MANUAL ABREVIADO DE ESTADÍSTICA MULTIVARIANTE

Jesús Montanero Fernández

Departamento de Matemáticas

Universidad de Extremadura

Introducción

El presente manual autoeditado y revisado periódicamente pretende constituir una introducción a las técnicas clásicas de la Estadística Multivariante, con breves incursiones en métodos más novedosos. Está concebido en principio como apoyo a la docencia en una asignatura de unas sesenta horas lectivas, correspondiente al cuarto curso del Grado en Estadística de la Universidad de Extremadura, e impartida por tanto a alumnos con cierta formación matemática en general y estadística en particular. Consta de un capítulo inicial enfocado mayormente a la comprensión de los modelos lineales univariante y multivariante, que se tratan en los dos capítulos siguientes. El estudio del modelo lineal es interesante de por sí para cualquier estadístico, pero en nuestro contexto debe entenderse mayormente como una herramienta teórica que nos permite comprender mejor el problema de clasificación, que se afronta en el capítulo 4. Se trata del tema más interesante desde el punto de vista práctico, junto con el capítulo 5, que aborda un problema genuinamente multivariante como es el de simplificar observaciones multidimensionales con el fin último de entender datos complejos. El último capítulo está dedicado el análisis de conglomerados, técnica que puede estudiarse también desde el punto de vista de la Minería de Datos.

La intención a la hora de de elaborar este manual ha sido exponer los contenidos de manera breve, pero indicando al lector referencias bibliográficas oportunas para profundizar en el estudio de la Estadística Multivariante. En particular, no se incluyen las demostraciones de los resultados teóricos. Algunas son asequibles y se proponen como ejercicio; otras pueden encontrarse en la bibliografía recomendada, por ejemplo en los manuales 56 y 59 del Servicio de Publicaciones de la UEx, que podrían considerarse como versiones extendidas del presente volumen. Tampoco pretende ser exhaustivo. De hecho, ciertas técnicas que podemos catalogar de multivariantes, como el análisis de componentes principales no lineal, el escalamiento multidimensional, el análisis de correspondencias y las redes neuronales se introducen a título meramente informativo. El lector interesado puede encontrar información más amplia sobre todas ellas en Hastie et al. (2008), Gifi (1990), Hair et al. (1999) y Uriel y Aldás (2005). Debe mencionarse también que, de todas las referencias indicadas en la bibliografía, los que han tenido una influencia más patente en la redacción de este volumen ha sido Arnold (1981) y Hatie et al. (2008).

Por último, hacemos constar que los diferentes gráficos y tablas de resultados que aparecen a lo largo del volumen han sido obtenidos mediante los programas SPSS y R. De hecho, se incluyen algunas capturas de pantallas para ilustrar la ejecución de las técnicas más relevantes con SPSS.

Badajoz, junio de 2019

Índice general

Capítulo 1

Preliminares

En este capítulo intentaremos fijar la notación, así como definir e interpretar conceptos fundamentales en el contexto de la Estadística Multivariante, muchos de los cuales deben ser conocidos. También llevaremos a cabo un breve repaso de Álgebra Lineal.

1.1. Notación

En general, solemos manejar en estadística dos tipos de lenguajes: probabilístico y muestral. El primero sirve para expresar las propiedades de la población objeto del estudio, entendiendo población en un sentido amplio; el segundo se utiliza para expresar las propiedades de una muestra de n datos extraídos, se supone que aleatoriamente, de dicha población.

El marco formal en el que se desarrolla el estudio poblacional es el espacio L^2 de funciones reales de cuadrado integrable, definidas sobre cierto espacio de probabilidad. Queremos decir que las variables aleatorias que estudiemos se identificarán con elementos de L^2. El estudio muestral tiene lugar en el espacio Euclídeo $\mathbb{R}^n$, es decir que, dada una variable aleatoria $X \in L^2$, una muestra aleatoria de tamaño n de dicha variable se identificará con un vector $\mathtt{X}$ de $\mathbb{R}^n$, cuyas componentes $\mathtt{X}_i$ serán las distintas mediciones de la misma. Obsérvese que hemos utilizado distintas fuentes de letra para denotar ambos conceptos, norma que intentaremos seguir en la medida de lo posible.

En el contexto del análisis multivariante, X puede denotar con frecuencia un vector aleatorio p-dimensional de componentes $X[1], \ldots, X[p]$. En tal caso, una muestra aleatoria de tamaño n para dicho vector aleatorio se expresará mediante la matriz $\mathtt{X} \in \mathcal{M}_{n\times p}$ que descompone así:

$$\mathtt{X} = (\mathtt{X}[1], \ldots, \mathtt{X}[p]) = \begin{pmatrix} \mathtt{X}_1[1] & \ldots & \mathtt{X}_1[p] \\ \vdots & & \vdots \\ \mathtt{X}_n[1] & \ldots & \mathtt{X}_n[p] \end{pmatrix} = \begin{pmatrix} \mathtt{X}_1' \\ \vdots \\ \mathtt{X}_n' \end{pmatrix} \tag{1.1}$$

A título de ejemplo, en el cuadro 1.1 se expone una muestra de tamaño $n = 38$ de un vector aleatorio de dimensión $p = 8$. Los datos corresponden a medidas de la motilidad de espermatozoides en moruecos y fueron recogidos por J.A. Bravo en el CENSYRA de Badajoz.

L^2 forma parte de una categoría de espacios que generalizan el concepto de espacio Euclídeo por estar también dotados de un producto interior. Concretamente, dados $f, g \in L^2$, se define

$$\langle f, g\rangle = \mathtt{E}_P[f \cdot g] \tag{1.2}$$

$\mathtt{E}_P$ se entiende como el funcional que asigna a cada variable aleatoria su integral respecto a la probabilidad P definida en el espacio de origen. El subíndice P suele omitirse. En el contexto estadístico en podemos considerar en $\mathbb{R}^n$ el siguiente producto interior proporcional al conocido como producto escalar:

$$\langle \mathtt{a}, \mathtt{b} \rangle = \frac{1}{n} \sum_{i=1}^{n} \mathtt{a}_i \mathtt{b}_i \tag{1.3}$$

En ambos espacios, los respectivos productos inducen sendas normas (al cuadrado), definidas en general mediante $\|a\|^2 = \langle a, a \rangle$ y, en consecuencia, sendas métricas basadas en la norma al cuadrado de las diferencias:

$$d^2(X, Y) = \mathtt{E}[(X - Y)^2], \qquad X, Y \in L^2 \tag{1.4}$$

$$\mathtt{d}_n^2(\mathtt{X}, \mathtt{Y}) = \frac{1}{n} \sum_i (\mathtt{X}_i - \mathtt{Y}_i)^2, \qquad \mathtt{X}, \mathtt{Y} \in \mathbb{R}^n \tag{1.5}$$

La segunda es, salvo una homotecia, la distancia Euclídea al cuadrado en $\mathbb{R}^n$. El uso de estas distancias para cuantificar errores se asocia al denominado método de Mínimos Cuadrados. Por otra parte, del producto interior se deriva a su vez una noción de ortogonalidad o perpendicularidad. En $\mathbb{R}^n$ decimos que $\mathtt{a}$ y $\mathtt{b}$ son ortogonales entre sí cuando $\langle \mathtt{a}, \mathtt{b} \rangle = 0$, en cuyo caso se denota $\mathtt{a} \perp \mathtt{b}$. En L^2 se define de manera análoga.

Proyección ortogonal: La noción de perpendicularidad se relacionará bajo ciertas condiciones con los conceptos estadísticos de incorrelación y de independencia. Además, da pie a considerar un tipo de función lineal denominada proyección ortogonal. Concretamente, si V es un subespacio lineal cerrado del espacio E (E en nuestro caso puede tratarse de L^2 o de $\mathbb{R}^n$), se define P_V como la aplicación que asigna a cada elemento e del espacio el único elemento de V tal que $e - P_V e \perp V$, en cuyo caso la distancia entre e y $P_V e$ es la mínima posible entre e y un elemento de V. Si V_1 y V_2 son dos subespacios ortogonales de E, se verifica que que $P_{V_1 \oplus V_2} = P_{V_1} + P_{V_2}$. Además, $\|P_{V_1 \oplus V_2} e\|^2 = \|P_{V_1} e\|^2 + \|P_{V_2} e\|^2$. Para $V \subset \mathbb{R}^n$, dado que P_V es una aplicación lineal se identificará con una matriz $n \times n$ que se denotará de la misma forma.

▪ *Ejercicio* 1. Dado $V \subset \mathbb{R}^m$, probar que $\mathtt{tr}(P_V) = \dim V$, y que todos lo elementos de la diagonal de P_V pertenecen al intervalo $[0, 1]$.

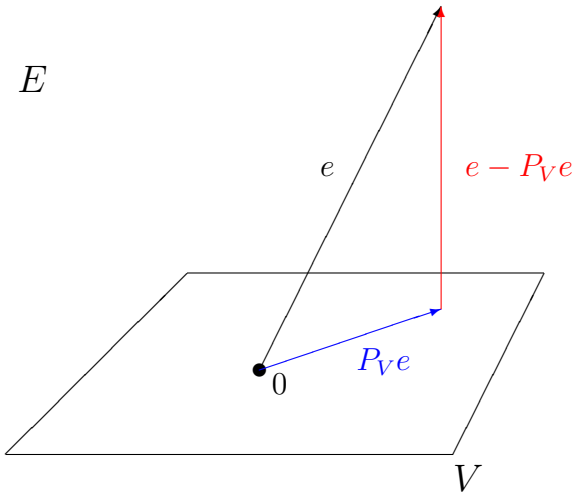


Figura 1.1: Proyección ortogonal

La colección de resultados teóricos conocida como Leyes de los Grandes Números y Teorema Central del Límite establecen una clara conexión entre los espacios $\mathbb{R}^n$ y L^2, si entendemos $\mathtt{X} \in \mathbb{R}^n$ como una muestra aleatoria simple de una variable aleatoria $X \in L^2$. La idea germinal de estos resultados y, por lo tanto, la conexión entre Estadística y Probabilidad, podría ilustrarse mediante el triángulo de Pascal y enunciarse en términos intuitivos así: si asumimos una completa simetría en la elección de n individuos (equiprobabilidad e independencia), la media de sus respectivos datos se estabiliza conforme n aumenta, debido a la conmutatividad (invarianza ante permutaciones) de la operación suma.

En todo caso, lo más importante en esta sección es resaltar que todos las definiciones en L^2 expresadas en términos del producto interior pueden traducirse automáticamente al lenguaje muestral e interpretarse de manera completamente análoga. Por ello, en este capítulo nos centraremos principalmente en el estudio de los parámetros probabilísticos o poblacionales (salvo en la sección 1.3, donde se fijará la notación de los parámetros muestrales), dejando como ejercicio para el lector el estudio paralelo en términos muestrales. Por lo general seguiremos la costumbre habitual de expresar los parámetros probabilísticos mediante letras griegas y sus homólogos muestrales con notación latina.

Si $\mathcal{X}$ es una familia de k elementos, bien sean de L^2 o de $\mathbb{R}^n$ (en el segundo caso puede identificarse con una matriz $n \times k$), se denota por $\langle \mathcal{X} \rangle$ su expansión lineal. En el espacio L^2 se denotará por $\mathtt{1}$ la variable aleatoria con valor constante 1, siendo entonces $\langle \mathtt{1} \rangle$ el subespacio unidimensional de las funciones constantes en L^2; se denotará por $\langle \mathtt{1} \rangle^\perp$ su ortogonal, que es un hiperplano de L^2. Análogamente, se denotará por 1_n al vector de $\mathbb{R}^n$ cuyas componentes son todas 1, siendo por tanto $\langle 1_n \rangle$ la recta de los vectores constantes y $\langle 1_n \rangle^\perp$ su ortogonal, de dimensión $(n-1)$.

1.2. Principales parámetros probabilísticos

En esta sección definiremos los parámetros relacionados con los momentos de orden uno y dos. Con ello estamos centrando indirectamente nuestro estudio en el ámbito de la distribución normal y de las relaciones de tipo lineal.

Media: Primeramente definimos la media de una variable aleatoria X como su esperanza, es decir, su integral respecto a la probabilidad considerada. Se denota por $\mathtt{E}[X]$ o por la letra μ, acompañada si es necesario de un subíndice aclaratorio. Si X es un vector p-dimensional, su media es el vector p-dimensional compuesto por las respectivas medias, y se denotará de forma idéntica.

Varianza: Dada una variable aleatoria $X \in L^2$ de media μ, se define su varianza mediante

$$\mathtt{var}[X] = \mathtt{E}[(X-\mu)^2] \tag{1.6}$$

denotándose también por la letra σ^2. La raíz cuadrada positiva de la varianza se denomina desviación típica. Nótese que la varianza está bien definida al ser X de cuadrado integrable. De hecho, puede expresarse mediate

$$\mathtt{var}[X] = \mathtt{E}[X^2] - \mathtt{E}[X]^2 \tag{1.7}$$

▪ *Ejercicio* 2. Probar (1.7).

- *Ejercicio* 3. Probar que $\mu = P_{\langle \mathtt{1} \rangle} X$.

Del ejercicio 3 se deduce que la media μ de una variable aleatoria X es la función constante más próxima en términos de la distancia (1.4). Así pues μ se interpreta como la constante más próxima a X. La diferencia $X - \mu \in \langle \mathtt{1} \rangle^{\perp}$ se denomina variabilidad de X. La distancia respecto a esa constante más próxima es precisamente la varianza:

$$\mathtt{var}[X] = d^2(X, \mathtt{E}[X]) \tag{1.8}$$

Varianza total: Si $\mathtt{X}$ es un vector aleatorio p-dimensional de componentes $X[1], \ldots, X[p]$, se define la varianza total de X mediante

$$\mathtt{var}_T[X] = \sum_{j=1}^{p} \mathtt{var}\big[X[j]\big] \tag{1.9}$$

Este parámetro puede interpretarse en términos de la distancia $d^2_{[p]}$ definida en el espacio de los p-vectores aleatorios con componentes en L^2 mediante

$$d^2_{[p]}(X, Y) = \mathtt{E}_P \left[\|X - Y\|^2_{\mathbb{R}^p} \right] \tag{1.10}$$

- *Ejercicio* 4. Probar que $\mathtt{E}[X]$ es el vector aleatorio constante que más se aproxima a X en términos de la distancia (1.10) y que, además,

$$\mathtt{var}_T[X] = d^2_{[p]}(X, \mathtt{E}[X]) \tag{1.11}$$

Covarianza: Dado un vector aleatorio p-dimensional X, se define la covarianza entre dos componentes cualesquiera $X[i]$ y $X[j]$ del mismo como el producto interior de sus respectivas variabilidades, es decir,

$$\mathtt{cov}\big[X[i], X[j]\big] = \langle X[i] - \mu_i, X[j] - \mu_j \rangle \tag{1.12}$$

denotándose también por σ_{ij}. Se trata de una generalización de la varianza, pues $\sigma_{ii} = \sigma_i^2$, que describe, según veremos en la próxima sección, el grado de relación lineal existente entre las variabilidades, es decir, el grado de relación afín existente entre las variables originales. Se dice que dos variables son incorreladas cuando su covarianza es nula, es decir, cuando sus variabilidades son ortogonales.

- *Ejercicio* 5. Probar que $-\sigma_i \sigma_j \leq \sigma_{ij} \leq \sigma_i \sigma_j$

Coeficiente de correlación: La desigualdad anterior invita a definir el denominado coeficiente de correlación lineal

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{1.13}$$

que tiene la virtud de ser adimensional y estar comprendido en todo caso entre -1 y 1. La incorrelación se identifica con $\rho_{ij} = 0$. Procuraremos utilizar los subíndices sólo cuando sea estrictamente necesario.

Dado un vector aleatorio p-dimensional X, las posibles covarianzas componen una matriz simétrica que puede definirse mediante

$$\mathtt{Cov}[X] = \mathtt{E}[(X - \mu)(X - \mu)'] \tag{1.14}$$

cuya diagonal está compuesta por las diferentes varianzas. Suele denotarse por la letra Σ. Lo mismo ocurre con los coeficientes de correlación, que componen una matriz de correlaciones $p \times p$ simétrica cuya diagonal está compuesta por unos.

▪ *Ejercicio* 6. ¿Por qué es simétrica Σ? ¿Por qué la diagonal de la matriz de correlaciones está compuesta por unos?

Es muy frecuente contemplar transformaciones de un vector aleatorio del tipo $\tilde{X} = A'X + b$, con $A \in \mathcal{M}_{p\times m}$ y $b \in \mathbb{R}^m$.

▪ *Ejercicio* 7. Probar que, en ese caso, el vector m-dimensional $\tilde{X}$ verifica

$$\mathtt{E}[\tilde{X}] = A'\mathtt{E}[X] + b, \qquad \mathtt{Cov}[\tilde{X}] = A'\mathtt{Cov}[X]A \tag{1.15}$$

También es frecuente considerar una partición del vector aleatorio p-dimensional X en dos vectores X_1 y X_2 de dimensiones p_1 y p_2, respectivamente, lo cual da lugar a su vez a particiones obvias de la media y la matriz de covarianzas:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right) \tag{1.16}$$

En el caso particular $p_1 = 1$, es decir, cuando X_1 es una variable aleatoria real y X_2 un vector aleatorio $(p-1)$-dimesional, la descomposición de Σ será de la forma

$$\Sigma = \left(\begin{array}{c|c} \sigma_1^2 & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right) \tag{1.17}$$

En tal caso cabe definir el coeficiente de correlación lineal múltiple (al cuadrado) entre X_1 y X_2 mediante

$$\rho_{12}^2 = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_1^2} \tag{1.18}$$

Se trata de una generalización del coeficiente de correlación simple (al cuadrado) que interpretaremos en la siguiente sección.

1.3. Principales parámetros muestrales

Aunque los conceptos que introduciremos a continuación son completamente análogos a los propuestos en la sección anterior (con la salvedad de que se definen en esta ocasión en el espacio Euclídeo $\mathbb{R}^n$, correspondiente a las muestras aleatoria de tamaño n), conviene dedicarle al menos un pequeño espacio aunque sea sólo por aclarar la notación.

En el caso univariante ($p = 1$), dado un vector $\mathtt{X} \in \mathbb{R}^n$ definimos su media y varianza mediante

$$\overline{\mathtt{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathtt{x}_i, \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathtt{x}_i - \overline{\mathtt{x}})^2 \tag{1.19}$$

Estos parámetros gozan de interpretación análoga a la de sus homólogos probabilísticos. En general, si $\mathtt{X} = (\mathtt{X}[1], \ldots, \mathtt{X}[p])$ denota una matriz de dimensiones $n \times p$, $\overline{\mathtt{x}}$ denotará al vector de $\mathbb{R}^p$ (matriz de dimensiones $p \times 1$) cuyas componentes son las medias de las diferentes columnas de $\mathtt{X}$ ordenadas, es decir, $\overline{\mathtt{x}} = (\overline{\mathtt{x}}[1], \ldots, \overline{\mathtt{x}}[p])'$. En ese caso, se denotará por $\overline{\mathtt{X}}$ la matriz $n \times p$ definida mediante $1_n \cdot \overline{\mathtt{x}}'$, es decir, la que consiste en reemplazar cada dato de $\mathtt{X}$ por la media aritmética de su repectiva columna.

Covarianza muestral: Dados $\mathtt{X}, \mathtt{Y} \in \mathbb{R}^n$, se denota

$$\begin{aligned} s_{xy} &= n^{-1}\langle \mathtt{X} - \overline{\mathtt{X}}, \mathtt{Y} - \overline{\mathtt{Y}} \rangle \\ &= n^{-1}(\mathtt{X} - \overline{\mathtt{X}})'(\mathtt{Y} - \overline{\mathtt{Y}}) \\ &= n^{-1}\sum_{i=1}^{n}(\mathtt{X}_i - \overline{\mathtt{x}})(\mathtt{Y}_i - \overline{\mathtt{y}}) \end{aligned}$$

En general, si $\mathtt{X}$ denota una matriz dimensiones $n \times p$, se define la matriz de covarianzas muestral

$$S = \mathtt{n}^{-1}(\mathtt{X} - \overline{\mathtt{X}})'(\mathtt{Y} - \overline{\mathtt{Y}}) \in \mathcal{M}_{p\times p} \tag{1.20}$$

Si se contempla una partición concreta de las columnas de $\mathtt{X}$, la matriz S se descompondría de manera análoga a (1.17), lo cual da pie a la definición de nuevos parámetros como el coeficiente de correlación múltiple.

Coeficientes de correlación simple y múltiple: Dados $\mathtt{X}_1 \in \mathbb{R}^n$ y $\mathtt{X}_2 \in \mathcal{M}_{n\times q}$ se define R^2_{12} o R^2 a secas mediante

$$R^2 = \frac{S_{12}S_{22}^{-1}S_{21}}{s_1^2} \tag{1.21}$$

En el caso particular $q = 1$ estaríamos hablando del cuadrado del coeficiente de correlación simple, que se define mediante $r = s_{xy}/s_x s_y$. Por último, si $\mathtt{X}$ es una matriz $p \times p$ se denota por $\mathtt{R}$ la matriz compuesta por los coeficientes de correlación lineal simple entre sus diferentes columnas, que puede expresarse matricialmente mediante $\mathtt{R} = S^{-1/2} \cdot S \cdot S^{-1/2}$.

Análogamente a (1.15), las transformaciones lineales de un vector aleatorio p-dimensional X, es decir, del tipo $A'X$, con $A \in \mathcal{M}_{p\times m}$, se traducen en términos de una muestra de tamaño $\mathtt{X} \in \mathcal{M}_{n\times p}$ de tamaño n en una transformación del tipo $\mathtt{X}A$.

