	46.20	1.79	3.71	13.61	29.30	134.03
China	10.51	21.04	47.30	1.81	3.73	13.90	29.13	133.53
Colombia	10.43	21.05	46.10	1.82	3.74	13.49	27.88	131.35
Cook-Islas	12.18	23.20	52.94	2.02	4.24	16.70	35.38	164.70
Costa	10.94	21.90	48.66	1.87	3.84	14.03	28.81	136.58
Checoslov.	10.35	20.65	45.64	1.76	3.58	13.42	28.19	134.32
Dinamarca	10.56	20.52	45.89	1.78	3.61	13.50	28.11	130.78
Rep. Dom.	10.14	20.65	46.80	1.82	3.82	14.91	31.45	154.12
Finlandia	10.43	20.69	45.49	1.74	3.61	13.27	27.52	130.87
Francia	10.11	20.38	45.28	1.73	3.57	13.34	27.97	132.30
RDA	10.12	20.33	44.87	1.73	3.56	13.17	27.42	129.92
RFA	10.16	20.37	44.50	1.73	3.53	13.21	27.61	132.23
UK	10.11	20.21	44.93	1.70	3.51	13.01	27.51	129.13
Grecia	10.22	20.71	46.56	1.78	3.64	14.59	28.45	134.60
Guatemala	10.98	21.82	48.40	1.89	3.80	14.16	30.11	139.33
Hungria	10.26	20.62	46.02	1.77	3.62	13.49	28.44	132.58
India	10.60	21.42	45.73	1.76	3.73	13.77	28.81	131.98

País	100m	200m	400m	800m	1500m	5Km	10Km	Maratón
Indonesia	10.59	21.49	47.80	1.84	3.92	14.73	30.79	148.83
Irlanda	10.61	20.96	46.30	1.79	3.56	13.32	27.81	132.35
Israel	10.71	21.00	47.80	1.77	3.72	13.66	28.93	137.55
Italia	10.01	19.72	45.26	1.73	3.60	13.23	27.52	131.08
Japon	10.34	20.81	45.86	1.79	3.64	13.41	27.72	128.63
Kenya	10.46	20.66	44.92	1.73	3.55	13.10	27.38	129.75
Korea	10.34	20.89	46.90	1.79	3.77	13.96	29.23	136.25
RD-Korea	10.91	21.94	47.30	1.85	3.77	14.13	29.67	130.87
Luxemb.	10.35	20.77	47.40	1.82	3.67	13.64	29.08	141.27
Malasia	10.40	20.92	46.30	1.82	3.80	14.64	31.01	154.10
Mauricio	11.19	22.45	47.70	1.88	3.83	15.06	31.77	152.23
Mexico	10.42	21.30	46.10	1.80	3.65	13.46	27.95	129.20
Holanda	10.52	20.95	45.10	1.74	3.62	13.36	27.61	129.02
N.Zelanda	10.51	20.88	46.10	1.74	3.54	13.21	27.70	128.98
Noruega	10.55	21.16	46.71	1.76	3.62	13.34	27.69	131.48
Papua-N.G.	10.96	21.78	47.90	1.90	4.01	14.72	31.36	148.22
Filipinas	10.78	21.64	46.24	1.81	3.83	14.74	30.64	145.27
Polonia	10.16	20.24	45.36	1.76	3.60	13.29	27.89	131.58
Portugal	10.53	21.17	46.70	1.79	3.62	13.13	27.38	128.65
Rumania	10.41	20.98	45.87	1.76	3.64	13.25	27.67	132.50
Singapur	10.38	21.28	47.40	1.88	3.89	15.11	31.32	157.77
España	10.42	20.77	45.98	1.76	3.55	13.31	27.73	131.57
Suecia	10.25	20.61	45.63	1.77	3.61	13.29	27.94	130.63
Suiza	10.37	20.46	45.78	1.78	3.55	13.22	27.91	131.20
Taiwan	10.59	21.29	46.80	1.79	3.77	14.07	30.07	139.27
Tailandia	10.39	21.09	47.91	1.83	3.84	15.23	32.56	149.90
Turquia	10.71	21.43	47.60	1.79	3.67	13.56	28.58	131.50
USA	9.93	19.75	43.86	1.73	3.53	13.20	27.43	128.22
USSR	10.07	20.00	44.60	1.75	3.59	13.20	27.53	130.55
Samoa	10.82	21.86	49.00	2.02	4.24	16.28	34.71	161.83

Fuente: ?