- *Ejercicio* 8. Probar que, en ese caso, se verifica

$$\overline{\mathtt{X} \cdot A} = \overline{\mathtt{X}} \cdot A, \qquad S_{\mathtt{X}\cdot A} = A' S_{\mathtt{X}} A \tag{1.22}$$

1.4. Regresión lineal

Consideremos un vector aleatorio X descompuesto en X_1 y X_2 según (1.16) con $p_1 = 1$, es decir, tenemos una variable aleatoria real X_1 y un vector X_2 p_2-dimensional. Nuestra intención es explicar la variabilidad de X_1 como función lineal de la variabilidad de X_2, en la medida de lo posible. Por lo tanto, buscamos el vector $\beta \in \mathbb{R}^{p_2}$ que alcance el siguiente mínimo, en cuyo caso se denominará solución mínimo-cuadrática:

$$\text{mín}\left\{ \|X_1 - \mathtt{E}[X_1] - b'(X_2 - \mathtt{E}[X_2])\|^2 : \ b \in \mathbb{R}^{p_2} \right\} \tag{1.23}$$

Ecuación de regresión lineal: La solución se obtiene proyectando ortogonalmente el vector aleatorio $X_1 - \mathtt{E}[X_1]$ sobre el subespacio $\langle X_2 - \mathtt{E}[X_2] \rangle$, como indica la figura 1.2. Se trata pues de buscar el vector β tal que

$$X_1 - \mathtt{E}[X_1] - \beta'(X_2 - \mathtt{E}[X_2]) \ \perp \ \langle X_2 - \mathtt{E}[X_2] \rangle \tag{1.24}$$

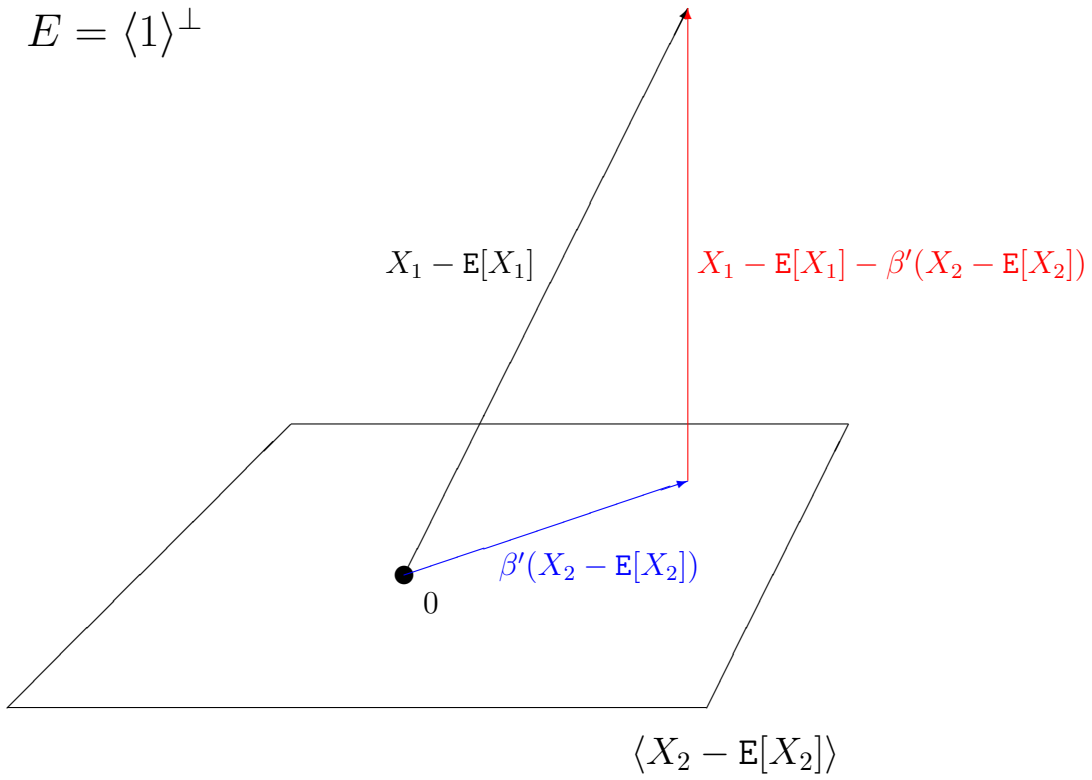


Figura 1.2: Ecuación de regresión lineal

En el caso $p_2 = 1$ (regresión simple), el problema se reduce a encontrar el escalar $\beta \in \mathbb{R}$ tal que

$$\langle X_1 - \mathtt{E}[X_1], X_2 - \mathtt{E}[X_2] \rangle = \beta \cdot \langle X_2 - \mathtt{E}[X_2], X_2 - \mathtt{E}[X_2] \rangle \tag{1.25}$$

que, según (1.12), es $\beta = \sigma_{22}^{-1}\sigma_{21}$.

▪ *Ejercicio* 9. Probar que, en general, la ortogonalidad (1.24) se alcanza con el vector

$$\beta = \Sigma_{22}^{-1}\Sigma_{21} \tag{1.26}$$

Por otra parte, si se define

$$\alpha = \mathtt{E}[X_1] - \beta'\mathtt{E}[X_2], \tag{1.27}$$

se verifica que $X_1 = \alpha + \beta' X_2 + \mathcal{E}$, donde $\mathcal{E}$, denominado variabilidad parcial de X_1 dado X_2, tiene media nula. Respecto a $\mathtt{var}[\mathcal{E}]$, denominada varianza parcial, dado que $\|X_1 - \mathtt{E}[X_1]\|^2 = \sigma_1^2$ y $\|\beta'(X_2 - \mathtt{E}[X_2])\|^2 = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, se deduce que la proporción de variabilidad de X_1 explicada linealmente por la variabilidad de X_2 es ρ_{12}^2, definido en (1.18). En definitiva, tal y como se ilustra en la figura 1.3,

$$\begin{aligned} \mathtt{var}[\mathcal{E}] &= \sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} && (1.28) \\ &= \sigma_1^2(1 - \rho_{12}^2) && (1.29) \end{aligned}$$

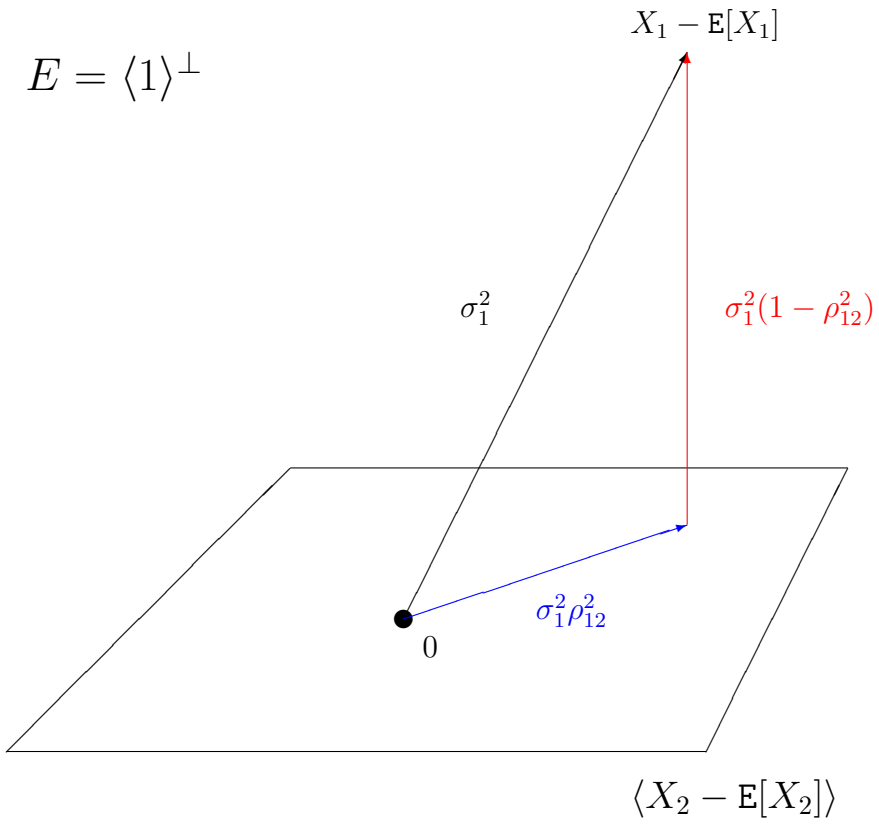


Figura 1.3: Descomposición de la varianza

Coeficientes de correlación parcial Razonando de manera análoga en el caso general $p_1 \geq 1$, obtenemos que la matriz de covarianzas de $\mathcal{E}$, denominada matriz de covarianzas parciales, es la siguiente

$$\mathtt{Cov}[\mathcal{E}] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{1.30}$$

que se denota $\Sigma_{11\cdot 2}$. Si $p_1 > 1$, los coeficientes de correlación asociados a esta matriz se denominan correlaciones parciales entre las componentes de X_1 dado el vector aleatorio X_2, que equivalen pues a las correlaciones simples entre las componentes del vector aleatorio $\mathcal{E}$. Así pues, si consideramos dos componentes de $X_1[i]$ y $X_1[j]$ de X_1, su correlación parcial al cuadrado dado X_2, que se denota por $\rho^2_{ij\cdot 2}$, se interpreta como la proporción de la variabilidad parcial de $X_1[i]$ dado X_2 explicada linealmente por la variabilidad parcial de $X_1[j]$ dado X_2, y viceversa. En un lenguaje más intuitivo podríamos definirlo como la proporción de variabilidad que comparten "en exclusiva", al margen de lo ya explicado por X_2.

Asimismo, podríamos generalizar el coeficiente de correlación múltiple como la matriz $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, pero de dicha matriz sólo nos interesarán sus autovalores, que denominaremos en el capítulo 3 coeficientes de correlación canónica.

▪ *Ejercicio* 10. Probar que $\alpha + \beta' X_2$ es la función afín de X_2 que minimiza la distancia respecto a X_1.

Incorrelación e independencia: Se dice que X_1 y X_2 son incorreladas cuando $\Sigma_{12} = 0$, lo cual equivale a $\beta = 0$ o, si $p_1 = 1$, $\rho_{12} = 0$. Se traduce por tanto en la imposibilidad de

explicar parte alguna de la variabilidad de X_1 como función lineal de la variabilidad de X_2. Geométricamente puede definirse así:

$$X_1 \text{ y } X_2 \text{ incorreladas } \Leftrightarrow \langle X_1 - \mathtt{E}[X_1]\rangle \perp \langle X_2 - \mathtt{E}[X_2]\rangle \tag{1.31}$$

La independencia supone sin embargo una propiedad estrictamente más fuerte que la incorrelación. Efectivamente, puede ocurrir que entre X_1 y X_2 no se dé relación afín alguna pero que, sin embargo, exista entre ambas una relación de otro tipo, que podría ser incluso funcional. $\mathtt{E}[X_1|X_2] \circ X_2$ es la función medible de X_2 que mejor se aproxima a X_1 según la métrica (1.4) y no podemos en general afirmar que se trate de una función afín. Eso sí ocurre bajo el supuesto de $(p_1 + p_2)$-normalidad, como veremos en el próximo capítulo. En ese caso, debe verificarse entonces $\mathtt{E}[X_1|X_2] \circ X_2 = \alpha + \beta X_2$, con α y β definidas como antes.

El concepto probabilístico de independencia lo suponemos conocido. Desde un punto de vista geométrico, podría definirse como sigue: primeramente, dado un vector k-dimensional Y con componentes en L^2, denótese por $\mathcal{M}(Y)$ el espacio de las variables en $\langle \mathtt{1}\rangle^\perp$ que son funciones medibles de Y. En tal caso, se verifica

$$X_1 \text{ y } X_2 \text{ independientes } \Leftrightarrow \mathcal{M}(X_1) \perp \mathcal{M}(X_2) \tag{1.32}$$

Al igual que podemos establecer una conexión entre el coeficiente de correlación y el concepto de independencia, podríamos asociar el coeficiente de correlación parcial al de independencia condicional.

▪ *Ejercicio* 11. Probar (1.31) y (1.32). Deducir entonces que la independencia implica incorrelación.

▪ *Ejercicio* 12. Interpretar los parámetros muestrales definidos en la sección 3 en los términos de la sección 4.

1.5. Nociones básicas de Álgebra Lineal

Aparte de los conceptos introducidos en la primera sección debemos destacar algunas nociones y resultados propios del Álgebra Lineal que se manejan con frecuencia en nuestra teoría. Hemos de tener presente en todo momento tres observaciones: primero que, fijada una base vectorial en $\mathbb{R}^m$, las aplicaciones lineales de $\mathbb{R}^m$ en $\mathbb{R}^m$ se identifican con las matrices cuadradas de orden m; segundo, que una vez fijado un orden de lectura, el conjunto $\mathcal{M}_{n\times p}$ de matrices de dimensión $n \times p$ se identifica con $\mathbb{R}^{np}$; tercero, que dicha identificación permite definir un producto interior en $\mathcal{M}_{n\times p}$ mediante

$$\begin{aligned} \langle A, B\rangle &= \mathtt{tr}(A'B) && (1.33)\\ &= \sum_{i,j} a_{ij}b_{ij} && (1.34) \end{aligned}$$

Este producto interior permite generalizar la distancia (1.5) al conjunto $\mathcal{M}_{n\times p}$ mediante:

$$\begin{aligned} \mathtt{d}^2_{n,p}(A, B) &= n^{-1}\mathtt{tr}[(A - B)'(A - B)] && (1.35)\\ &= n^{-1}\sum_{i=1}^{n} \|a_i - b_i\|^2_{\mathbb{R}^p} && (1.36) \end{aligned}$$

donde a_i' y b_i' denotan las filas de A y B, respectivamente. Esta distancia generalizada puede entenderse a su vez como una versión muestral de la distancia (1.10). Entre otras propiedades, podemos destacar que $\mathtt{tr}(A'B) = \mathtt{tr}(B'A)$ y que, si A, B, C son matrices cuadradas de orden m, se verifica que $\mathtt{tr}(ABC) = \mathtt{tr}(CAB) = \mathtt{tr}(BCA)$.

▪ *Ejercicio* 13. Probar (1.34) y (1.36).

▪ *Ejercicio* 14. Dada una matriz de datos $\mathtt{X} \in \mathcal{M}_{n\times p}$, probar que la varianza total muestral de $\mathtt{X}$, definida de manera análoga a (1.9) como la suma de las varianzas muestrales de sus p-componentes, verifica

$$s_T^2 = \mathtt{d}_{n,p}^2(\mathtt{X}, \overline{\mathtt{X}}) \tag{1.37}$$

Matriz positiva: En el conjunto de matrices cuadradas $m \times m$, podemos definir el siguiente preorden que generaliza el orden natural en $\mathbb{R}$: decimos que $A \geq B$ cuando $x'Ax \geq x'Bx$ para todo $x \in \mathbb{R}^m$. Así mismo, decimos que $A > B$ cuando la desigualdad anterior es estricta si $x \neq 0$. En consecuencia, $A \geq 0$ cuando $x'Ax \geq 0$ para todo $x \in \mathbb{R}^m$, en cuyo caso se dice que A es semidefinida positiva. Si $A > 0$ se dice definida positiva.

Distancia de Mahalanobis: Dada una matriz $A \in \mathcal{M}_{m\times m}$ simétrica y positiva podemos definir en $\mathbb{R}^m$ la distancia de Mahalanobis D_A^2 mediante

$$D_A^2(x, y) = (x - y)'A^{-1}(x - y), \qquad x, y \in \mathbb{R}^m \tag{1.38}$$

Se trata de una generalización de la métrica Euclídea, que se obtendría en el caso $A = \mathtt{Id}$.

Matriz ortogonal: Se dice que una matriz $\Gamma \in \mathcal{M}_{m\times m}$ es ortogonal cuando sus columnas constituyen una base ortonormal de $\mathbb{R}^m$, es decir, cuando $\Gamma' = \Gamma^{-1}$. El conjunto de matrices ortogonales de orden m se denotará por $\mathcal{O}_m$.

Matriz idempotente: Se dice que una matriz $A \in \mathcal{M}_{m\times m}$ es idempotente cuando $A^2 = A$. Puede probarse que, si V es un subespacio lineal de $\mathbb{R}^m$ y $B \in \mathcal{M}_{m\times \dim V}$ es una base de V (entendemos con esto que las columnas de B constituyen una base de V, es decir, que $V = \langle B \rangle$), entonces la matriz P_V que se identifica con la proyección ortogonal sobre V puede calcularse mediante

$$P_V = B(B'B)^{-1}B' \tag{1.39}$$

Se trata pues de una matriz simétrica e idempotente.

▪ *Ejercicio* 15. Probar (1.39). Es más, probar que una matriz $A \in \mathcal{M}_{m\times m}$ simétrica e idempotente se identifica con la proyección ortogonal sobre $V = \langle A \rangle$.

Autovalores y autovectores: Dada una matriz $A \in \mathcal{M}_{m\times m}$, se dice que $\delta \in \mathbb{R}$ es un autovalor real de A y $\gamma \in \mathbb{R}^m$ un autovector asociado cuando se verifica que $Ae = \delta \cdot \gamma$. En tal caso, δ debe ser necesariamente una raíz del polinomio $p(x) = |A - x \cdot \mathtt{Id}|$ y $\langle \gamma \rangle$ debe estar incluido en $\ker(A - \delta \cdot \mathtt{Id})$. Puede probarse que, si A es simétrica, las m raíces de $p(x)$ son reales, lo cual equivale a la existencia de m autovalores reales contados con su multiplicidad. El siguiente resultado, conocido como teorema de diagonalización de una matriz simétrica, aclara la estructura de la familia de autovectores asociados.

Teorema 1.5.1. Dada una matriz $A \in \mathcal{M}_{m\times m}$ simétrica, si Δ denota la matriz diagonal compuesta por los autovalores $\delta_1, \ldots, \delta_m$ de A ordenados de mayor a menor y contados con su multiplicidad, existe una matriz $\Gamma \in \mathcal{O}_m$, cuyos vectores columnas se denotan por $\gamma_1, \ldots, \gamma_m$, tal que

$$A = \Gamma\Delta\Gamma' \tag{1.40}$$

Se verifica además que $\delta_1 = \text{máx}\{\gamma' A\gamma : \ \|\gamma\| = 1\}$, que se alcanza con $\gamma = \gamma_1$, y que, para todo $j = 2, \ldots, m$, $\delta_j = \text{máx}\{\gamma' A\gamma : \ \|\gamma\| = 1, \ \gamma \perp \langle\gamma_1, \ldots, \gamma_{j-1}\rangle\}$, alcanzándose con $\gamma = \gamma_j$.

Del teorema se sigue directamente que las columnas de Γ constituyen una base ortonormal de autovectores asociados a los correspondientes autovalores. También podemos de deducir de (1.40) que $\Delta = \Gamma^{-1}A\Gamma$. Por lo tanto, la aplicación lineal identificada con la matriz A para la base vectorial original admite una expresión diagonal respecto a una base ortonormal canónica de autovectores. Es decir, el cambio a la base de autovectores permite expresar la matriz de forma sencilla. A modo de ejemplo, podemos utilizar ese procedimiento para demostrar las siguientes propiedades;

- *Ejercicio* 16. Dada una matriz simétrica A, probar:

  (i) Si A es simétrica, su rango coincide con el número de autovalores no nulos.

  (ii) Si $A \geq 0$, sus autovalores son todos no negativos. Si $A > 0$, son todos estrictamente positivos.

  (iii) Si $A \geq 0$, existe una matriz simétrica $A^{1/2}$ tal que $A = A^{1/2}A^{1/2}$. Si $A > 0$, existe también una matriz simétrica $A^{-1/2}$ tal que $A^{-1} = A^{-1/2}A^{-1/2}$.

  (iv) Si $A \geq 0$, existe una matriz X con las mismas dimensiones tal que $A = X'X$.

  (v) La traza de una matriz simétrica es la suma de sus autovalores y el determinante, el producto de los mismos.

  (vi) La inversa de una matriz simétrica positiva también es positiva.

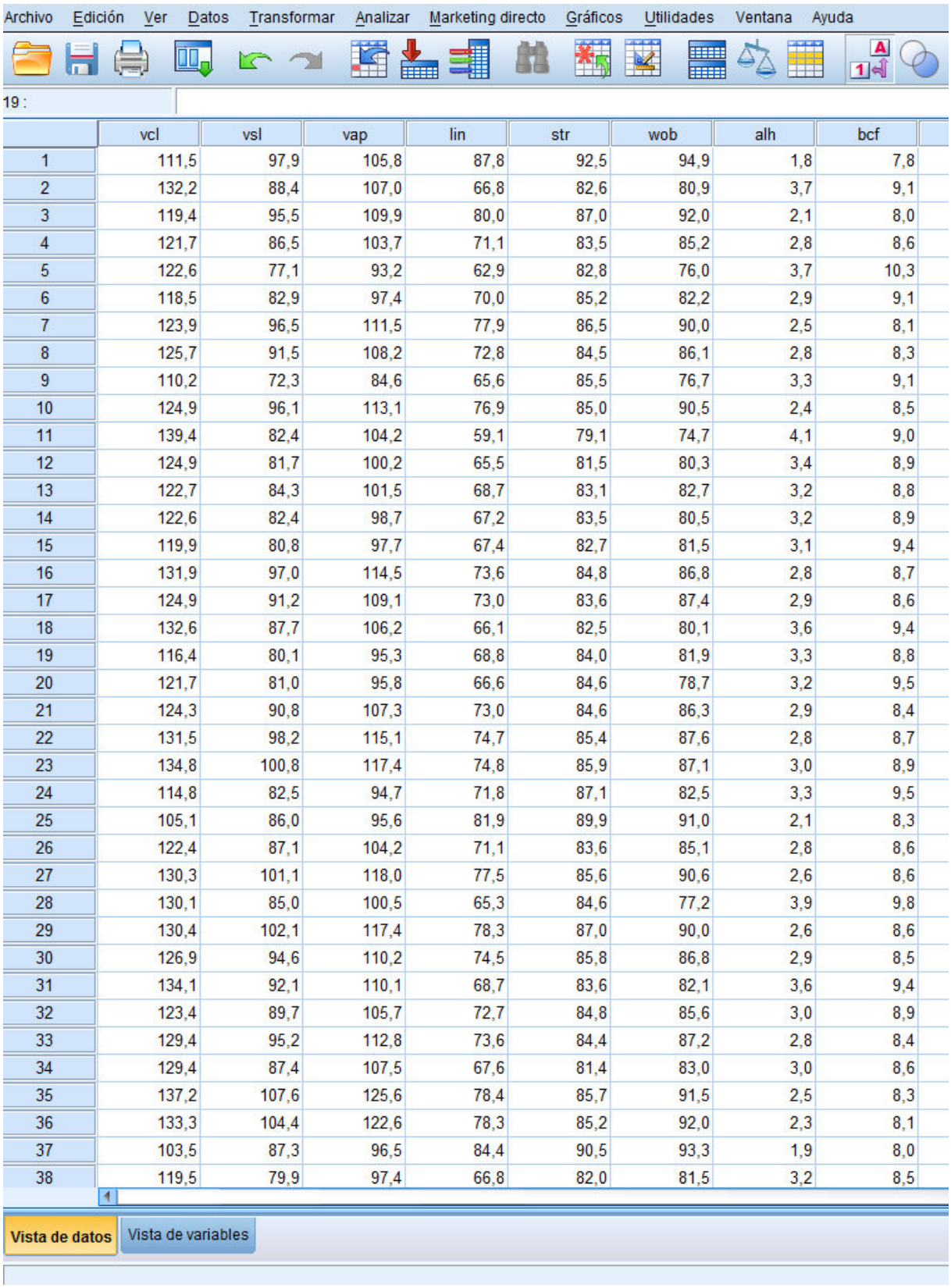
  (vii) La matriz P_V de la proyección ortogonal sobre el subespacio lineal $V \subset \mathbb{R}^n$ consta de tantos autovalores 1 como $\dim V$, siendo nulos los $n - \dim V$ restantes.

A partir del teorema 1.5.1 y del ejercicio 1 podemos probar el siguiente resultado en el cual se fundamenta el capítulo 5:

Lema 1.5.2. En las condiciones del teorema 1.5.1 y dado $k \leq m$, si Γ_1 es la matriz con los autovectores asociados a los k primeros autovalores de A, se verifica que

$$\text{máx}\{\mathtt{tr}(B'AB) : \ B \in \mathcal{M}_{m\times k}, \ B'B = \mathtt{Id}\} = \sum_{i=1}^{k} \delta_i \tag{1.41}$$

y se alcanza en $B = \Gamma_1$.



	vcl	vsl	vap	lin	str	wob	alh	bcf
1	111,5	97,9	105,8	87,8	92,5	94,9	1,8	7,8
2	132,2	88,4	107,0	66,8	82,6	80,9	3,7	9,1
3	119,4	95,5	109,9	80,0	87,0	92,0	2,1	8,0
4	121,7	86,5	103,7	71,1	83,5	85,2	2,8	8,6
5	122,6	77,1	93,2	62,9	82,8	76,0	3,7	10,3
6	118,5	82,9	97,4	70,0	85,2	82,2	2,9	9,1
7	123,9	96,5	111,5	77,9	86,5	90,0	2,5	8,1
8	125,7	91,5	108,2	72,8	84,5	86,1	2,8	8,3
9	110,2	72,3	84,6	65,6	85,5	76,7	3,3	9,1
10	124,9	96,1	113,1	76,9	85,0	90,5	2,4	8,5
11	139,4	82,4	104,2	59,1	79,1	74,7	4,1	9,0
12	124,9	81,7	100,2	65,5	81,5	80,3	3,4	8,9
13	122,7	84,3	101,5	68,7	83,1	82,7	3,2	8,8
14	122,6	82,4	98,7	67,2	83,5	80,5	3,2	8,9
15	119,9	80,8	97,7	67,4	82,7	81,5	3,1	9,4
16	131,9	97,0	114,5	73,6	84,8	86,8	2,8	8,7
17	124,9	91,2	109,1	73,0	83,6	87,4	2,9	8,6
18	132,6	87,7	106,2	66,1	82,5	80,1	3,6	9,4
19	116,4	80,1	95,3	68,8	84,0	81,9	3,3	8,8
20	121,7	81,0	95,8	66,6	84,6	78,7	3,2	9,5
21	124,3	90,8	107,3	73,0	84,6	86,3	2,9	8,4
22	131,5	98,2	115,1	74,7	85,4	87,6	2,8	8,7
23	134,8	100,8	117,4	74,8	85,9	87,1	3,0	8,9
24	114,8	82,5	94,7	71,8	87,1	82,5	3,3	9,5
25	105,1	86,0	95,6	81,9	89,9	91,0	2,1	8,3
26	122,4	87,1	104,2	71,1	83,6	85,1	2,8	8,6
27	130,3	101,1	118,0	77,5	85,6	90,6	2,6	8,6
28	130,1	85,0	100,5	65,3	84,6	77,2	3,9	9,8
29	130,4	102,1	117,4	78,3	87,0	90,0	2,6	8,6
30	126,9	94,6	110,2	74,5	85,8	86,8	2,9	8,5
31	134,1	92,1	110,1	68,7	83,6	82,1	3,6	9,4
32	123,4	89,7	105,7	72,7	84,8	85,6	3,0	8,9
33	129,4	95,2	112,8	73,6	84,4	87,2	2,8	8,4
34	129,4	87,4	107,5	67,6	81,4	83,0	3,0	8,6
35	137,2	107,6	125,6	78,4	85,7	91,5	2,5	8,3
36	133,3	104,4	122,6	78,3	85,2	92,0	2,3	8,1
37	103,5	87,3	96,5	84,4	90,5	93,3	1,9	8,0
38	119,5	79,9	97,4	66,8	82,0	81,5	3,2	8,5

Cuadro 1.1: Matriz de datos muestra tamaño $n = 38$ y dimensión $p = 8$

Capítulo 2

Modelo lineal multivariante

En este capítulo expondremos los aspectos más generales del modelo lineal normal multivariante. Previamente, estudiaremos con brevedad las distribuciones de probabilidad relacionadas con este modelo así como el modelo lineal normal (univariante) que pretende generalizar.

2.1. Distribución normal multivariante

La distribución normal multivariante p-dimensional o p-normal se trata de una generalización natural de la distribución normal que servirá como hipótesis de partida en el modelo estadístico objeto de estudio.

Dados $\mu \in \mathbb{R}^p$ y $\Sigma \in \mathcal{M}_{p\times p}$ simétrica y semidefinida positiva, se dice que un vector aleatorio X p-dimensional sigue un modelo de distribución $N_p(\mu, \Sigma)$ cuando admite la siguiente función característica:

$$\varphi_X(t) = \exp\left\{it'\mu - \frac{1}{2}t'\Sigma t\right\}, \qquad t \in \mathbb{R}^p \tag{2.1}$$

En ese caso se denota $X \sim N_p(\mu, \Sigma)$ y puede comprobarse trivialmente que generaliza la distribución normal unidimensional. Vamos a enunciar a continuación las propiedades fundamentales de esta distribución. Las dos siguientes se siguen de las propiedades de la función característica.

Proposición 2.1.1. Si $X \sim N_{p_2}(\mu, \Sigma)$, $A \in \mathcal{M}_{p_1\times p_2}$ y $b \in \mathbb{R}^{p_1}$, entonces

$$AX + b \sim N_{p_1}(A\mu + b, A\Sigma A') \tag{2.2}$$

Proposición 2.1.2. Si $Z[1], \ldots, Z[p]$ iid N(0,1), entonces $Z = (Z[1], \ldots, Z[p])' \sim N_p(0, \mathtt{Id})$

A partir de las dos propiedades anteriores podemos construir cualquier vector normal:

Proposición 2.1.3. Dados μ y Σ como en la definición, si consideramos el vector aleatorio Z anterior, la descomposición $\Sigma = \Gamma\Delta\Gamma'$ y se denota $A = \Gamma\Delta^{1/2}$, se sigue que $AZ + \mu \sim N_p(\mu, \Sigma)$.

En consecuencia, se sigue de (1.15) el siguiente resultado:

Proposición 2.1.4. Si $X \sim N_p(\mu, \Sigma)$, $\mathtt{E}[X] = \mu$ y $\mathtt{Cov}[X] = \Sigma$.

También es consecuencia de la proposición 2.1.1 que, si $X \sim N(\mu, \Sigma)$, cada componente $X[i]$ de X sigue un modelo de distribución $N(\mu_i, \sigma_i^2)$. Sin embargo, el recíproco no es cierto. Hemos de tener en cuenta que la componente $X[i]$ puede obtenerse mediante $e_i'X$, siendo e_i el vector unidad en el eje de coordenadas i-ésimo, y que la siguiente afirmación puede probarse con relativa facilidad:

Proposición 2.1.5. Dado un vector aleatorio p-dimensional X, cualquiera de las condiciones siguientes garantizan la p-normalidad del mismo:

(i) $a'X$ es 1-normal, para todo $a \in \mathbb{R}^p$.

(ii) Sus componentes son todas normales e independientes entre sí.

El siguiente resultado puede probarse también a través de la función característica y establece la equivalencia entre incorrelación e independencia bajo la hipótesis de normalidad.

Proposición 2.1.6. Si descomponemos un vector (p_1+p_2)-normal X con matriz de covarianzas Σ en X_1 de dimensión p_1 y X_2 de dimensión p_2, entonces X_1 y X_2 son independientes s, y sólo si, $\Sigma_{12} = 0$.

Si la matriz de covarianzas es estrictamente positiva, la distribución p-normal es dominada por la medida de Lebesgue en $\mathbb{R}^p$. Teniendo en cuenta las proposiciones 2.1.1, 2.1.2 y el teorema del cambio de variables, podemos obtener la densidad de dicha distribución:

Proposición 2.1.7. Si $X \sim N_p(\mu, \Sigma)$ con $\Sigma > 0$ admite la siguiente función de densidad:

$$f(\mathtt{x}) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left\{-\frac{1}{2}(\mathtt{x}-\mu)'\Sigma^{-1}(\mathtt{x}-\mu)\right\}, \quad \mathtt{x} \in \mathbb{R}^n. \tag{2.3}$$

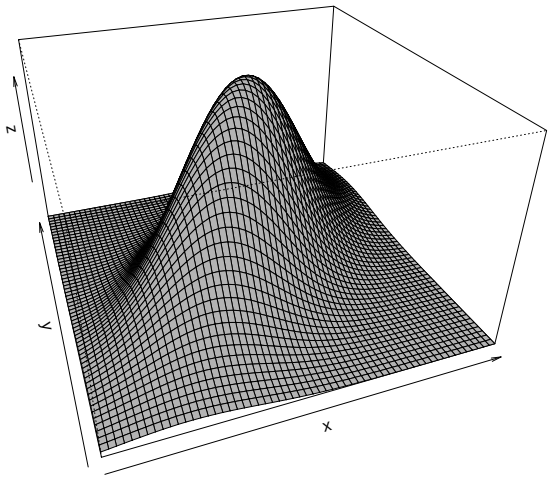


Figura 2.1: Función de densidad distribución 2-normal

▪ *Ejercicio* 17. Probar las siete proposiciones anteriores.

Nótese que en la función de verosimilitud determinada por (2.3) la observación $\mathtt{x}$ y los parámetros (μ, Σ) que caracterizan la distribución de probabilidad se relacionan a través de la distancia de Mahalanobis $D^2_\Sigma(\mathtt{x}, \mu)$. Concretamente, para cada $k \in [0, [(2\pi)^p|\Sigma|]^{-1/2}]$, la región de los puntos $\{\mathtt{x} \in \mathbb{R}^p : f(\mathtt{x}) = k\}$, es decir, aquéllos cuya densidad es igual a k, es el elipsoide siguiente:

$$\{\mathtt{x} \in \mathbb{R}^p : D^2_\Sigma(\mathtt{x}, \mu) = \tilde{k}\} \tag{2.4}$$

para $\tilde{k} = -2\log\left(k\sqrt{(2\pi)^p|\Sigma|}\right)$.

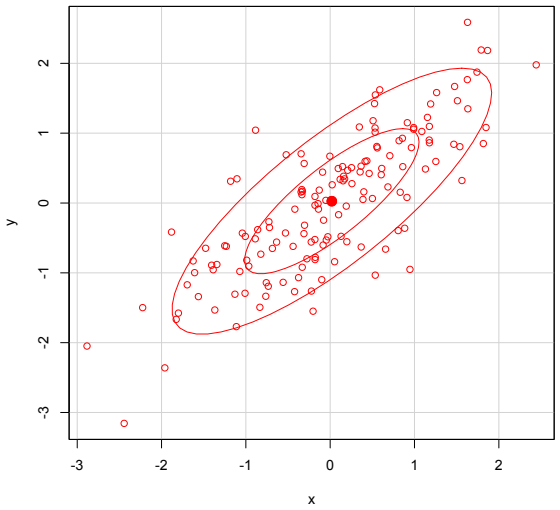


Figura 2.2: Contornos distribución 2-normal

En la figura 2.1 se aprecia la función de densidad de la distribución bidimensional

$$N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right) \tag{2.5}$$

mientras que en la figura 2.2 se podemos ver un diagrama de dispersión con una muestra aleatoria simple de tamaño $n = 150$ de dicha distribución en la que aparecen marcados dos contornos elípticos de la misma.

Consideremos un vector aleatorio X $(p_1 + p_2)$-normal que descompone de la forma

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{p_1+p_2}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \tag{2.6}$$

El siguiente resultado puede probarse teniendo en cuenta el hecho conocido de que la densidad de la distribución condicional $P^{X_1|X_2}$ puede calcularse mediante

$$f_{X_1|X_2=\mathtt{x}_2}(\mathtt{x}_1) = \frac{f_{X_1,X_2}(\mathtt{x}_1, \mathtt{x}_2)}{f_{X_2}(\mathtt{x}_2)} \tag{2.7}$$

Proposición 2.1.8. Si $\Sigma_{22} > 0$, se verifica

$$X_1|X_2 = \mathtt{x}_2 \sim N_{p_1}(\alpha + \beta'\mathtt{x}_2, \Sigma_{11\cdot 2}) \tag{2.8}$$

con α, β y $\Sigma_{11\cdot 2}$ definidas según (1.27), (1.26) y (1.30), respectivamente.

Como consecuencia se deduce que, bajo el supuesto de normalidad, $\mathtt{E}[X_1|X_2] \circ X_2 = \alpha + \beta' X_2$. Es más, podemos garantizar que

$$X_1 = \alpha + \beta' X_2 + \mathcal{E}, \qquad \mathcal{E} \text{ y } X_2 \text{ independientes, } \mathtt{E}[\mathcal{E}] = 0, \ \mathtt{Cov}[\mathcal{E}] = \Sigma_{11\cdot 2} \tag{2.9}$$

Esta afirmación puede probarse también teniendo en cuenta la proposición 2.1.6, (1.31) y (1.32). En definitiva, establece una clara conexión entre los conceptos de normalidad y linealidad.

▪ *Ejercicio* 18. Si X denota un vector 2-normal siguiendo un modelo de distribución (2.5), razonar qué modelo de distribución sigue en cada caso el vector Y indicando, si procede, su función de densidad:

(a) $Y[1] = 1 + 2X[1] + 3X[2]$; $Y[2] = 4 - X[1] + X[2]$

(b) $Y[1] = 2 + 5X[1] - 4X[2]$

(c) $Y[1] = 1 + 2X[1] + 3X[2]$; $Y[2] = 4 - 4X[1] - 6X[2]$

▪ *Ejercicio* 19. Simular de manera aproximada una muestra de tamaño $n = 200$ de la distribución (2.5).

Desde el punto de vista estadístico, podemos proponer tests para contrastar la hipótesis inicial de normalidad multivariante. En Bilodeau y Brenner (1999) se recoge un test que se basa en el hecho de que, para una muestra aleatoria simple de tamaño n de una distribución p-normal, las distancias de Mahalanobis entre las observaciones y la media aritmética de la misma dada la matriz de covarianzas muestral siguen una distribución de la familia Beta y tienden a la incorrelación conforme aumenta el tamaño de muestra. Desde una perspectiva eminentemente práctica, si realmente tenemos la intención de utilizar alguno de los procedimientos de tipo paramétrico que expondremos a continuación, resulta más realista comprobar que los diagramas de dispersión entre las diferentes componentes revelan al menos relaciones de tipo lineal, estando muy pendiente de la presencia de sesgos, que pueden conducirnos a transformar las variables originales, o fragmentaciones de los datos, que pueden conducirnos a introducir factores cualitativos en el modelo.

2.2. Modelo lineal

Antes de abordar el estudio del modelo lineal multivariante, repasaremos muy brevemente el modelo lineal en dimensión 1, empezando por las distribuciones de probabilidad asociadas al mismo.

2.2.1. Distribuciones asociadas al modelo

No pretendemos aquí describir con detalle las diferentes distribuciones pues se suponen conocidas (ver por ejemplo Nogales (1998)), sino interpretarlas desde un punto de vista geométrico.

Distribución normal esférica: El punto de partida del modelo es la distribución normal multivariante esférica, que se obtiene cuando la matriz de covarianzas del vector es de la forma $\Sigma = \sigma^2 \mathtt{Id}$, para algún $\sigma^2 > 0$. Efectivamente, puede comprobarse que, en ese caso, la distancia de Mahalanobis D^2_Σ es, salvo una homotecia, la distancia Euclídea, por lo que los elipsoides (2.4) son en particular esferas cuyo centro es la media. En virtud de la proposición 2.1.5-(ii), $\mathtt{Y} \sim N_n(\mu, \sigma^2 \mathtt{Id})$ si, y sólo si, sus componentes $\mathtt{Y}_1, \ldots, \mathtt{Y}_n$ son independientes, normales y con idéntica varianza (homocedásticos). Es decir, que esta distribución está asociada a una muestra de tamaño n en sentido amplio, de ahí la notación utilizada.

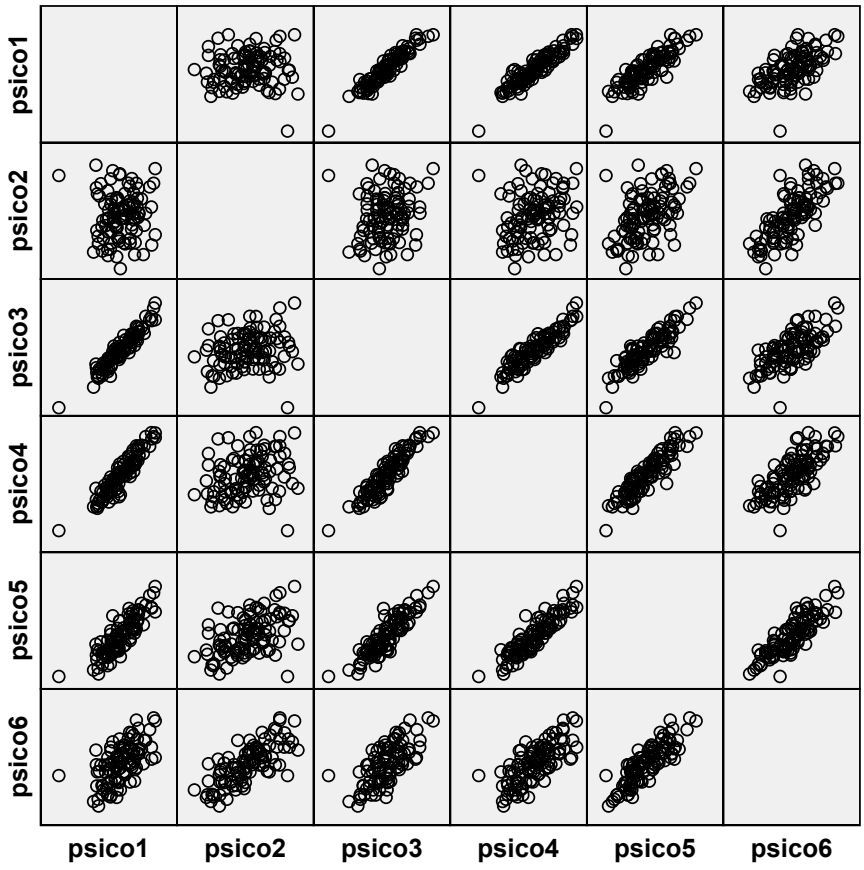


Figura 2.3: Simulación muestra $n = 100$ distribución 6-normal

Distribución χ^2: Cuando la media es 0, la distribución normal esférica tiene además la particularidad de ser invariante ante cualquier transformación mediante una matriz ortogonal, es decir, que la verosimilitud de cada observación $\mathtt{Y}$ depende exclusivamente de $\|\mathtt{Y}\|^2$. Eso explica nuestro interés en estudiar la distribución de $\|\mathtt{Y}\|^2$ bajo la hipótesis de normalidad esférica. Efectivamente, si $\mathtt{Y} \sim N_n(0, \mathtt{Id})$, se dice que $\|\mathtt{Y}\|^2 \sim \chi^2_n$; la anterior distribución denominada χ^2 central puede generalizarse si consideramos la norma Euclídea al cuadrado de un vector $\mathtt{Y} \sim N_n(\mu, \mathtt{Id})$, que se distribuye según un modelo $\chi^2_n(\delta)$, con $\delta = \|\mu\|^2$. Si $\mathtt{Y} \sim N_n(\mu, \sigma^2\mathtt{Id})$, entonces $\sigma^{-2}\|\mathtt{Y}\|^2/ \sim \chi^2_n(\delta)$ con $\delta = \|\mu\|^2/\sigma^2$, lo cual se denota por $\|\mathtt{Y}\|^2 \sim \sigma^2\chi^2_n(\delta)$.

- *Ejercicio* 20. Probar que, en general, si $\mathtt{Y} \sim N_n(\mu, \sigma^2\mathtt{Id})$ y $E \subset \mathbb{R}^n$,

$$\frac{1}{\sigma^2}\|P_E\mathtt{Y}\|^2 \sim \chi^2_{\dim E}(\delta), \qquad \delta = \frac{1}{\sigma^2}\|P_E\mu\|^2 \tag{2.10}$$

Por otra parte, se verifica además que $\mathtt{E}\left[\|P_E\mathtt{Y}\|^2\right] = (\dim E)\sigma^2(1+\delta)$. Como caso particular, si $\mu \in E^\perp$, entonces $\|P_E\mathtt{Y}\|^2 \sim \sigma^2\chi^2_{\dim E}$.

- *Ejercicio* 21. Probar que, dados $E_1 \perp E_2$ y $X \sim N_n(\mu, \sigma^2\mathtt{Id})$, se verifica que $\|P_{E_i}\mathtt{Y}\|^2 \sim \sigma^2\chi^2_{\dim E_i}(\|P_{E_i}\mu\|^2/\sigma^2)$, para $i = 1, 2$, y son independientes.

Distribución F: Este modelo probabilístico de distribución surge de manera natural de la aplicación del Principio de Invarianza en el problema estadístico de contraste de una hipótesis lineal para la media que veremos más adelante. Efectivamente, en la resolución de dicho problema nos vemos avocados a considerar, dado un vector $\mathtt{Y} \sim N_n(\mu, \sigma^2\mathtt{Id})$, el cociente entre $\|P_{E_1}\mathtt{Y}\|^2$ y $\|P_{E_2}\mathtt{Y}\|^2$ para ciertos subespacios E_1 y E_2 ortogonales entre sí y tales que $\mu \in E_2^{\perp}$. No obstante, ambos términos se normalizan dividiéndolos por sus respectivas dimensiones, de manera que la distribución F se obtiene mediante

$$\frac{\|P_{E_1}\mathtt{Y}\|^2/\dim E_1}{\|P_{E_2}\mathtt{Y}\|^2/\dim E_2} \sim F_{\dim E_1, \dim E_2}(\delta), \quad \delta = \frac{\|P_{E_1}\mu\|^2}{\sigma^2} \tag{2.11}$$

Nótese que el cociente entre las medias del numerador y el denominador es $(1+\delta)$ y, por lo tanto, 1 cuando $\delta = 0$. La distribución $m \cdot F_{m,n}$ converge a χ^2_m cuando n tiende a infinito.

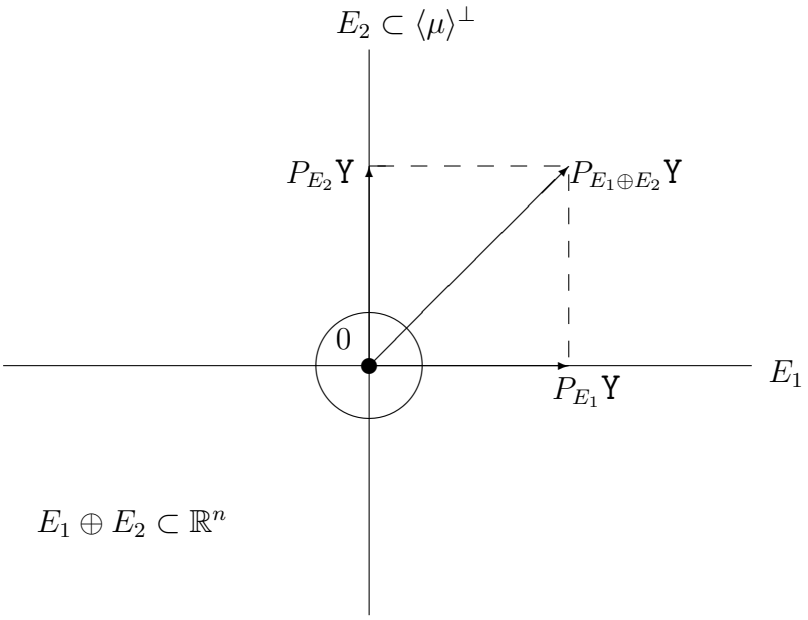


Figura 2.4: Interpretación geométrica de la distribución F

Distribución t: En esta sección interpretaremos la conocida distribución t-Student como un caso particular de la distribución F. Concretamente, decimos que una variable real t sigue un modelo de distribución $t_m(\delta)$ cuando es simétrico respecto a 0 y tal que $t^2 \sim F_{1,m}(\delta)$. De esta forma, nos encontraremos con dicha distribución cuando operemos como en la figura 2.4 con $\dim E_1 = 1$.

2.2.2. El modelo y ejemplos

El modelo lineal (normal) consiste en una estructura o experimento estadístico en $\mathbb{R}^n$ donde la distribución de probabilidad es normal esférica $N_n(\mu, \sigma^2\mathtt{Id})$. No se impone ninguna condición respecto al parámetro σ^2 pero si se impone una restricción de tipo lineal para el parámetro μ, pues se supondrá por hipótesis que $\mu \in V$ para un cierto subespacio lineal conocido $V \subset \mathbb{R}^n$. Se denota mediante

$$\mathtt{Y} \sim N_n(\mu, \sigma^2), \quad \mu \in V,\ \sigma^2 > 0 \tag{2.12}$$

La restricción lineal $\mu \in V$ vendrá dada, bien por la presencia de factores cualitativos, bien por la relación lineal respecto a otras variables numéricas con valores conocidos.

Si una matriz $\mathtt{X} \in \mathcal{M}_{n\times \dim V}$ constituye una base de V, podemos parametrizar el modelo (2.12) a través de las coordenadas β de μ respecto a $\mathtt{X}$, es decir, $\mathtt{Y} \sim N_n(\mathtt{X}\beta, \sigma^2\mathtt{Id})$, o equivalentemente,

$$\mathtt{Y} = \mathtt{X}\beta + \mathcal{E}, \quad \beta \in \mathbb{R}^{\dim V},\ \mathcal{E} \sim N_n(0, \sigma^2\mathtt{Id}),\ \sigma^2 > 0 \tag{2.13}$$

Enunciaremos a continuación cuatro ejemplo de problemas estadísticos que se formalizan mediante el modelo lineal:

Ejemplo 1. [Muestra aleatoria simple de una distribución normal] Consideremos $\mathtt{Y}_1, \ldots, \mathtt{Y}_n$ iid $N(\nu, \sigma^2)$. En ese caso, el vector aleatorio $\mathtt{Y} = (\mathtt{Y}_1, \ldots, \mathtt{Y}_n)'$ sigue una distribución $N_n(\mu, \sigma^2\mathtt{Id})$ con $\mu \in V = \langle 1_n \rangle$ y $\sigma^2 > 0$.

Ejemplo 2. [Muestras independientes de distribuciones normales con idéntica varianza] Consideremos ahora, para $i = 1, 2$, sendas muestras independientes entre sí $\mathtt{Y}_{i1}, \ldots, \mathtt{Y}_{in_i}$ iid $N(\mu_i, \sigma^2)$. Si se denota $n = n_1 + n_2$ e $\mathtt{Y} = (\mathtt{Y}_{11}, \ldots, \mathtt{Y}_{2n_2})'$, se verifica que $\mathtt{Y} \sim N_n(\mu, \sigma^2\mathtt{Id})$ con $\sigma^2 > 0$ y $\mu \in V = \langle \mathtt{v}_1, \mathtt{v}_2 \rangle$, donde $\mathtt{v}_1$ denota el vector de $\mathbb{R}^n$ cuyas n_1 primeras componentes son 1 y el resto 0. De manera análoga se define $\mathtt{v}_2$.

Ejemplo 3. [Diseño completamente aleatorizado] Se trata de una generalización del problema anterior para $r \geq 2$ muestras independientes $\mathtt{Y}_{i1}, \ldots, \mathtt{Y}_{in_i}$ iid $N(\mu_i, \sigma^2)$. En este caso, si $n = \sum_i n_i$ e $\mathtt{Y} = (\mathtt{Y}_{11}, \ldots, \mathtt{Y}_{rn_r})'$, se verifica que $\mathtt{Y} \sim N_n(\mu, \sigma^2\mathtt{Id})$ con $\sigma^2 > 0$ y $\mu \in V = \langle \mathtt{v}_1, \ldots, \mathtt{v}_r \rangle$.

Ejemplo 4. [Regresión lineal múltiple] Supongamos que se recogen n observaciones independientes que pueden calcularse mediante una relación afín con los valores de otras q variables numéricas controladas en el experimento, salvo errores independientes, normalmente distribuidos y homocedásticos. Es decir,

$$\begin{array}{ccccccccc} \mathtt{Y}_1 & = & \beta_0 & + & \beta_1\mathtt{Z}_1[1] & +\ldots+ & \beta_q\mathtt{Z}_1[q] & +\mathcal{E}_1 \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots \\ \mathtt{Y}_n & = & \beta_0 & + & \beta_1\mathtt{Z}_n[1] & +\ldots+ & \beta_q\mathtt{Z}_n[q] & +\mathcal{E}_n \end{array} \tag{2.14}$$

donde $\mathcal{E}_1, \ldots, \mathcal{E}_n$ iid $N(0, \sigma^2)$. Si se denota

$$\mathtt{Y} = \begin{pmatrix} \mathtt{Y}_1 \\ \vdots \\ \mathtt{Y}_n \end{pmatrix}, \quad \mathtt{X} = \begin{pmatrix} 1 & \mathtt{Z}_1[1] & \ldots & \mathtt{Z}_1[q] \\ \vdots & \vdots & & \vdots \\ 1 & \mathtt{Z}_n[1] & \ldots & \mathtt{Z}_n[q] \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \mathcal{E}_1 \\ \vdots \\ \mathcal{E}_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_q \end{pmatrix} \tag{2.15}$$

el modelo puede expresase de la forma (2.13) con $V = \langle \mathtt{X} \rangle$. En lo sucesivo se denotará $\underline{\beta} = (\beta_1, \ldots, \beta_q)' \in \mathbb{R}^q$. Este vector expresa la influencia de los vectores explicativos $\mathtt{Z}[1], \ldots, \mathtt{Z}[q]$ en la predicción de la respuesta $\mathtt{Y}$. Así mismo, se denota por $\mathtt{Z}$ la matriz que resulta al eliminar de $\mathtt{X}$ el término independiente 1_n. De esta, forma, el modelo puede expresarse también mediante

$$\mathtt{Y} = \beta_0 \cdot 1_n + \mathtt{Z}\underline{\beta} + \mathcal{E} \tag{2.16}$$

Regresión respecto a variables dummys: Cualquiera de los problemas considerados anteriormente puede entenderse como un problema de regresión lineal, es decir, pueden parametrizarse de la forma (2.13) para una base $\mathtt{X}$ con término independiente. Así, en el caso del ejemplo 3, podemos considerar entre otras posibilidades la matriz $\mathtt{X} = (1_n, \mathtt{v}_1, \ldots, \mathtt{v}_{r-1})$. Con esta parametrización particular, la relación entre μ y β es la siguiente:

$$\beta_0 = \mu_r, \quad \beta_j = \mu_j - \mu_r, \ j = 1, \ldots, r-1 \tag{2.17}$$

- *Ejercicio* 22. Probar (2.17). Indicar así mismo cómo se relacionaría μ con β si consideráramos la base natural $\tilde{X} = (1_n, \mathtt{v}_1 - \mathtt{v}_r, \ldots, \mathtt{v}_{r-1} - \mathtt{v}_r)$.

Los vectores $\mathtt{Z}[1], \ldots, \mathtt{Z}[r-1]$ de $\mathtt{X}$ en la parametrización anterior recogen valores concretos de unas variables denominadas dummys que indican la muestra o categoría a la que pertenece cada dato. Que las medias $\mu_1, \ldots, \mu_r$ sean idénticas, es decir, que las muestras procedan de una única distribución común, equivale a que $\underline{\beta}$ sea nulo, independientemente de la parametrización particular considerada. En otras palabras, la ausencia de relación entre el factor cualitativo que distingue las muestras con la variable numérica Y equivale a la ausencia de relación de ésta con las variables numéricas dummys.

- *Ejercicio* 23. Desarrollar con detalle los modelos asociados a los cuatro ejemplos anteriores.

2.2.3. Estimación y contraste de hipótesis

Dado que suponemos $\mu \in V$ y que, al seguir $\mathtt{Y}$ un modelo de distribución n-normal, es más verosímil que la observación $\mathtt{Y}$ sea próxima a la media que lo contrario, parece razonable estimar μ mediante un vector de V próximo a $\mathtt{Y}$. De hecho, definimos el estimador

$$\hat{\mu} = P_V \mathtt{Y} \tag{2.18}$$

En tal caso, resulta también razonable estimar σ^2 mediante la distancia (1.5) entre $\mathtt{Y}$ y $\hat{\mu}$, es decir, $\hat{\sigma}^2_{MV} = n^{-1}\|P_{V^\perp}\mathtt{Y}\|^2$. Puede probarse que ambos estimadores son independientes y que constituyen un estadístico suficiente y completo. Se sigue entonces del teorema de Lehmann-Scheffe, que $\hat{\mu}$ es el estimador insesgado de mínima varianza (EIMV) de μ. También puede probarse a partir de (2.3) que $(\hat{\mu}, \hat{\sigma}^2_{MV})$ constituyen un estimador de máxima verosimilitud (EMV) de (μ, σ^2). Sin embargo, $\hat{\sigma}^2_{MV}$ no es insesgado, de ahí que se proponga el siguiente estimador que es, según el teorema de Lehmann-Scheffe, EIMV:

$$\hat{\sigma}^2 = \frac{1}{n - \dim V}\|P_{V^\perp}\mathtt{Y}\|^2 \tag{2.19}$$

Si el modelo está parametrizado de la forma (2.13), podemos estimar β como las coordenadas del estimador de μ, es decir:

$$\hat{\beta} = (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'\mathtt{Y} \tag{2.20}$$

En definitiva, los estimadores de μ y σ^2 pueden entenderse geométricamente según la figura 2.4 con $E_1 = V$ y $E_2 = V^\perp$.

- *Ejercicio* 24. Obtener los estimadores $\hat{\mu}$ y $\hat{\sigma}^2$ para los ejemplos 1 y 2.
- *Ejercicio* 25. Obtener $\hat{\mu}$ para el ejemplo 3. Probar que, en dicho ejemplo, el EIMV de σ^2 es

$$\hat{\sigma}^2 = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\mathtt{Y}_{ij} - \overline{\mathtt{y}}_{i\cdot})^2 \tag{2.21}$$

▪ *Ejercicio* 26. Probar que, en el ejemplo 4, podemos estimar β a partir de las medias aritméticas del vector $\mathtt{Y}$ y la matriz $\mathtt{Z}$, así como de la matriz de covarianza muestral conjunta mediante

$$\hat{\underline{\beta}} = S_{zz}^{-1} S_{zy}, \qquad \hat{\beta}_0 = \overline{\mathtt{y}} - \overline{\mathtt{z}}' \hat{\underline{\beta}} \tag{2.22}$$

Para resolver el anterior ejercicio se aconseja utilizar un razonamiento análogo al que se ilustra en la figura 1.2, pero en términos muestrales, y tener en cuenta la descomposición ortogonal $\langle \mathtt{X} \rangle = \langle 1_n \rangle \oplus \langle \mathtt{Z} - \overline{\mathtt{Z}} \rangle$.

▪ *Ejercicio* 27. Probar que, en el ejemplo 4, $\hat{\sigma}^2_{MV}$ puede relacionarse con la varianza[1] s_y^2 del vector $\mathtt{Y}$ y el coeficiente de correlación múltiple al cuadrado R^2 de $\mathtt{Y}$ respecto a $\mathtt{Z}$, definido en (1.21), mediante

$$\hat{\sigma}^2_{MV} = n^{-1}(n-1) s_y^2 (1 - R^2) \tag{2.23}$$

El problema de contraste de hipótesis relativas al parámetro σ^2 no será expuesto aquí debido a que los tests que los resuelven son sensibles ante la violación del supuesto de normalidad. No ocurre lo mismo con el test F o anova que resuelve el contraste de hipótesis de tipo lineal sobre el parámetro μ pues, tal y como se prueba en Arnold (1981), es asintóticamente válido aunque no se verifique el supuesto de normalidad. Además, es relativamente robusto ante la heretocedasticidad. Lo mismo ocurre en el modelo multivariante.

Anova: Nos ocuparemos pues del contraste de hipótesis tipo $H_0 : \mu \in W$, para algún subespacio lineal $W \subset V$. Veamos ejemplos de hipótesis de este tipo:

▪ *Ejercicio* 28. En el ejemplo 1 podemos contrastar si la media ν de la distribución es nula. Probar que se corresponde con $W = 0$.

▪ *Ejercicio* 29. En los ejemplos 2 y 3 podemos contrastar si todas las muestras consideradas provienen de una misma distribución de probabilidad. Probar que en ambos casos se corresponde con $W = \langle 1_n \rangle$.

▪ *Ejercicio* 30. En el ejemplo 4 podemos contrastar si los vectores explicativos $\mathtt{Z}[1], \ldots, \mathtt{Z}[q]$ no intervienen en la explicación de Y, lo cual equivale a $\underline{\beta} = 0$. Probar que se corresponde con $W = \langle 1_n \rangle$. Dicho contraste se denomina total.

▪ *Ejercicio* 31. En las condiciones del ejemplo 4 podemos contrastar también hipótesis del tipo $\beta_j = 0$. Probar que, por ejemplo, en el caso $j = q$, se corresponde con $W = \langle 1_n, \mathtt{Z}[1], \ldots, \mathtt{Z}[q-1] \rangle$. Dicho contraste se denomina parcial.

Si se denota $V|W = W^\perp \cap V$, la hipótesis inicial $H_0 : \mu \in W$ equivale a $P_{V|W}\mu = 0$. Ello invita a descomponer $\mathbb{R}^n$ en tres subespacios ortogonales: $\mathbb{R}^n = W \oplus V|W \oplus V^\perp$. De dicha descomposición se deriva la siguiente descomposición ortogonal del vector de observaciones:

$$\mathtt{Y} = P_W \mathtt{Y} + P_{V|W} \mathtt{Y} + P_{V^\perp} \mathtt{Y} \tag{2.24}$$

Si $W = 0$, como en el ejercicio 28, la descomposisión (2.24) se reduce a los dos últimos sumandos.

Teniendo en cuenta el Principio de Invarianza y el hecho de que $(\hat{\mu}, \hat{\sigma}^2)$ es suficiente, se sigue que la decisión correspondiente al contraste de H_0 debe depender de la observación Y a través del cociente entre $\|P_{V|W} Y\|^2$ y $\|P_{V^\perp} Y\|^2$. Si dividimos ambos números positivos por

[1] Nos referimos al estimador insesgado de la varianza $s_y^2 = (n-1)^{-1} \sum_{i=1}^n (\mathtt{Y}_i - \overline{\mathtt{y}})^2$.

sus respectivos grados de libertad obtenemos la distribución F. En definitiva, consideramos el estadístico de contraste

$$F(\mathtt{Y}) = \frac{n - \dim V}{\dim V|W} \cdot \frac{\|P_{V|W}\mathtt{Y}\|^2}{\|P_{V^\perp}\mathtt{Y}\|^2} \tag{2.25}$$

o, equivalentemente,

$$F(\mathtt{Y}) = \frac{\|P_{V|W}\mathtt{Y}\|^2/(\dim V|W)}{\hat{\sigma}^2} \tag{2.26}$$

que, según (2.11), sigue en general un modelo de distribución $F_{\dim V|W, n-\dim V}(\delta)$, con $\delta = \|P_{V|W}\mu\|^2/\sigma^2$. En particular, bajo la hipótesis inicial sigue una distribución $F_{\dim V|W, n-\dim V}$. Teniendo en cuenta el Principio de Máxima Verosimilitud, construimos el denominado test F o anova de manera que se rechace la hipótesis inicial si el estadístico F toma valores extremos, los cuales serán tanto más verosímiles cuanto más nos alejemos de la hipótesis inicial ($\delta = 0$). Del lema fundamental de Neyman-Pearson se deduce que el test así construido es UMP-invariante; además, es el test de la razón de verosimilitudes (TRV). Desde el punto de vista geométrico puede entenderse según la figura 2.4 con $E_1 = V|W$ y $E_2 = V^\perp$.

El caso de mayor interés práctico es $W = \langle 1_n \rangle$ (como en los ejercicios 29 y 30), en el cual la decomposión (2.24) de $\|\mathtt{Y}\|^2$ es la siguiente:

$$\begin{aligned} \|\mathtt{Y}\|^2 &= \|P_{\langle 1_n\rangle}\mathtt{Y}\|^2 + \|P_{\langle 1_n\rangle^\perp}\mathtt{Y}\|^2 && (2.27)\\ &= n\overline{\mathtt{y}}^2 + \|P_{V|\langle 1_n\rangle}\mathtt{Y}\|^2 + \|P_{V^\perp}\mathtt{Y}\|^2 && (2.28) \end{aligned}$$

▪ *Ejercicio* 32. Probar que, si consideramos cualquier matriz $\mathtt{Z}$ tal que de $\langle (1_n|\mathtt{Z})\rangle = V$ y R^2 denota el coeficiente de correlación múltiple entre $\mathtt{Y}$ y $\mathtt{Z}$, como en (2.23), la igualdad (2.28) puede expresarse también así

$$\|\mathtt{Y}\|^2 = n\overline{\mathtt{y}}^2 + (n-1)\left[R^2 s_y^2 + (1-R^2)s_y^2\right] \tag{2.29}$$

Nótese que el Principio de Invarianza nos conduce a desentendernos del primer sumando, correspondiente a un vector en $\langle 1_n\rangle$, y concentranos en el cociente entre el segundo y el tercero, tal y como se ilustra mediante la figura 2.5, que es la versión muestral de la figura 1.3.

Este razonamiento nos ayuda a comprender que el valor del coeficiente de correlación lineal múltiple R^2 depende exclusivamente del subespacio V (conteniendo a $\langle 1_n\rangle$) sobre el que deseemos proyectar las observaciones, independientemente de la matriz $(1_n|Z)$ que escojamos como base del mismo, lo cual invita a entender de manera más general la definición propuesta en (1.21). Concretamente, dados $Y \in \mathbb{R}^n$ y $V \subset \mathbb{R}^n$ conteniendo a $\langle 1_n\rangle$, se puede definir el coeficiente de correlación de Y respecto a V mediante

$$R^2 = \frac{\|P_{V|\langle 1_n\rangle}\mathtt{Y}\|^2}{\|P_{\langle 1_n\rangle^\perp}\mathtt{Y}\|^2} \tag{2.30}$$

En el caso particular $\dim V|W = 1$, es decir, cuando W es un hiperplano de V, el estadístico de contrate se denota por t^2 en lugar de F pues se confronta con la distribución $t^2_{n-\dim V}$, dando lugar a lo que conocemos como test de Student.

▪ *Ejercicio* 33. Relacionar la descomposición (2.28) con los términso de la tabla 2.1.

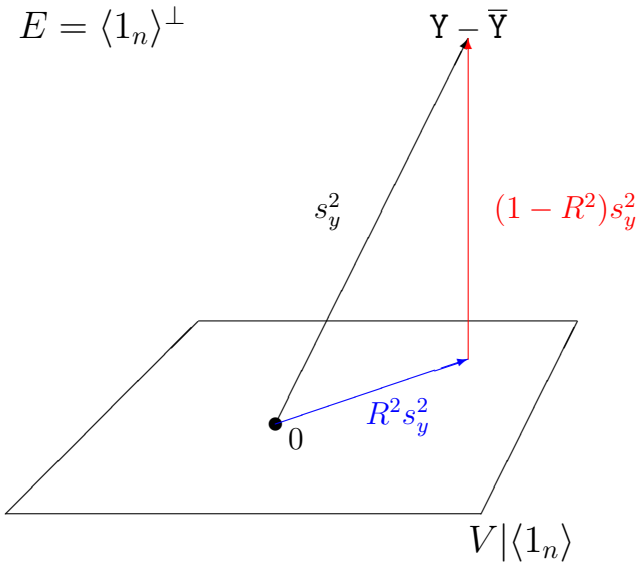


Figura 2.5: Descomposición de la varianza muestral

▪ *Ejercicio* 34. Resolver el contrate de la hipótesis inicial $H_0 : \nu = 0$ en el ejemplo 1; resolver el contrate de la hipótesis inicial $H_0 : \mu_1 = \mu_2$ en el ejemplo 2.

▪ *Ejercicio* 35. Probar que el test anova que resuelve el contraste $H_0 : \mu_1 = \ldots = \mu_r$ en el ejemplo 3 consiste en confrontar con la distribución $F_{r-1,n-r}$ el estadístico de contraste

$$F = \frac{SCH/(r-1)}{SCE/(n-r)}, \tag{2.31}$$

donde

$$SCH = \sum_i n_i (\overline{\mathtt{Y}}_{i\cdot} - \overline{\mathtt{y}}_{\cdot\cdot})^2, \tag{2.32}$$

$$SCE = \sum_i \sum_j (\mathtt{Y}_{ij} - \overline{\mathtt{y}}_{i\cdot})^2 \tag{2.33}$$

▪ *Ejercicio* 36. Probar que el test anova que resuelve el contrate para $W = \langle 1_n \rangle$ consiste en confrontar con la distribución $F_{\dim V - 1, n - \dim V}$ el estadístico de contraste

$$F = \frac{n - \dim V}{\dim V - 1} \cdot \frac{R^2}{1 - R^2}. \tag{2.34}$$

con R^2 definido según (2.30). En particular, para contrastar $H_0 : \underline{\beta} = 0$ en el ejemplo 4 se utiliza el estadístico

$$F = \frac{n - (q+1)}{q} \cdot \frac{R^2}{1 - R^2}. \tag{2.35}$$

¿Qué sucede en el caso particular $q = 1$?

▪ *Ejercicio* 37. En las condiciones del ejemplo 4, ¿qué distribución sigue bajo la hipótesis inicial $H_0 : \beta_q = 0$ el estadístico de contraste del anova correspondiente? Probar que el estadístico de contraste correspondiente puede expresarse de manera análoga a (2.35) pero mediante el

coeficiente de correlación parcial muestral entre Y y $\mathsf{Z}[q]$, dados $\mathsf{Z}[1], \ldots \mathsf{Z}[q-1]$, que se denota por $r_{\mathsf{y},\mathsf{z}[q]\cdot(\mathsf{z}[1],\ldots,\mathsf{z}[q-1])}$, o abreviadamente por r_{q*}, mediante

$$F = [n - (q+1)] \cdot \frac{r_{q*}^2}{1 - r_{q*}^2} \tag{2.36}$$

▪ *Ejercicio* 38. Probar que[2] el estadístico de contraste (2.37) puede expresarse también a través de $\hat{\beta}_q$ y del coeficiente de correlación múltiple de $\mathsf{Z}[q]$ respecto al resto de variables explicativas $\mathsf{Z}[1], \ldots, \mathsf{Z}[q-1]$, que se denota abreviadamente por R_{q*}, mediante

$$F = n \cdot \frac{\hat{\beta}_q^2}{\hat{\sigma}^2 \cdot s_{\mathsf{Z}[q]}^{-2}} \cdot (1 - R_{q*}^2) \tag{2.37}$$

El término $1 - R_{q*}^2$ que aparece a la derecha en (2.37) se denomina tolerancia y su inverso, factor de inflación de la varianza (FIV). (2.37) y (2.36) vienen a expresar pues , en términos intuitivos, que una redundancia lineal entre las variables explicativas, que a su vez puede asociarse a bajas correlaciones parciales, resta fiabilidad a las estimaciones de los coeficientes de regresión.

Por otra parte, teniendo en cuenta que, en las condiciones del ejemplo 3, la hipótesis inicial $H_0 : \mu_1 = \ldots = \mu_r$ equivale a $\underline{\beta} = 0$ para cualquier parametrización del modelo mediante variables dummys, se sigue de (2.35) que la decisión al respecto se puede expresar a través de la correlación múltiple R^2 de Y con dichas variables dummys, independientemente de cuáles se elijan. Esto hecho, que tendrá lugar igualmente en el modelo multivariante, justifica el estudio de los coeficientes de correlación canónicos.

En la salida de SPSS recogida en el cuadro 2.1 podemos apreciar muchos de los ingredientes estudiados en la sección.

Variable dependiente: sepleng

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	63,212[a]	2	31,606	119,265	,000
Intersección	5121,682	1	5121,682	19326,505	,000
species	63,212	2	31,606	119,265	,000
Error	38,956	147	,265		
Total	5223,850	150			
Total corregida	102,168	149			

a. R cuadrado = ,619 (R cuadrado corregida = ,614)

Cuadro 2.1: Tabla anova; sepleng vs especies en irisdata.sav

▪ *Ejercicio* 39. Construye mediante SPSS dos variables dummys para distinguir las tres especies de flores de irisdata y comprueba que el coeficiente de correlación múltiple R^2 entre sepleng y dichas variables es el que aparece en la tabla 2.1.

[2] Ver Manual de Modelos Lineales, Montanero (2008).

▪ *Ejercicio* 40. En el cuadro 2.2 aparece el resultado del anova para comparar los valores medios de glucemia de cuatro categorías de recién nacidos (control, respiratoria, metabólica y mixta) mediante un diseño equilibrado. Relacionar los valores que aparecen en dicha tabla con las columnmas de la matriz de datos del cuadro 2.6, donde a los datos originales se les ha añadido las proyecciones relacionadas con la descomposición (2.24), las sumas de cuadrados correspondientes a las descomposición (2.28) y las variables dummys asociadas a la parametrización del ejercicio 22.

Pruebas de los efectos inter-sujetos

Variable dependiente:Nivel de glucemia en el cordón umbilical

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	643,930[a]	3	214,643	4,557	,009
Intersección	170156,250	1	170156,250	3612,563	,000
Enfermedad	643,930	3	214,643	4,557	,009
Error	1507,240	32	47,101		
Total	172307,420	36			
Total corregida	2151,170	35			

a. R cuadrado = ,299 (R cuadrado corregida = ,234)

Cuadro 2.2: Tabla anova; glucemia vs enfermedad en acidosis-SPSS.sav

2.3. Modelo multivariante

Una vez repasado el modelo lineal univariante estamos en condiciones de generalizarlo al caso multivariante, en el cual no contamos con una sino con p variables respuesta. Previamente, debemos introducir con brevedad las distribuciones de probabilidad asociadas al nuevo modelo. Para un estudio más detallado, consultar Arnold (1981), Anderson (1958) y Mardia et al. (1979). En lo que sigue supondremos en todo momento $\Sigma > 0$ y $n \geq p$.

2.3.1. Distribuciones asociadas al modelo

Seguimos el mismo esquema que en el caso unidimensional, con la salvedad de que no existe una distribución que generalice unívocamente la distribución F. Debemos tener en cuenta que, en nuestro modelo estadístico, la observación es una matriz $\mathtt{Y} \in \mathcal{M}_{n\times p}$ de datos como la que aparece en el cuadro 1.1, que se denotará como en (1.1).

Distribución normal matricial: Se trata de la distribución de partida del modelo, al igual que la normal esférica lo era en el caso univariante. Dados $\mu \in \mathcal{M}_{n\times p}$ y $\Sigma \in \mathcal{M}_{p\times p}$ simétrica y definida positiva, se dice que $\mathtt{Y} \sim N_{n,p}(\mu, \mathtt{Id}, \Sigma)$ cuando $\mathtt{Y}_i \sim N_p(\mu_i, \Sigma)$, $i = 1, \ldots, n$, siendo todas independientes. Esta distribución es un caso particular de otra más general que se trata

	Enfermedad	Glucemia	W	W_ortog	V_W	V_ortog	Dummy_1	Dummy_2	Dummy_3
1	Control	67,00	68,75	-1,75	-4,32	2,57	1	0	0
2	Control	59,60	68,75	-9,15	-4,32	-4,83	1	0	0
3	Control	57,30	68,75	-11,45	-4,32	-7,13	1	0	0
4	Control	64,00	68,75	-4,75	-4,32	-,43	1	0	0
5	Control	79,00	68,75	10,25	-4,32	14,57	1	0	0
6	Control	70,30	68,75	1,55	-4,32	5,87	1	0	0
7	Control	52,80	68,75	-15,95	-4,32	-11,63	1	0	0
8	Control	67,40	68,75	-1,35	-4,32	2,97	1	0	0
9	Control	62,50	68,75	-6,25	-4,32	-1,93	1	0	0
10	Respiratoria	68,20	68,75	-,55	2,08	-2,63	0	1	0
11	Respiratoria	70,40	68,75	1,65	2,08	-,43	0	1	0
12	Respiratoria	78,00	68,75	9,25	2,08	7,17	0	1	0
13	Respiratoria	72,20	68,75	3,45	2,08	1,37	0	1	0
14	Respiratoria	64,00	68,75	-4,75	2,08	-6,83	0	1	0
15	Respiratoria	82,40	68,75	13,65	2,08	11,57	0	1	0
16	Respiratoria	65,70	68,75	-3,05	2,08	-5,13	0	1	0
17	Respiratoria	65,80	68,75	-2,95	2,08	-5,03	0	1	0
18	Respiratoria	70,80	68,75	2,05	2,08	-,03	0	1	0
19	Metabólica	73,30	68,75	4,55	5,92	-1,37	0	0	1
20	Metabólica	81,30	68,75	12,55	5,92	6,63	0	0	1
21	Metabólica	70,20	68,75	1,45	5,92	-4,47	0	0	1
22	Metabólica	73,60	68,75	4,85	5,92	-1,07	0	0	1
23	Metabólica	73,60	68,75	4,85	5,92	-1,07	0	0	1
24	Metabólica	74,30	68,75	5,55	5,92	-,37	0	0	1
25	Metabólica	73,30	68,75	4,55	5,92	-1,37	0	0	1
26	Metabólica	77,20	68,75	8,45	5,92	2,53	0	0	1
27	Metabólica	75,20	68,75	6,45	5,92	,53	0	0	1
28	Mixta	50,90	68,75	-17,85	-3,68	-14,17	-1	-1	-1
29	Mixta	71,20	68,75	2,45	-3,68	6,13	-1	-1	-1
30	Mixta	60,20	68,75	-8,55	-3,68	-4,87	-1	-1	-1
31	Mixta	71,20	68,75	2,45	-3,68	6,13	-1	-1	-1
32	Mixta	60,80	68,75	-7,95	-3,68	-4,27	-1	-1	-1
33	Mixta	78,70	68,75	9,95	-3,68	13,63	-1	-1	-1
34	Mixta	61,00	68,75	-7,75	-3,68	-4,07	-1	-1	-1
35	Mixta	57,40	68,75	-11,35	-3,68	-1,00	-1	-1	-1
36	Mixta	74,20	68,75	5,45	-3,68	9,13	-1	-1	-1
37	.	172307,42	170156,25	2151,17	643,93	1507,24	.	.	.

Figura 2.6: Proyecciones en acidosis-SPSS.sav

con detalle en Arnold (1981). La función de densidad se define, para cada matriz $\mathtt{X} \in \mathcal{M}_{n\times p}$, mediante

$$f(\mathtt{x}) = \frac{1}{(2\pi)^{np}|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}[(\mathtt{X}-\mu)\Sigma^{-1}(\mathtt{X}-\mu)']\right\} \tag{2.38}$$

Distribución de Wishart: Generaliza la distribución χ^2. Dado $\mathtt{Y} \sim N_{n,p}(\mu, \mathtt{Id}, \Sigma)$, puede probarse que la distribución de $\mathtt{Y}'\mathtt{Y}$ depende de μ a través de $\mu'\mu$. Teniendo en cuenta eso y dado $E \subset \mathbb{R}^n$, se define la distribución de Wishart mediante $Y'P_EY \sim W_p(\dim E, \delta, \Sigma)$, con $\delta = \mu'P_E\mu$. Si $\delta = 0$ se denota $W_p(\dim E, \Sigma)$. Las propiedades de la distribución de Wishart son por completo análogas a la de la distribución χ^2 y se estudian con detalle en Arnold (1981).

- *Ejercicio* 41. Comprobar que $W_1(m, \delta, \sigma^2) = \sigma^2\chi^2_m(\delta/\sigma^2)$

Distribución T^2 de Hotelling: Dados $X \sim N_p(\nu, \Sigma)$ y $W \sim W_p(m, \Sigma)$ independientes, se define la distribución T^2-Hotelling mediante

$$mX'W^{-1}X \sim T^2_{p,m}(\delta), \qquad \delta = \nu'\Sigma^{-1}\nu \tag{2.39}$$

En el caso $\delta = 0$ se denota $T^2_{p,m}$. En Arnold (1981) se prueba que esta distribución no es en esencia nueva, sino que se identifica, salvo un factor escala, con un modelo tipo F, lo cual garantiza que está bien definida. Concretamente

$$T^2_{p,m}(\delta) = \frac{mp}{m-p+1} F_{p,m-p+1}(\delta) \tag{2.40}$$

En particular, se verifica que $T^2_{1,m} = t^2_m$, por lo que debemos entender la distribución T^2 una generalización en sentido estadístico de la distribución t^2. Es decir, que se utilizará en aquellos problemas multivariantes cuyos análogos univariantes precisen de la distribución t-Student, concretamente, en el contraste de hipótesis del tipo $H_0 : \mu \in W$ con $\dim V|W = 1$. Veremos que en tales casos el estadístico de contraste puede entenderse geométricamente como una distancia de Mahalanobis. Además, puede probarse que $T^2_{p,m}$ converge en distribución a χ^2_p conforme m tiende a infinito.

Distribuciones de Wilks, Lawley-Hotelling, Roy y Pillay: Pueden entenderse como cuatro formas diferentes de generalizar la distribución F en el caso multivariante. Se estudian con detalle en Arnold (1981). Al igual que ocurre con la distribución F, convergen en distribución a $\chi^2_{p\cdot\dim V|W}$ conforme aumenta el segundo grado de libertad, por lo cual omitiremos aquí su estudio.

2.3.2. El modelo y ejemplos

Dada una matriz $A \in \mathcal{M}_{n\times p}$ y $E \subset \mathbb{R}^n$, se denota $A \in E$ cuando cada columna de A pertenece al subespacio E. Dicho esto, el modelo lineal normal multivariante viene dado por una matriz de datos $\mathtt{Y} \sim N_{n,p}(\mu, \mathtt{Id}, \Sigma)$, con $\Sigma > 0$ y la restricción $\mu \in V$ para algún $V \subset \mathbb{R}^n$ conocido. Por lo tanto, $\mathtt{Y}$ constituye una matriz como la que aparece en el cuadro 1.1 que recoge una muestra (en sentido amplio) de n observaciones $\mathtt{Y}_i \sim N_p(\mu_i, \Sigma)$ independientes. Si consideramos una base $\mathtt{X}$ de V, el modelo puede parametrizarse también de la forma

$$\mathtt{Y} = \mathtt{X}\beta + \mathcal{E}, \qquad \mathcal{E} \sim N_{n,p}(0, \mathtt{Id}, \Sigma),\ \beta \in \mathcal{M}_{\dim V\times p},\ \Sigma > 0 \tag{2.41}$$

Los cuatro problemas univariantes (ejemplos 1-4) considerados en el apartado 2.2.2 se generalizan al caso multivariante dando lugar a los siguientes problemas estadísticos multivariantes que se estudiarán con más detalle en el siguiente capítulo. Basta tener en cuenta que la variable respuesta Y se convierte en este caso en un vector respuesta p-dimensional de componentes $Y[1], \ldots, Y[p]$.

Ejemplo 5. [Muestra aleatoria simple de una distribución p-normal] Consideremos $\mathtt{Y}_1, \ldots, \mathtt{Y}_n$ iid $N_p(\nu, \Sigma)$. En ese caso, la matriz aleatoria $\mathtt{Y} = (\mathtt{Y}_1, \ldots, \mathtt{Y}_n)'$ sigue un modelo de distribución $N_{n,p}(\mu, \mathtt{Id}, \Sigma)$ con $\mu \in V = \langle 1_n \rangle$ y $\Sigma > 0$. Efectivamente, se verifica que cada columna $\mu[j]$ de μ, que corresponde a la componente $Y[j]$ del vector Y, pertenece a V.

Ejemplo 6. [Muestras independientes de p-normales con idénticas matrices de covarianzas] Consideremos, para $i = 1, 2$, sendas muestras independientes $\mathtt{Y}_{i1}, \ldots, \mathtt{Y}_{in_i}$ iid $N_p(\mu_i, \Sigma)$. Si se denota $n = n_1 + n_2$ e $\mathtt{Y} = (\mathtt{Y}_{11}, \ldots, \mathtt{Y}_{2n_2})'$, se verifica que $\mathtt{Y} \sim N_{n,p}(\mu, \mathtt{Id}, \Sigma)$ con $\Sigma > 0$ y $\mu \in V = \langle \mathtt{v}_1, \mathtt{v}_2 \rangle$.

Ejemplo 7. [Diseño completamente aleatorizado multivariante] Se generaliza el caso univariante como en los ejemplos 5 y 6.

Ejemplo 8. [Regresión lineal multivariante] A diferencia del ejemplo 4 univariante, se pretende explicar p variables respuesta, $Y[1], \ldots, Y[p]$, a partir de q variables explicativas, lo cual nos lleva a un modelo tipo (2.41) donde $\mathtt{Y}$ es la matriz $n \times p$ de observaciones respuesta, expresada como en (1.1), $\mathcal{E}$ la matriz $n \times p$ de errores, $\mathtt{X}$ es la misma matriz que aparece en (2.15) y β es la matriz $(q+1) \times p$ siguiente

$$\beta = \left(\begin{array}{ccc} \beta_0[1] & \ldots & \beta_0[p] \\ \hline \beta_1[1] & \ldots & \beta_1[p] \\ \vdots & \ldots & \vdots \\ \beta_q[1] & \ldots & \beta_q[p] \end{array}\right) = \left(\begin{array}{c} \beta_0' \\ \hline \beta_1' \\ \vdots \\ \beta_q' \end{array}\right) \tag{2.42}$$

Para cada coeficiente $\beta_i[j]$, el subíndice i y el índice entre corchetes j indican, respectivamente, a qué vector explicativo y a qué vector respuesta hace referencia. La primera fila, relativa al término independiente, se denota por β_0, y el resto de la matriz por $\underline{\beta}$.

Al igual que en el caso univariante, un problema como el del ejemplo 7 puede parametrizarse de idéntica forma mediante variables dummys para convertirse en un problema de regresión lineal multivariante, donde el contraste de la igualdad de las r medias equivale al contraste total de la hipótesis $\underline{\beta} = 0$.

Estos cuatro problemas se abordarán con más detalle en el siguiente capítulo. A continuación estudiaremos brevemente la solución teórica a los problemas de estimación y contraste de hipótesis.

2.3.3. Estimación y contraste de hipótesis

Los estimadores de μ y σ^2 en el modelo univariante pueden generalizarse de manera natural mediante

$$\hat{\mu} = P_V \mathtt{Y}, \tag{2.43}$$

$$\hat{\Sigma} = \frac{1}{n - \dim V} \mathtt{Y}' P_{V^\perp} \mathtt{Y} \tag{2.44}$$

Puede probarse que, así definidos, $\hat{\mu}$ y $\hat{\Sigma}$ son EIMV de μ y Σ y que, si reemplazamos en $\hat{\Sigma}$ el denominador $n - \dim V$ por n, constituyen el EMV. El estimador μ consiste en estimar la media de las distintas componentes por separado. Si el modelo está parametrizado de la forma (2.41), el estimador de β será igualmente

$$\hat{\beta} = (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'\mathtt{Y} \tag{2.45}$$

- *Ejercicio* 42. Probar que $(n - \dim V)\hat{\Sigma} \sim W_p(n - \dim V, \Sigma)$

En lo referente al contraste de hipótesis tipo $H_0 : \mu \in W$, las afirmaciones de los ejercicios (28), (29) y (30) pueden extenderse trivialmente al caso multivariante. El test que resuelve el contraste se denomina manova.

Manova: Consideraremos nuevamente la descomposición ortogonal $\mathbb{R}^n = W \oplus V | W \oplus V^\perp$, que da pie a definir las siguientes matrices $p \times p$ simétricas y semidefinidas positivas:

$$\mathtt{S}_1 = \mathtt{Y}' P_W \mathtt{Y}, \qquad \mathtt{S}_2 = \mathtt{Y}' P_{V|W} \mathtt{Y}, \qquad \mathtt{S}_3 = \mathtt{Y}' P_{V^\perp} \mathtt{Y} \tag{2.46}$$

▪ *Ejercicio* 43. Probar que,las tres matrices aleatorias siguen las siguientes distribuciones, independientes entre sí:

$$\mathsf{S}_1 \sim W_p(\dim W, \mu' P_W \mu, \Sigma), \quad \mathsf{S}_2 \sim W_p(\dim V|W, \mu' P_{V|W} \mu, \Sigma), \quad \mathsf{S}_3 \sim W_p(n - \dim V, \Sigma) \qquad (2.47)$$

Las matrices S_i, $i = 1, 2, 3$ en (2.46), de dimensión $p \times p$ y semidefinidas positivas, pueden expresarse mediante $\mathsf{Z}_i'\mathsf{Z}_i$, siendo $\mathsf{Z}_i = G_i'\mathsf{Y}$ y G_i una matriz de n filas cuyas columnas constituyan una base ortonormal de W, $V|W$ y $V^\perp$, respectivamente. En lo sucesivo prescindiremos de S_1 por argumentos de invarianza. Es más, la aplicación del Principio de Invarianza de manera análoga al caso univariante nos conduce a desechar todo test cuyo estadístico de contraste no pueda expresarse en función de los autovalores de $\mathsf{S}_3^{-1}\mathsf{S}_2$, que se denotan de mayor a menor mediante $t_1, \ldots, t_p$. Si los consideramos como vector aleatorio su distribución exacta se calcula partiendo de (2.47). En lo sucesivo, se denotará

$$b = \text{mín}\{p, \dim V|W\} \qquad (2.48)$$

▪ *Ejercicio* 44. Probar que p autovalores reales y no negativos de la matriz simétrica y semidefinida postiva $\mathsf{Z}_2\mathsf{S}_3^{-1}\mathsf{Z}_2'$ los son a su vez de la matriz $\mathsf{S}_3^{-1}\mathsf{S}_2$. Además, si $p > b$, los últimos $p-b$ autovalores de ambas matrices son necesariamente nulos. Lo mismo puede decirse de $(\theta_1, \ldots, \theta_p)$, definidos como los autovalores de la matriz $\Sigma^{-1}\mu' P_{V|W}\mu$, de los cuales $(t_1, \ldots, t_b)$ pueden considerarse estimadores. Nótese que la hipótesis inicial $H_0 : \mu \in W$ equivale a $\theta_1 = \ldots = \theta_b = 0$.

Así pues, el Principio de Invarianza nos conduce a considerar sólo los tests que se construyan a partir de los b primeros autovalores de $\mathsf{S}_3^{-1}\mathsf{S}_2$, $(t_1, \ldots, t_b)$. En el capítulo 3 se verá un interpretación precisa de estos autovalores. Sólo en el caso $b = 1$ estaremos en condiciones de formular directamente un test basado en la distribución de t_1. Se da tal situación cuando $p = 1$ o $\dim V|W = 1$:

(i) Si $p = 1$ las matrices de (2.46) son números positivos y t_1 es, salvo una constante, el estadístico F. Se trata pues del propio anova.

(ii) Si $\dim V|W = 1$ puede probarse que t_1 sigue, salvo una constante, una distribución T^2-Hotelling, lo cual permite formular un test UMP-invariante y de razón de verosimilitudes. Si, además, $p = 1$, estaremos hablando del test de Student.

Dado que en el caso $b > 1$ el Principio de Invarianza no propicia una simplificación completa de la información, el problema se ha abordado históricamente acogiéndose a otros diferentes principios estadísticos que conducen a respectivas soluciones razonables, que pueden expresarse a partir de los mencionados autovalores, es decir, que son invariantes. De esta manera aparecen en la literatura estadística cuatro tests diferentes: el test de Wilks, que es el TRV y, por lo tanto, responde al Principio de Máxima Verosimilitud; el test de Lawley-Hotelling, que responde al Principio de Sustitución (o método *plug-in*); el test de Pillay, que sigue la línea del test anterior pero utilizando los coeficientes de correlación canónica (ver apartado 3.4.1), y por último el de Roy, que obedece al Principio de Unión-Intersección[3]. La formulación precisa de estos tests como funciones de los autovalores $t_1, \ldots, t_b$ se comentará brevemente en el apartado 3.4.1. No obstante, nos centraremos aquí en el test de Wilks por dos razones: por obedecer al Principio de Máxima Verosimilitud que es, posiblemente, el más intuitivo en Estadística, y porque facilita

[3]Los detalles de los principios mencionados los podemos encontrarlos en Arnold (1981) o en el Manual 59 citado en la bibliografía.

el algoritmo de selección de variables en regresión lineal, lo cual es especialmente interesante en el análisis discriminante lineal.

De (2.38) se sigue que el estadístico de contraste del test de Wilks, es decir, la razón de verosimilitudes, es la siguiente:

$$\lambda(\mathsf{Y}) = \frac{|\mathsf{S}_3|}{|\mathsf{S}_2 + \mathsf{S}_3|} \tag{2.49}$$

- *Ejercicio* 45. Probar que $\lambda(\mathsf{Y})$ puede expresarse a través de $t_1, \ldots, t_b$ mediante

$$\lambda(\mathsf{Y}) = \prod_{i=1}^{b} (1 + t_i)^{-1} \tag{2.50}$$

Se demuestra en Arnold (1981) o, también, en el Manual 59, que bajo la hipótesis nula, $-(n - \mathtt{dim} V) \log \lambda$ converge en distribución a $\chi^2_{p \cdot \dim V|W}$ cuando n tiende a infinito. Este resultado es incluso cierto aunque no se respete el supuesto de normalidad, siempre y cuando el diseño de la muestra respete ciertas condiciones razonables. Los otros tres tests recogidos por la literatura (Lawley-Hotelling, Pillay y Roy) también convergen asintóticamente a la distribución $\chi^2_{p \cdot \dim V|W}$. En definitiva, para muestras de gran tamaño utilizaremos la distribución χ^2 como referencia, aunque el programa SPSS puede trabajar con otras aproximaciones a la distribución F. En el cuadro 2.3 podemos apreciar un esquema explicativo.

Exacta	$p = 1$	$p \geq 1$
$\dim V\|W = 1$	$t^2_{n-\dim \mathtt{v}}$	$T^2_{p,n-\dim \mathtt{v}}$
$\dim V\|W \geq 1$	$F_{\dim \mathtt{v}\|\mathtt{w},n-\dim \mathtt{v}}$	$\mathrm{Wilks}_{p,\dim \mathtt{v}\|\mathtt{w},n-\dim \mathtt{v}}$

$\xrightarrow{\mathtt{n}\uparrow}$

Asintótica	$p = 1$	$p \geq 1$
$\dim V\|W = 1$	χ^2_1	χ^2_p
$\dim V\|W \geq 1$	$\chi^2_{\dim \mathtt{v}\|\mathtt{w}}$	$\chi^2_{p \cdot \dim \mathtt{v}\|\mathtt{w}}$

Cuadro 2.3: Distribuciones en los contrastes de la media en función de las dimensiones del modelo lineal V y de la hipótesis inicial W, del número p de variables y del tamaño n de la muestra.

También se recogen en Arnold (1981), Dillon y Goldstein (1984), Flury (1996) y Rencher (1995), entre otras referencias, diversos tests para contrastes de hipótesis relativos a la matriz de covarianzas implementados en los programas estadísticos, como el test M de Box, el de esfericidad de Barlett y algunos otros, que no abordamos aquí por brevedad y dado que son sensibles ante la violación del supuesto de normalidad.

Capítulo 3

Aplicaciones del modelo

En este capítulo desarrollaremos los cuatro problemas estadísticos formulados en los ejemplos 5-8 de la página 33 del capítulo anterior, cuyo denominador común es que se formalizan mediante el modelo lineal multivariante. Añadimos además un apartado dedicado al análisis de correlación canónica, relacionado directamente con el problema de regresión lineal multivariante, y una sección dedicada al análisis de perfiles, relacionado con los tres problemas restantes. Por último, ilustraremos con un ejemplo algunas de las técnicas estudiadas. En los distintos casos se aplicarán los métodos teóricos de estimación y contraste de hipótesis expuestos en el capítulo anterior. Se da por supuesto que el lector conoce ya las técnicas univariante análogas (test de Student para muestras independientes y relacionadas, anova de una vía y estudio de regresión lineal múltiple), que puede consultar, por ejemplo, en Peña (2010). A lo largo del capítulo se hará uso del siguiente resultado, comúnmente conocido como teorema de los multiplicadores finitos de Langrange, que se deduce del teorema de la función implícita y permite obtener valores extremos para una función definida en $\mathbb{R}^p$ bajo una serie de restricciones.

Lema 3.0.1. Sean $k < p$ enteros y ϕ y f funciones derivables de $\mathbb{R}^p$ en $\mathbb{R}$ y $\mathbb{R}^k$, respectivamente, tales que existe máx$\{\phi(x) : \ f(x) = 0\}$ alcanzándose en $c \in \mathbb{R}^p$ tal que $\bigtriangledown f(c) \neq 0$. Entonces, existe $\eta \in \mathbb{R}^k$ tal que $\bigtriangledown(\phi - \eta' f)(c) = 0$.

3.1. Inferencia para una media

Desarrollamos aquí el ejemplo 5 de la página 33. Partimos pues de una muestra aleatoria simple de una distribución p-normal, es decir,

$$\mathtt{Y}_1, \ldots, \mathtt{Y}_n \text{ iid } N_p(\nu, \Sigma) \tag{3.1}$$

de tal forma que la matriz de datos $\mathtt{Y}$ sigue un modelo de distribución $N_{n,p}(\mu, \mathtt{Id}, \Sigma)$ con $\mu \in V = \langle 1_n \rangle$ y $\Sigma > 0$. Denótese por $\overline{\mathtt{y}}$ el vector de medias $(\overline{\mathtt{y}}[1], \ldots, \overline{\mathtt{y}}[p])'$ y por S la matriz de covarianzas muestral. Podemos probar entonces los siguientes resultados.

Proposición 3.1.1. Los EIMV de μ y Σ son $\hat{\mu} = 1_n \cdot \overline{\mathtt{y}}'$ y $\hat{\Sigma} = \frac{n}{n-1} S$, respectivamente.

Proposición 3.1.2. $n(\overline{\mathtt{y}} - \nu)' \hat{\Sigma}^{-1} (\overline{\mathtt{y}} - \nu) \sim T^2_{p,n-1}$

De la proposición 3.1.2 se sigue que el siguiente conjunto de $\mathbb{R}^p$ es una región de confianza a nivel $1 - \alpha$ para la media ν.

$$\mathcal{E}_\alpha(\mathtt{Y}) = \left\{ x \in \mathbb{R}^p : \ n(\overline{y} - x)' \hat{\Sigma}^{-1} (\overline{y} - x) \leq T^{2,\alpha}_{p,n-1} \right\} \tag{3.2}$$

Esta región geométrica es un elipsoide cuyo centro es $\overline{\mathrm{y}}$ y cuya forma viene dada por $\hat{\Sigma}$. Si pretendemos contrastar la hipótesis inicial $H_0 : \nu = 0$, que equivale a $\mu \in W = 0$, la proposición 3.1.2 invita confrontar con la distribución $T^2_{p,n-1}$ el estadístico de contraste

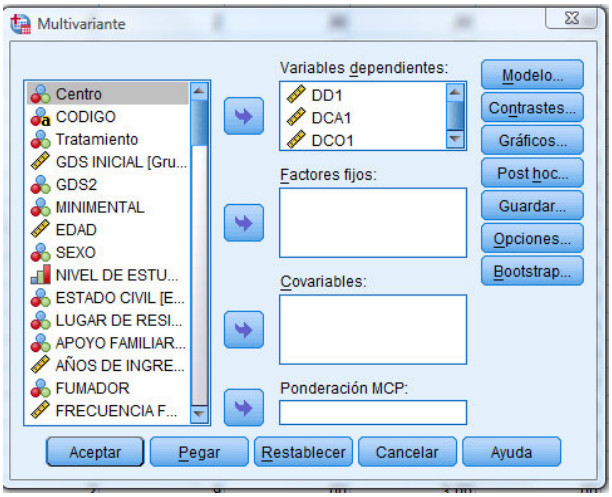
$$T^2(\mathrm{Y}) = n \cdot \overline{\mathrm{y}}' \hat{\Sigma}^{-1} \overline{\mathrm{y}} \tag{3.3}$$

Tanto el elipsoide (3.2) como el estadístico de contraste (3.3) pueden expresarse en términos de la distancia de Mahalanobis $D^2_{\hat{\Sigma}}$ definida en (1.38). Concretamente,

$$T^2(\mathrm{Y}) = n \cdot D^2_{\hat{\Sigma}}(\overline{\mathrm{y}}, 0) \tag{3.4}$$

Éste es precisamente el test UMP-invariante y de razón de verosimilitudes que se propone en el capítulo anterior para este caso particular, donde se da la circunstancia de que $\dim V|W = 1$. Concretamente, con la notación introducida en dicho capítulo se verifica que t_1 y θ_1, entendidos como funciones sobre el espacio de observaciones y el de parámetros probabilísticos, son iguales salvo escalares a $D^2_{\hat{\Sigma}}(\overline{\mathrm{y}}, 0)$ y $D^2_{\Sigma}(\mu, 0)$, respectivamente.

- *Ejercicio* 46. Probar que $\hat{\Sigma}^{-1} > 0$
- *Ejercicio* 47. Probar que, en , $p = 1$, el test (3.3) es el de Student para una muestra.
- *Ejercicio* 48. En el cuadro 3.1 se muestra el resultado de aplicar el test 3.3 con tres variables y 36 individuos (se muestra el correspondiente cuadro de diálogo de SPSS). Interpretar la tabla en función de los conocimientos teóricos.



Por otra parte, del Teorema Central el Límite y la Ley Débil de los Grandes Números se sigue:

Proposición 3.1.3. Si $\mathrm{Y}_1, \ldots, \mathrm{Y}_n$ iid con media ν y componentes en L^2, entonces se verifica la siguiente convergencia en distribución:

$$\lim_{n\to\infty} nD^2_{\hat{\Sigma}}(\overline{\mathrm{y}}, \nu) = \chi^2_p \tag{3.5}$$

Cuadro 3.1: Tabla Manova una muestra

Contrastes multivariados[b]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
Intersección	Traza de Pillai	,802	44,582[a]	3,000	33,000	,000
	Lambda de Wilks	,198	44,582[a]	3,000	33,000	,000
	Traza de Hotelling	4,053	44,582[a]	3,000	33,000	,000
	Raíz mayor de Roy	4,053	44,582[a]	3,000	33,000	,000

a. Estadístico exacto
b. Diseño: Intersección

Cuadro 3.2: Cuadro de diálogos Manova para una muestra

Este resultado otorga validez asintótica al test propuesto aunque no se verifique el supuesto de normalidad. Nótese también que podemos construir una región de confianza a nivel $1-\alpha$ sin utilizar técnicas multivariantes, calculando para cada componente del vector respuesta Y un intervalo de confianzas a nivel $1-\alpha^*$ y componiendo entonces un rectángulo en dimensión p. El valor de α^* puede determinarse mediante de manera conservadora mediante la desigualdad de Bonferroni:

$$P\big(\bigcap_{i=1}^{m} A_i\big) \geq 1 - \sum_{i=1}^{m} P\big(A_i^c\big) \tag{3.6}$$

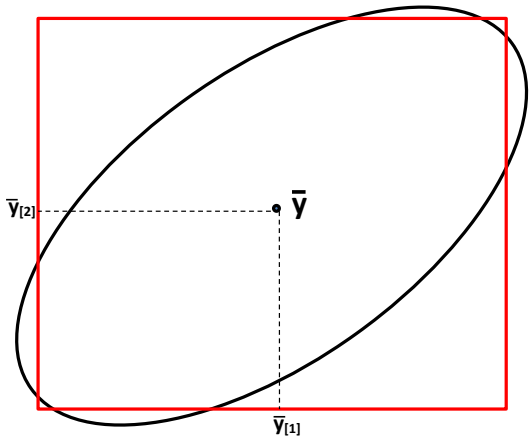


Figura 3.1: Rectángulo y elipse de confianza

El elipsoide (3.2) delimita una región del espacio de menor tamaño que el del rectángulo, siendo mayor su diferencia cuanto mayor sea la correlación entre las variables. Ello es debido a que el método univariante no hace uso en ningún momento de las covarianzas y, por lo tanto, emplea menos información que el multivariante. Si las componentes del vector aleatorio Y fueran incorreladas (independientes bajo el supuesto de p-normalidad) el rectángulo anterior podría construirse sin recurrir a la desigualdad de Bonferroni (3.6) y tendría un área similar al de la elipse, cuyos ejes coincidirían con los ejes de coordenadas. En ese caso no procedería el uso de métodos multivariantes.

3.2. Inferencia para dos medias

En esta sección desarrollamos el ejemplo 6 de la página 33. Se trata pues de estudiar la posible relación entre un vector respuesta p-dimensional Y y un factor cualitativo que distingue dos categorías. Partimos de dos muestras independientes de sendas distribuciones p-normales con matriz de covarianzas común

$$\begin{cases} \mathtt{Y}_{11}, \ldots, \mathtt{Y}_{1n_1} \text{ iid} & N_p(\mu_1, \Sigma) \\ \mathtt{Y}_{21}, \ldots, \mathtt{Y}_{2n_2} \text{ iid} & N_p(\mu_2, \Sigma) \end{cases} \tag{3.7}$$

La matriz de datos $\mathtt{Y}$ sigue un modelo de distribución $N_{n_1+n_2,p}(\mu, \mathtt{Id}, \Sigma)$ con $\mu \in V = \langle \mathtt{v}_1, \mathtt{v}_2 \rangle$ y $\Sigma > 0$.

- *Ejercicio* 49. Construir los EIMV de μ y Σ a partir de las medias aritméticas $\overline{\mathtt{y}}_1, \overline{\mathtt{y}}_2 \in \mathbb{R}^p$ de ambas muestras.

- *Ejercicio* 50. Probar que

$$\frac{n_1 n_2}{n_1 + n_2} \cdot D^2_{\hat{\Sigma}}(\overline{\mathtt{y}}_1, \overline{\mathtt{y}}_2) \sim T^2_{p,n_1+n_2-2}(\theta), \qquad \theta = D^2_{\Sigma}(\mu_1, \mu_2) \tag{3.8}$$

Si pretendemos contrastar la hipótesis inicial $H_0 : \mu_1 = \mu_2$, (3.8) invita a confrontar con la distribución T^2_{p,n_1+n_2} el estadístico de contraste

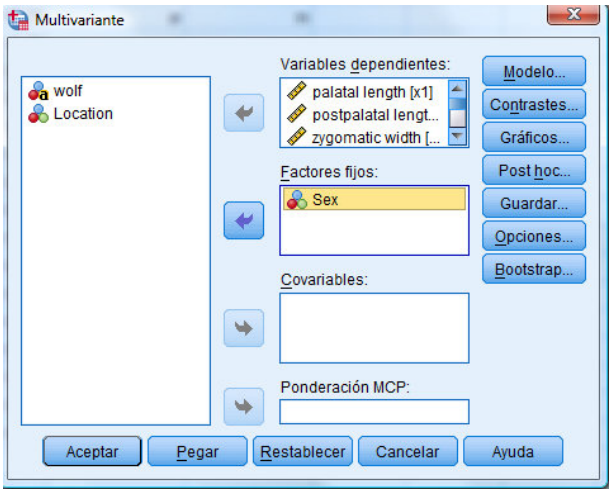
$$T^2(\mathtt{Y}) = \frac{n_1 n_2}{n_1 + n_2} \cdot D^2_{\hat{\Sigma}}(\overline{\mathtt{y}}_1, \overline{\mathtt{y}}_2) \tag{3.9}$$

En eso consiste precisamente el test UMP-invariante y de razón de verosimilitudes que se propone en el capítulo anterior para este caso particular, donde se da la circunstancia también de que $\dim V|W = 1$. Concretamente, se verifica que t_1 y θ_1, entendidos como funciones sobre el espacio de observaciones y el de parámetros probabilísticos, son iguales salvo escalares a $D^2_{\hat{\Sigma}}(\overline{\mathtt{y}}_1, \overline{\mathtt{y}}_2)$ y θ, respectivamente.

Como en la sección anterior, estamos también en condiciones de garantizar la validez asintótica del test aunque no se verifique el supuesto de p-normalidad si $n_1, n_2 \to \infty$; también podemos garantizarla aunque no se verifique el supuesto de homocedasticidad si, además, $n_1/n_2 \to 1$. Si $p = 1$ el test propuesto es el conocido test de Student para dos muestras independientes.

La hipótesis $H_0 : \mu_1 = \mu_2$ podría contrastarse prescindiendo de técnicas multivariantes aplicando de manera independiente sendos tests de Student para cada una de las p componentes del vector respuesta Y. En ese caso, los niveles de significación de cada test deberían calcularse de manera conservadora mediante la desigualdad de Bonferroni. No podemos descartar que el método multivariante (3.9) aprecie diferencias significativas entre ambas medias mientras que ninguno de los tests de Student univariantes logre diferenciar las componentes de las mismas. Hemos de ser conscientes, nuevamente, de que el método multivariante hace uso de la información que aportan las covarianzas, lo cual no se tiene en cuenta en ninguno de los p tests de Student.

- *Ejercicio* 51. Interprétese en los términos anteriores la tabla que aparece en el cuadro 3.4, obtenida según se indica en el cuadro 3.3, que corresponde a un estudio comparativo efectuado a 25 lobos en los se que relacionan de 9 variables numéricas con el sexo.



Cuadro 3.3: Cuadro de diálogos Manova para dos muestras

Contrastes multivariados[b]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
Intersección	Traza de Pillai	1,000	6886,561[a]	9,000	15,000	,000
	Lambda de Wilks	,000	6886,561[a]	9,000	15,000	,000
	Traza de Hotelling	4131,937	6886,561[a]	9,000	15,000	,000
	Raíz mayor de Roy	4131,937	6886,561[a]	9,000	15,000	,000
Sex	Traza de Pillai	,784	6,038[a]	9,000	15,000	,001
	Lambda de Wilks	,216	6,038[a]	9,000	15,000	,001
	Traza de Hotelling	3,623	6,038[a]	9,000	15,000	,001
	Raíz mayor de Roy	3,623	6,038[a]	9,000	15,000	,001

a. Estadístico exacto
b. Diseño: Intersección + Sex

Cuadro 3.4: Tabla Manova dos muestras

Nótese, por otra parte, que la j-ésima componente del vector respuesta, $Y[j]$ es la proyección del vector Y sobre el j-ésimo eje de coordenadas. Si e_j denota un vector unitario que lo determina, podemos expresar $Y[j] = e_j' Y$. En general, para cada eje $\langle a \rangle$ con $\|a\| = 1$, podemos considerar la proyección $a'Y$ sobre $\langle a \rangle$ que da lugar, teniendo en cuenta (1.15), a dos muestras independientes

$$\begin{cases} a'\mathtt{Y}_{11}, \ldots, a'\mathtt{Y}_{1n_1} \text{ iid } N_1(a'\mu_1, a'\Sigma a) \\ a'\mathtt{Y}_{21}, \ldots, a'\mathtt{Y}_{2n_2} \text{ iid } N_1(a'\mu_2, a'\Sigma a) \end{cases} \tag{3.10}$$

y a una hipótesis inicial $H_0^a : a'\mu_1 = a'\mu_2$, que puede contrastarse a partir de los datos proyectados mediante el test de Student. Concretamente, se confronta con la distribución $t_{n_1+n_2-2}$ el estadístico de contrate $t_{\langle a \rangle}(\mathtt{Y})$ definido como $t(\mathtt{Y}a)$. Conocido $\mathtt{Y}$, debe existir necesariamente un eje $\langle a_1 \rangle$ que aporte un valor máximo para $t_{\langle a \rangle}(\mathtt{Y})$. Teniendo en cuenta (1.15) y el lema 3.0.1 obtenemos la solución concreta[1]

$$\langle a_1 \rangle = S_c^{-1}(\overline{\mathtt{y}}_1 - \overline{\mathtt{y}}_2), \qquad S_c = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2} \tag{3.11}$$

[1] S_c es el EIMV de Σ que se pedía calcular en el ejercicio 49.

Es más, si se denota

$$\mathtt{W}_{ij}[1] = a_1' \mathtt{Y}_{ij}, \qquad i = 1, 2,\ j = 1, \ldots, n_i \tag{3.12}$$

se verifica entonces que $t^2(\mathtt{W}[1]) = T^2(\mathtt{Y})$. En ese sentido podemos afirmar que distinguir las dos muestras en dimensión p es equivalente a distinguirlas en dimensión 1 sobre el eje $\langle a_1 \rangle$, que se denomina (primer) eje discriminante. El vector de proyecciones $\mathtt{W}[1] = \mathtt{Y}a_1$ se denomina vector de las (primeras) puntuaciones discriminantes. En la figura 3.2 el eje discriminante se representa con líneas discontinuas:

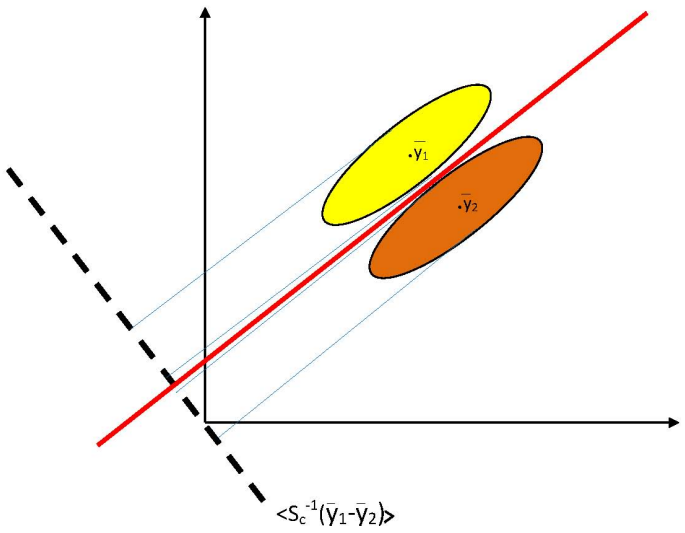


Figura 3.2: Eje discriminante

3.3. Manova de una vía

En esta sección desarrollaremos el ejemplo 7 de la página 33 y ampliaremos el concepto de eje discriminante. El problema supone una generalización del estudiado en la sección anterior, puesto que trata la relación entre un vector respuesta p-dimensional Y y un factor cualitativo que, en este caso, distingue entre $r \geq 2$ categorías. Por lo tanto, partimos de un diseño, denominado completamente aleatorizado, similar a (3.7) pero con r muestras independientes de n_i datos cada una. Mantendremos aquí la notación habitual del diseño de experimentos. En particular, n denotará la suma $\sum_{i=1}^{r} n_i$. La matriz de datos $\mathtt{Y}$ sigue entonces un modelo de distribución $N_n(\mu, \mathtt{Id}, \Sigma)$ con $\mu \in V = \langle \mathtt{v}_1, \ldots, \mathtt{v}_r \rangle$ y $\Sigma > 0$. La hipótesis inicial a contrastar en este caso es $H_0 : \mu_1 = \ldots = \mu_r$, que se corresponde con $\mu \in W = \langle 1_n \rangle$. Si $r > 2$ y $p > 1$ se verifica, a diferencia de los dos estudios anteriores, que b, según se define en (2.48), es mayor que 1.

A pesar de que, desde un punto de vista práctico, la comparación de 2 medias es un problema semejante a la comparaciones de $r \geq 3$ medias, el último estudio comporta una mayor complicación formal dado que no puede resolverse en términos de una distancia T^2 entre un único par de elementos. Por eso nos limitamos a aplicar la solución general del contraste expuesta en el capítulo anterior a este caso concreto: se obtienen $t_1 > \ldots > t_b > 0$, los autovalores

positivos[2] de $\mathsf{S}_3^{-1}\mathsf{S}_2$, donde S_2 y S_3 se calculan según (2.46) y, a partir de los mismos, obtenemos el valor del estadístico λ de Wilks definido según (2.50), cuya distribución exacta depende de $\theta_1 \geq \ldots \geq \theta_b \geq 0$, definidos en el ejercicio 44; por último, se confronta con la distribución $\chi^2_{p(r-1)}$ el valor $-(n-r)\log\lambda(\mathsf{Y})$.

En el caso $p = 1$ el test obtenido es el anova de una vía; en el caso $r = 2$ es el test (3.9); en general se denomina manova de una vía, que será asintóticamente válido aunque no se verifique el supuesto de normalidad si $n_1, \ldots, n_r$ tienden a infinito.

Desde este punto de vista, el problema de contrastar una hipótesis tipo $H_0 : \mu \in W$ se reduce a obtener las matrices S_2 y S_3 adecuadas. En este caso particular, pueden obtenerse trivialmente de manera similar a SCE y SCH en (2.31).

▪ *Ejercicio* 52. Probar que

$$\mathsf{S}_2 = \begin{pmatrix} SCH_{11} & \ldots & SCH_{1p} \\ \vdots & & \vdots \\ SCH_{1p} & \ldots & SCH_{pp} \end{pmatrix}, \quad \mathsf{S}_3 = \begin{pmatrix} SCE_{11} & \ldots & SCE_{1p} \\ \vdots & & \vdots \\ SCE_{1p} & \ldots & SCE_{pp} \end{pmatrix} \tag{3.13}$$

donde, para $h, k = 1, \ldots, p$,

$$SCH_{hk} = \sum_{i=1}^{r} n_i \big(\overline{\mathsf{y}}_{i\cdot}[h] - \overline{\mathsf{y}}_{\cdot\cdot}[h]\big) \cdot \big(\overline{\mathsf{y}}_{i\cdot}[k] - \overline{\mathsf{y}}_{\cdot\cdot}[k]\big) \tag{3.14}$$

$$SCE_{hk} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \big(\mathsf{Y}_{ij}[h] - \overline{\mathsf{y}}_{i\cdot}[h]\big) \cdot \big(\mathsf{Y}_{ij}[k] - \overline{\mathsf{y}}_{i\cdot}[k]\big) \tag{3.15}$$

Aunque no vamos a estudiar aquí diseños de experimentos multivariantes con dos o más factores, el lector debe percatarse de que, si es capaz de resolver el problema en el caso univariante, basta con proceder de manera análoga a (3.14) y (3.15) para obtener la solución general para el caso multivariante.

El interés de estas dos últimas secciones radica en la vinculación existente entre el manova de una vía (y también el test (3.9), entendido como caso particular) con el LDA (análisis discriminate lineal) de Fisher. Por otra parte, el problema de comparación de medias en un diseño completamente aleatorizado puede entenderse como un problema de regresión lineal, multivariante en este caso, respecto a $r - 1$ variables dummys de asignación a categorías, lo cual justifica a su vez el estudio del problema de regresión lineal multivariante que desarrollamos en la siguiente sección.

3.3.1. Ejes discriminantes

El concepto de eje discriminante introducido en la sección anterior puede ampliarse cuando el número de muestras es mayor que 2. Dado un eje $\langle a \rangle$ podemos considerar el estadístico de contraste $F_{\langle a \rangle}(\mathsf{Y})$ para la hipótesis inicial de igualdad de medias a partir de los datos proyectados sobre dicho eje. Nuestro primer objetivo es encontrar el eje $\langle a_1 \rangle$ que lo maximiza. En el caso $r = 2$ la solución es (3.11).

▪ *Ejercicio* 53. Probar que la solución general es el eje $\langle a_1 \rangle$ con

$$a_1 = \arg\max\{a'\mathsf{S}_2 a : \ a'\mathsf{S}_3 a = 1\} \tag{3.16}$$

[2]Nótese que descartamos la posibilidad de que dos autovalores sean iguales y de que alguno sea nulo. Se debe a que puede probarse que, bajo los supuestos del modelo, ello puede ocurrir con probabilidad nula.

▪ *Ejercicio* 54. Utilizando el lema 3.0.1 y teniendo en cuenta (1.22), probar que $F_{\langle a_1\rangle}(\mathtt{Y}) = \frac{n-r}{r-1}\cdot t_1$, siendo t_1 el primer autovalor de $\mathtt{S}_3^{-1}\mathtt{S}_2$ y a_1 un autovector asociado tal que $a_1'\mathtt{S}_3 a_1 = 1$.

De esta forma construimos el primer vector de puntuaciones discriminantes $\mathtt{W}[1] = \mathtt{Y}a_1$. El proceso puede continuar en principio hasta completar p ejes discriminantes con sus respectivas puntuaciones: el segundo eje discriminante $\langle a_2\rangle$ se define como aquél sobre el que debemos proyectar $\mathtt{Y}$ para obtener un vector de puntuaciones $\mathtt{W}[2] = \mathtt{Y}a_2$ incorrelado con $\mathtt{W}[1]$ y con $F_{\langle a_2\rangle}(\mathtt{Y})$ máximo, y así sucesivamente hasta obtener a_p y el vector de puntuaciones $\mathtt{W}[p] = \mathtt{Y}a_p$. Los ejes discriminantes son los p autovectores de $\mathtt{S}_3^{-1}\mathtt{S}_2$ y los valores máximos del estadístico F son, salvo el escalar $(n-r)/(r-1)$, sus respectivos autovalores $t_1,\ldots,t_p$. Dado que los $p-b$ últimos son necesariamente nulos, sólo se contemplan en la práctica los b primeros, de ahí que en el caso $r=2$ consideremos un único eje discriminante. En definitiva, si A denota la matriz $p\times p$ cuyas columnas son los vectores $a_1,\ldots,a_p$, podemos transformar la matriz de datos originales $\mathtt{Y}$ en una matiz de idénticas dimensiones con todas las puntuaciones discriminantes

$$\mathtt{W} = YA \tag{3.17}$$

donde A verifica

$$A'\mathtt{S}_3A = \mathtt{Id}, \qquad A'\mathtt{S}_2A = \left(\begin{array}{ccc|ccc} t_1 & & 0 & 0 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & t_b & 0 & & 0 \\ \hline 0 & & 0 & 0 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & 0 & 0 & & 0 \end{array}\right) \tag{3.18}$$

El siguiente resultado puede demostrarse a partir de (3.18) y (3.14) y es la clave definitiva para entender los ejes discriminantes y el significado de los autovalores $t_1,\ldots,t_b$:

▪ *Ejercicio* 55. Para todo $k = 1,\ldots,p$, se verifica:

$$\sum_{i=1}^{r} n_i \left(\overline{\mathtt{W}}_{i\cdot}[k] - \overline{\mathtt{W}}_{\cdot\cdot}[k]\right)^2 = t_k \tag{3.19}$$

Por otra parte, tal y como se adelantaba en el ejericio 44, los autovalores t_i pueden entenderse respectivamente como estimadores de los autovalores probabilísticos $\theta_1,\ldots,\theta_p$ de la matriz $\Sigma^{-1}\mu' P_{V|W}\mu$. La hipótesis inicial $H_0(1) : \theta_1 = 0$ equivale a $H_0 : \mu_1 = \ldots = \mu_r = 0$, y se contrasta mediante el manova de una vía a partir de $t_1,\ldots,t_b$, tomando como referencia la distribución $\chi^2_{p(r-1)}$. Sin embargo, la veracidad de la hipótesis inicial $H_0(2) : \theta_2 = 0$ equivale en términos intuitivos a que toda la discriminación entre las medias recaiga exclusivamente en el primer eje discriminante. La hipótesis $H_0(2)$ puede contrastarse a partir de $t_2,\ldots,t_p$ y tomando como referencia la distribución $\chi^2_{(p-1)(r-2)}$. De esta forma puede evaluarse la capacidad de discriminación de sucesivos ejes, aunque en la práctica la valoraremos directamente en términos muestrales ponderando los autovalores $t_1,\ldots,t_b$.

▪ *Ejercicio* 56. Interpretar en los términos de la teoría los cuadros 3.5 y 3.6, correspondientes a la comparación multivariante de medias entre las tres especies de flores de irisdata.

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	32,192[a]	99,1	99,1	,985
2	,285[a]	,9	100,0	,471

a. Se han empleado las 2 primeras funciones discriminantes canónicas en el análisis.

Cuadro 3.5: Autovalores y correlaciones canónicas

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1 a la 2	,023	546,115	8	,000
2	,778	36,530	3	,000

Cuadro 3.6: Test de Wilks

3.4. Regresión multivariante

Desarrollamos aquí el ejemplo 8, lo cual da pie al análisis de correlación canónica. El problema se expresa formalmente así: $\mathtt{Y} = \mathtt{X}\beta + \mathcal{E}$, donde $\mathcal{E} \sim N_{n,p}(0, \mathtt{Id}, \Sigma)$ con $\Sigma > 0$ y siendo β una matriz de dimensiones $(q+1) \times p$ del tipo (2.42). El problema de estimación de β queda resuelto en (2.45). En lo referente al problema de contraste de hipótesis, consideraremos dos casos de especial interés.

3.4.1. Contraste total: análisis de correlación canónica

Estudiamos primeramente el contraste de la hipótesis inicial $H_0 : \underline{\beta} = 0$ que, en términos de la media $\mu = \mathtt{X}\beta$, se expresa mediante $H_0 : \mu \in W = \langle 1_n \rangle$. Por lo tanto, $\dim V|W = q$ y $b = \text{mín}\{p, q\}$. Se denotará por $\lambda(\mathtt{Y})(\mathtt{Z})$ el estadístico λ de Wilks para el contraste total.

▪ *Ejercicio* 57. Probar que, en este caso, se verifica

$$\mathtt{S}_2 = nS_{yz}S_{zz}^{-1}S_{zy} \tag{3.20}$$
$$\mathtt{S}_3 = n[S_{yy} - S_{yz}S_{zz}^{-1}S_{zy}] \tag{3.21}$$

El test de Wilks consiste en confrontar $-[n-(q+1)]\log\lambda(\mathtt{Y})(\mathtt{Z})$ con con la distribución χ^2_{pq}, donde

$$\lambda(\mathtt{Y})(\mathtt{Z}) = \prod_{i=1}^{b}(1+t_i)^{-1}, \qquad t_1 > \ldots > t_b > 0 \text{ autovalores positivos de } \mathtt{S}_3^{-1}\mathtt{S}_2 \tag{3.22}$$

▪ *Ejercicio* 58. En el caso $p = 1$, que se corresponde con el problema de regresión múltiple, tenemos un único número

$$t_1 = \frac{R^2}{1-R^2} \tag{3.23}$$

Es decir que, si $p = 1$, el test total puede expresarse en función del coeficiente de correlación múltiple (al cuadrado) definido en (1.21), según (3.23). En el caso multivariante $p \geq 1$ podemos generalizar la relación anterior si definimos $r_1^2 > \ldots > r_b^2 > 0$ como los autovalores positivos de $S_{yy}^{-1} S_{yz} S_{zz}^{-1} S_{zy}$.

▪ *Ejercicio* 59. Probar que

$$r_i^2 = \frac{t_i}{1+t_i}, \text{ es decir, } t_i = \frac{r_i^2}{1-r_i^2}, \quad i = 1, \ldots, b \tag{3.24}$$

Los autovalores $r_1^2 > \ldots > r_b^2 > 0$ se denominan coeficientes de correlación canónica muestrales (al cuadrado) y, según hemos visto, contienen información relevante en el contraste de la hipótesis $H_0 : \underline{\beta} = 0$. De hecho, podemos considerarlos como estimadores de los coeficientes de correlación canónica probabilísticos $\rho_1^2 \geq \ldots \geq \rho_b^2$, que se definen según el teorema 3.4.1, de manera, que para todo i, $\theta_i = \rho_i^2/(1-\rho_i^2)$. Es decir, que la hipótesis inicial puede expresarse en términos de correlaciones canónicas mediante $H_0 : \rho_1^2 = \ldots = \rho_b^2 = 0$. No obstante, podemos ofrecer una interpretación más clara de los coeficientes de correlación canónica.

En lenguaje probabilístico, si Y y Z son vectores aleatorios de dimensiones p y q, respectivamente, buscamos $\alpha_1 \in \mathbb{R}^p$ y $\beta_1 \in \mathbb{R}^q$ tales que las variables $U_1 = \alpha_1' Y$ y $V_1 = \beta_1' Z$ tengan varianza 1 y su correlación sea máxima entre todas las proyecciones de Y y Z sobre sendos ejes de $\mathbb{R}^p$ y $\mathbb{R}^q$. En ese caso, los ejes obtenidos, $\langle \alpha_1 \rangle$ y $\langle \beta_1 \rangle$, se denominan primer par de ejes canónicos, y (U_1, V_1), el primer par de variables canónicas. La correlación(al cuadrado) entre ambas se denota por ρ_1^2 y se denomina primer coeficiente de correlación canónica. El siguiente paso es determinar otro par de ejes y, por lo tanto, otro par de proyecciones (U_2, V_2), incorreladas con (U_1, V_1) y con una correlación entre sí, ρ_2^2, máxima, y así sucesivamente hasta llegar a $b = \text{mín}\{p, q\}$. Consideremos las siguientes matrices cuadradas de orden p y q, respectivamente, y rango $b = \text{mín}\{p, q\}$:

$$\Sigma_{yy}^{-1} \Sigma_{yz} \Sigma_{zz}^{-1} \Sigma_{zy} \tag{3.25}$$

$$\Sigma_{zz}^{-1} \Sigma_{zy} \Sigma_{yy}^{-1} \Sigma_{yz} \tag{3.26}$$

▪ *Ejercicio* 60. Probar que los b primeros autovalores de las matrices (3.25) y (3.26) coinciden (no así sus respectivos autovectores).

La demostración del siguiente resultado, que se recoge en el manual 59 de la UEx, se basa fundamentalmente en el lema 3.0.1:

Teorema 3.4.1. Con las notaciones precedentes se verifica:

(i) Los coeficientes de correlación canónicas $\rho_1^2 \ldots, \rho_b^2$ son los b primeros autovalores de la matriz (3.25).

(ii) Los vectores $\alpha_1, \ldots, \alpha_b$ que determinan los ejes canónicos asociados a Y pueden obtenerse como autovectores de la matriz (3.25) asociados a $\rho_1^2 \ldots, \rho_b^2$, respectivamente. Análogamente, los vectores $\beta_1, \ldots, \beta_b$ que determinan los ejes canónicos para Z pueden obtenerse como autovectores de la matriz (3.26) asociados a $\rho_1^2 \ldots, \rho_b^2$, respectivamente.

En definitiva, los ejes canónicos permiten entender de manera más natural la correlación lineal entre las variables respuestas y las explicativas.

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_q \end{pmatrix} \longrightarrow \left(\begin{array}{c} V_1 \\ \hline \vdots \\ \hline V_b \end{array}\right) \begin{array}{c} \overset{\rho_1}{\longleftrightarrow} \\ \\ \overset{\rho_b}{\longleftrightarrow} \end{array} \left(\begin{array}{c} U_1 \\ \hline \vdots \\ \hline U_b \end{array}\right) \longleftarrow \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$$

▪ *Ejercicio* 61. Expresar la definición y el teorema anteriores en términos muestrales.

▪ *Ejercicio* 62. Probar que, dada una variable aleatoria real Y y un vector aleatorio Z de dimensión q, la máxima correlación lineal simple entre Y y una combinación lineal de las componentes de Z, $\beta' Z$, es el coeficiente (1.18), y se obtiene con β según (1.26).

Sabemos que la hipótesis inicial $H_0 : \mu_1 = \ldots = \mu_r$ en un diseño completamente aleatorizado equivale a $H_0 : \underline{\beta} = 0$ si parametrizamos el modelo como una regresión lineal multivariante respecto a $r-1$ variables dummys. En ese caso, los autovalores $t_1, \ldots, t_b$ correspondientes al manova de una vía, que expresan la capacidad de discriminación de los ejes discriminantes, pueden calcularse a partir de S_2 y S_3 definidas según (3.20) y (3.21), siendo Z el vector de variables dummys. No obstante, dichos autovalores se relacionan con los coeficientes de correlación canónica según (3.24). Por lo tanto, el propio manova de una vía puede expresarse en términos de los coeficientes de correlación canónica, calculados a partir de las variables dummys, de la misma forma que el anova de una vía se expresa en términos del coeficiente de correlación múltiple R^2. Además, r_i expresa, al igual que t_i, el poder de discriminación del eje $\langle a_i \rangle$, con la ventaja a la hora de interpretarlo de que está acotado entre 0 y 1.

▪ *Ejercicio* 63. Probar que los ejes discriminantes son los propios ejes canónicos para Y que se obtienen considerando como Z el vector de variables dummys (ver figura 4.2). ¿Cómo podrían obtenerse mediante SPSS los ejes canónicos corerspondientes a las variables dummys?

▪ *Ejercicio* 64. Interpretar en los términos de la teoría los coeficientes de correlación canónica que aparecen en el cuadro 3.5.

Estamos ya en condiciones de interpretar los cuatro tests del manova propuestos en la página 35: dado que los autovalores $t_1, \ldots, t_b$ o, mejor, los coeficientes de correlación canónica asociados, $r_1^2, \ldots, r_b^2$, resumen el grado de correlación entre el vector numérico respuesta Y y el vector explicativo Z (que en la práctica suele corresponderse con las variables dummys asociadas a un factor cualitativo) falta por determinar cómo se evalúan exactamente. Pues bien, el test de Roy considera únicamente el primer autovalor t_1; el test de Lawley-Hotelling considera la suma de todos, $\sum_{i=1}^b t_i$; el de Pillay hace lo mismo pero tomando las correlaciones canónicas, es decir, $\sum_{i=1}^b r_i^2$; el test de Wilks, que es el que utilizamos aquí, los multiplica tras sumarles el elemento neutro del producto, es decir, $\prod_{i=1}^b (1+t_i)^{-1}$ (se invierte porque el TRV se ha expresado así tradicionalmente).

3.4.2. Contrastes parciales: método Lambda de Wilks

El otro tipo de contraste de interés está relacionado con la depuración del modelo mediante los algoritmos de selección de variables. Se trata en esta ocasión de contrastar hipótesis iniciales del tipo $H_0 : \beta_{j_1} = \ldots = \beta_{j_k} = 0$ para $k < q$ y $j_1, \ldots, j_k \in \{1, \ldots, q\}$. La veracidad de esa hipótesis conllevaría suprimir del modelo un parte de la matriz Z que se denota por Z_D y está compuesta por las columnas $j_1, \ldots, j_k$, dando lugar a un modelo reducido con un nueva matriz $\mathsf{Z}_R \in \mathcal{M}_{n\times(q-k)}$.

▪ *Ejercicio* 65. Probar que, si $k = 1$, el problema puede resolverse haciendo uso de la distribución $T^2_{p,n-(q+1)}$ que, salvo una constante, coincide con $F_{p,n-p-q}$.

En todo caso, se denota por $\lambda(Y)(\mathsf{Z}_R|\mathsf{Z}_D)$ el estadístico de Wilks que resuelve este contraste. El método de Wilks oferece una ventaja a la hora de elaborar un algoritmo de selección de variables, pues los estadísticos de contraste para los tests parciales pueden obtenerse a partir de los correspondientes a los test totales para los diferentes modelos reducidos.

▪ *Ejercicio* 66. Probar que

$$\lambda(\mathtt{Y})(\mathtt{Z}_R|\mathtt{Z}_D) = \frac{\lambda(\mathtt{Y})(\mathtt{Z})}{\lambda(\mathtt{Y})(\mathtt{Z}_R)} \tag{3.27}$$

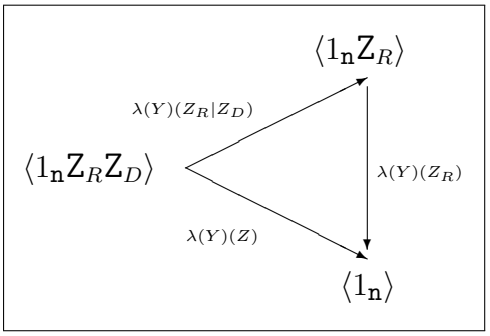


Figura 3.3: Test de Wilks parcial

En el caso de la regresión lineal multivariante, pomos considerar, además de los conocidos algoritmos de selección de variables explicativas (hacia adelante, hacia atrás, pasos sucesivos), otros para la selección de variables respuesta: dado cualquier $j = 1, \ldots, p$, entendemos que el vector $\mathtt{Y}[j]$ es prescindible en el modelo cuando, si consideramos un modelo de regresión lineal múltiple con $\mathtt{Y}[j]$ como variable respuesta y $\mathtt{Z}, \mathtt{Y}_R$ como explicativas, $\mathtt{Z}$ debería ser eliminada según el test parcial. Este criterio a la hora de seleccionar variables se relaciona con el concepto probabilístico de independencia condicional.

▪ *Ejercicio* 67. Probar que el contraste para $\mathtt{Y}[j]$ puede resolverse haciendo uso de la distribución $F_{q,n-(p+q)}$.

Se denota no obstante mediante $\lambda(\mathtt{Y}_R|\mathtt{Y}[j])(\mathtt{Z})$ el estadístico de Wilks que resuelve este contraste .

▪ *Ejercicio* 68. Teniendo en cuenta (3.27), probar que

$$\lambda(\mathtt{Y}_R|\mathtt{Y}[j])(\mathtt{Z}) = \frac{\lambda(\mathtt{Y})(\mathtt{Z})}{\lambda(\mathtt{Y}_R)(\mathtt{Z})} \tag{3.28}$$

Si estamos relacionando un vector numérico Y con un factor cualitativo que distingue r categorías y parametrizamos el modelo mediante $r-1$ variables dummys recogidas en una matriz $\mathtt{Z}$, podemos aplicar una selección de variables respuesta para determinar qué componentes de Y guardan una relación esencial con el factor. El método Lambda de Wilks se define como el algoritmo de selección hacia adelante de variables respuestas según el test (3.28), y será de utilidad en el capítulo 4.

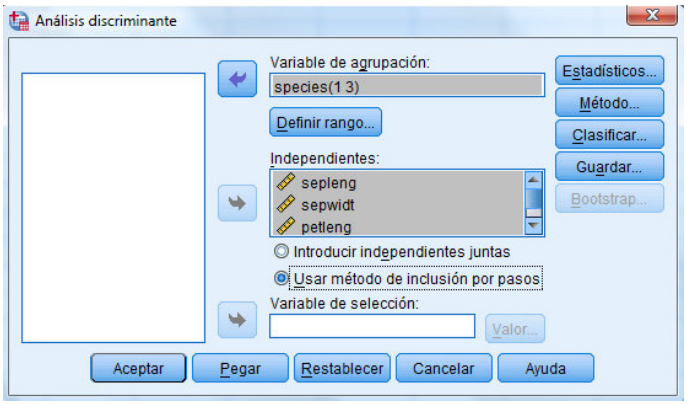
▪ *Ejercicio* 69. Probar que, en la fase j-ésima del algoritmo Lambda de Wilks, se introduce la variable que, añadida a las $j-1$ ya incluidas anteriormente, aporta una resultado más significativo en el manova de una vía, es decir, un valor mínimo en el estadístico lambda de Wilks, siempre y cuando resulte significativo el test parcial que se resuelve mediante la distribución del ejercicio 67.

El cuadro 3.7 refleja en qué orden se van introduciendo las variables numéricas (todas) de irisdata según el algoritmo Lambda de Wilks para la discriminación entre especies. En el cuadro de diálogo de SPSS 3.8 se indica cómo ejecutar el análisis discriminate.

Variables no incluidas en el análisis

Paso		F para entrar	Lambda de Wilks
0	sepleng	119,265	,381
	sepwidt	49,160	,599
	petleng	1180,161	,059
	petwidt	960,007	,071
1	sepleng	34,323	,040
	sepwidt	43,035	,037
	petwidt	24,766	,044
2	sepleng	12,268	,032
	petwidt	34,569	,025
3	sepleng	4,721	,023

Cuadro 3.7: Método Lambda de Wilks para irisdata



Cuadro 3.8: Cuadro de diálogos Lambda de Wilks

3.5. Análisis de perfiles

Se trata de una técnica que generaliza el test de Student para muestras relacionadas y da sentido al contraste $H_0 : \nu = 0$ estudiado en la primera sección del capítulo. Este método puede considerarse una alternativa más robusta al análisis de medidas repetidas (ver Arnold (1981) y Hair et al. (1999)).

En ocasiones resulta interesante estudiar la evolución de una carácter numérico a lo largo de una secuencia temporal con p mediciones. En ese caso, contaremos con un vector Y p-dimensional, de manera que la hipótesis inicial $H_0 : \mu[1] = \ldots = \mu[p]$ se interpreta como una ausencia de evolución, al menos por término medio. También puede ser interesante comparar las evoluciones en distintas categorías de un factor cualitativo, como en el ejemplo que se recoge en la figura 3.4, que corresponde del dolor durante seis meses distinguiendo tres tratamientos[3]

En este caso, que los tres tratamientos tengan efectos idénticos por término medio equivale a la hipótesis inicial $H_0 : \mu_1 = \mu_2 = \mu_3$ del diseño completamente aleatorizado, que se contrasta

[3] J. Rodríguez Mansilla et al. Clinical Rehabilitation (2014).

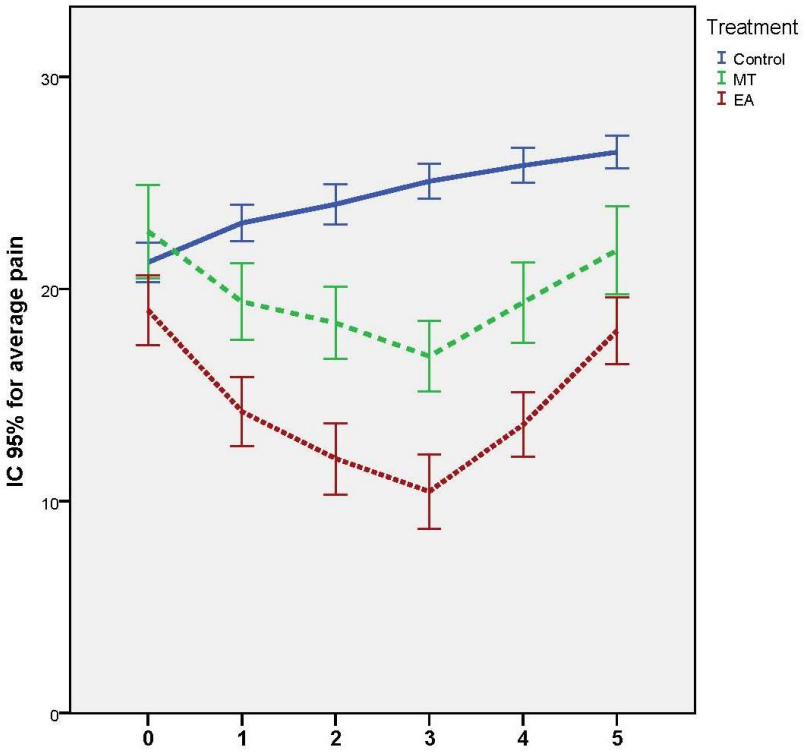


Figura 3.4: Perfiles dolor por tratamientos

mediante el manova de una vía. No obstante, también puede resultar de interés contrastar, por ejemplo, el paralelismo de los perfiles, que se interpreta como una evolución similar desde la fase inicial. Si contamos con sólo $p = 2$ mediciones, una inicial y otra final, estaremos ante un diseño conocido como de muestras relacionadas. Se resuelve calculando la diferencia D, con media ν, entre las dos fases. De esta forma, la hipótesis inicial $H_0 : \mu[1] = \mu[2]$ equivale a $\nu = 0$ y se contrasta mediante el test de Student para una muestra aplicado a D. La hipótesis inicial de paralelismo entre los r perfiles equivale a $\nu_1 = \ldots = \nu_r$ y se contrasta mediante el anova de una vía.

Sin embargo, cuando consideramos más de 2 fases debemos calcular la diferencia entre cada variable y la anterior, dando lugar a un vector D en dimensión $p - 1$. La hipótesis inicial $H_0 : \mu[1] = \ldots = \mu[p]$ se contrasta mediante el test (3.3) aplicado a D, y la de paralelismo entre los r perfiles, mediante el manova de una vía.

Abordar un análisis de perfiles mediante un manova es sólo una de las posibles opciones y, seguramente, no la más popular. Los supuestos es los que basa son la normalidad multivariante y la igualdad de matrices de covarianzas (en el caso de incluir un factor intersujeto en el modelo, como es el tratamiento en el estudio del dolor). Del primero sabemos que puede obviarse asintóticamente, lo cual justifica la robustez del modelo. Como principal alternativa podemos destacar[4] el modelo de medidas repetidas que, en principio, supone además dos condiciones adicionales sobre la matriz o matrices de covarianzas: la igualdad de las varianzas de las componentes, por un lado, y la igualdad de las covarianzas por otro. Un caso particular de esta hipótesis es el supuesto de esfericidad (homocedasticidad y covarianzas nulas), que conduciría a aplicar un test F, pudiendo aplicarse correcciones en los grados de libertad tanto del

[4]Rencher (1996), sección 6.9.

numerador como del denominador en dichos test en función del grado de desviación respecto al modelo esférico. En eso consiste en la práctica el análisis de medidas repetidas. Si no existe un factor intergrupo y no estamos dispuestos a asumir hipótesis relativas a la distribución del vector (salvo la continuidad del mismo) contamos con la alternativa de Friedman basada en rangos.

Capítulo 4

Problema de clasificación

En este capítulo vamos a abordar el problema de clasificación de una unidad experimental respecto a r categorías posibles a partir de la medición de p variables numéricas. Por ejemplo, mediante los datos recogidos en el archivo irisdata podemos elaborar una estrategia para determinar a qué especie (setosa, virgínica o vesicolor) pertenece un lirio a partir de la observación de sus cuatro medidas morfológicas (longitud y anchura de pétalo y sépalo). Desde el punto de vista formal y al igual que sucede en el diseño completamente aleatorizado, contamos con un vector numérico Y p-dimensional y un factor cualitativo I con r categorías que se identifican con sendas distrubuciones en $\mathbb{R}^p$. Debería existir una fuerte relación entre Y y el factor cualitativo para que pudiéramos discriminar entre las categorías a partir de Y. En ese sentido, el problema de clasificación puede entenderse como el problema inverso a la comparación de medias que vimos en el capítulo anterior. La diferencia entre ambos problemas estriba en los roles que desempeñan el vector numérico y el factor cualitativo en cada caso, como se ilustra en la figura 4.1.

El problema de clasificación se enmarca en el contexto teórico de la Teoría de la Decisión, pues buscamos una estrategia adecuada para decidir a qué distribución, pertenece una observación en $\mathbb{R}^p$. Puede tratarse tanto a partir de un modelo frecuentista como Bayesiano. El punto de vista Bayesiano se caracteriza por la posibilidad de contemplar preferencias *a priori* hacia ciertas categorías. El modelo clásico o frecuentista puede entenderse como un modelo condicional dado I del Bayesiano, dadas las distribuciones concretas a las que pertenecen las observaciones numéricas. En un contexto Bayesiano la solución óptima al problema de clasificación es la denominada estrategia de Bayes. A pesar de las discrepancias a nivel filosófico entre los marcos teóricos Bayesiano y frecuentista, la estrategia de Bayes podría ser también asumida según el segundo punto de vista si consideramos una distribución *a priori* uniforme. Precisamente, dicha estrategia coincide con la estrategia óptima según el criterio minimax, propia del marco frecuentista. Bajo ciertas condiciones, la estrategia minimax consiste, como veremos, en el Análisis Discriminate Lineal (LDA) de Fisher, que constituye el núcleo del capítulo. Veremos también que la estrategia LDA puede entenderse como el reverso de una moneda cuyo anverso es el manova de una vía, lo cual da pleno sentido al estudio del modelo lineal multivariante.

En general, los elementos básicos de la Teoría de la Decisión e Inferencia Bayesiana pueden encontrarse en Nogales (1998). Para un desarrollo más detallado de este problema concreto remitimos al lector a Anderson (1958).

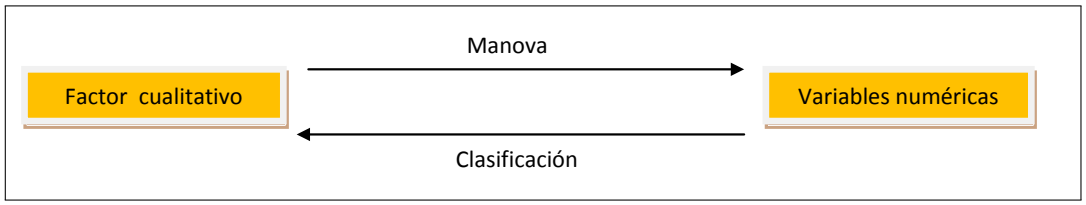


Figura 4.1: Manova y clasificación

4.1. Planteamiento general

Nuestro propósito inicial es entender cómo se afronta este problema en el caso sencillo de que existan sólo dos categorías para extenderlo después de manera natural al caso general de r categorías. No consideraremos aquí ninguna función de pérdida o costes para nuestro problema de decisión y supondremos inicialmente que las distribuciones de probabilidad del modelo son continuas.

Si tenemos que decidir si una observación $y \in \mathbb{R}^p$ proviene de un modelo de distribución P_1, con densidad p_1 o, por el contrario, se explica por un modelo P_2, con densidad p_2, podemos distinguir entre dos tipos de estrategias: una estrategia aleatoria S consisten en calcular la probabilidad de que la observación provenga de distribución P_1 $(I = 1)$ en función de la observación Y; una estrategia no aleatoria $\mathtt{S}$ sólo probabilidad 1 ó 0, es decir, se decanta directamente por P_1 o P_2 en función de Y. Los riesgos o probabilidades de error $R_{\mathtt{S}}(1)$ y $R_{\mathtt{S}}(2)$ asociados a una estrategia no aleatoria se pueden calcular entonces a partir de p_1 y p_2 de forma trivial. Obviamente, cualquier estrategia aleatoria S lleva asociada otra no aleatoria $\mathtt{S}$ basada en un punto de corte en la probabilidad condicional, que por defecto es 0.5.

Podemos considerar un preorden $\succeq$ en el conjunto de las estrategias no aleatorias, de manera que $\mathtt{S}_1 \succeq \mathtt{S}_2$ cuando $R_{\mathtt{S}_1}(i) \leq R_{\mathtt{S}_2}(i)$, para $i = 1, 2$. Se dice $\mathtt{S}_1 \succ \mathtt{S}_2$ cuando alguna de las desigualdades es estricta. Encontrar un elemento maximal es posible sólo en ciertos casos triviales. Dada una distribución distribución *a priori* Q concreta, que se identifica con el número $q = Q(\{P_1\}) \in [0, 1]$, definimos su riesgo de Bayes para q como el error cuadrático medio respecto a la función indicador para $\{I = 1\}$

$$R_S^q = q \cdot \int (S(y) - 1)^2 \cdot p_1(y)\, dy + (1 - q) \cdot \int S^2(y) \cdot p_2(y)\, dy \tag{4.1}$$

En el caso particular de que la estrategia sea no aleatoria, se trata de la combinación convexa de sus riesgos $qR_{\mathtt{S}}(1) + (1-q)R_{\mathtt{S}}(2)$. En todo caso, el riesgo se minimiza mediante la esperanza condicional, es decir, la distribución *a posteriori* determinada por la Regla de Bayes , que se denomina estrategia de Bayes para q denotándose por S_q:

$$S_q(y) = P(I = 1|Y = y) = \frac{q \cdot p_1(y)}{q \cdot p_1(y) + (1 - q) \cdot p_2(y)} \tag{4.2}$$

La estrategia no aleatoria asociada $\mathtt{S}_q$ consistirá en asignar $y \in \mathbb{R}^p$ a P_1 cuando

$$\frac{p_1(y)}{p_2(y)} \geq \frac{1 - q}{q} \tag{4.3}$$

▪ *Ejercicio* 70. Probar que $\mathtt{S}_q$ minimiza el riesgo de Bayes para q en el subconjunto de estrategias no aleatorias.

Puede probarse también que la clase estrategias no aleatorias $\{\mathtt{S}_q : \ q \in [0,1]\}$ constituyen una familia completa maximal, es decir: cualquier estrategia no aleatoria que no pertenece a la famila es mejorada estrictamente por alguna de la familia, y no existe ninguna dentro de la familia mejorada estrictamente por alguna otra. Así pues, deberíamos seleccionar una estrategia de este tipo, pero la elección depende del valor de q que queramos considerar y ello nos situaría en un marco Bayesiano. Si no estamos en condiciones de proponer una distribución *a priori*, podemos optar por escoger la estrategia no aleatoria minimax, que es el elemento maximal para el orden definido a partir del máximo de los riesgos. Puede probarse que es la estrategia no aleatoria asociada a la estrategia Bayes para una distribución *a priori* uniforme, es decir, $\mathtt{S}_{0.5}$, y que $R_{\mathtt{S}_{0.5}}(1) = R_{\mathtt{S}_{0.5}}(2)$. Ésta es la que adoptaremos por defecto si no estamos dispuestos a asumir el marco teórico Bayesiano, teniendo en cuenta que, en todo caso, la introducción de una distribución *a priori* en el modelo se traduciría simplemente en una corrección trivial de la estrategia minimax según (4.3). En definitiva, la estrategia minimax consiste en asignar y a P_1 cuando se verifica

$$p_1(y) \geq p_2(y) \tag{4.4}$$

Es decir, se asigna la observación a la distribución que la hace más verosímil. Se trata pues de una aplicación directa del Principio de Máxima Verosimilitud y ésta es la idea fundamental que debe prevalecer. En el caso general de r categorías se procede de forma idéntica, asignando y a P_i cuando

$$p_i(y) \geq p_j(y), \qquad \forall j \neq i \tag{4.5}$$

Método núcleo de estimación de densidades: Teniendo en cuenta (4.5), es obvio cómo deberíamos resolver el problema si dispusiéramos de adecuadas estimaciones de las funciones de densidad: asignando y a P_i cuando

$$\hat{p}_i(y) \geq \hat{p}_j(y) \qquad \forall j \neq i \tag{4.6}$$

Describiremos aquí heurísticamente el denominado método del núcleo de estimación de densidades, empezando por el caso univariante. Para un estudio más detallado remitimos al lector a Silverman (1986).

Supongamos que contamos con una muestra aleatoria $\mathtt{y}_1, \ldots, \mathtt{y}_n$, correspondiente a una determinada distribución continua con función de densidad p y queremos estimar el valor de p en y, que se denota $\hat{p}(y)$. Para ello escogemos un número $\delta > 0$, que denominaremos ancho de banda, y consideramos el intervalo $[y-\delta, y+\delta]$, de amplitud 2δ. Si $N(y)$ denota la cantidad de datos de la muestra en el anterior intervalo y n es suficientemente grande, se sigue de la Ley Débil de los Grandes Números que $P\big([y-\delta, y+\delta]\big) \simeq N(y)/n$. Por otra parte, si δ es pequeño se verifica por el Teorema Fundamental del Cálculo que $P\big([y-\delta, y+\delta]\big) \simeq p(y) \cdot 2\delta$, lo cual nos induce a definir para cada $y \in \mathbb{R}^p$ el estimador $\hat{p}(y) = N(y)/(2n\delta)$. Si queremos expresar $\hat{p}$ en función de los datos de la muestra, hemos de tener en cuenta que un dato $\mathtt{y}_i$ pertenece al intervalo anterior si, y sólo si, $\delta^{-1}|\mathtt{y}_i - y| \leq 1$. Definimos entonces la función (denominada núcleo)

$$K(u) = \begin{cases} \frac{1}{2} \text{ si } |u| \leq 1 \\ 0 \text{ si } |u| > 1 \end{cases}, \quad u \in \mathbb{R}. \tag{4.7}$$

De esta forma,

$$\hat{p}(y) = \frac{1}{n\delta} \sum_{i=1}^{n} K\left(\frac{y - \mathtt{y}_i}{\delta}\right), \qquad x \in \mathbb{R} \tag{4.8}$$

En el caso multivariante (dimensión p) no consideramos intervalos de amplitud 2δ centrados en y sino cubos de volumen $2^p\delta^p$, y el núcleo K^p asigna el valor 2^{-p} a un punto u cuando $\|u\|_\infty \leq 1$. De esta forma, la función de densidad se estima reemplazando en (4.8) K por K^p y δ por δ^p. No obstante, la función de densidad estimada será de tipo escalonado. Un procedimiento comúnmente utilizado para suavizarla es considerar, en vez del núcleo anterior, el siguiente:

$$\tilde{K}(u) = \frac{1}{(2\pi S)^{p/2}} \exp\left\{-\frac{1}{2}u'S^{-1}u\right\}, \qquad u \in \mathbb{R}^p, \tag{4.9}$$

donde S es la matriz de covarianzas muestral. Así, la función de densidad se estima mediante

$$\hat{p}(y) = \frac{1}{n\delta^p\,(2\pi S)^{p/2}} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2\delta^2}(y - \mathtt{y}_i)'S^{-1}(y - \mathtt{y}_i)\right\} \tag{4.10}$$

Podemos comprobar que la función anterior se trata, efectivamente, de una densidad. Una vez estimadas las densidades de las distintas categorías procederemos a establecer las regiones de clasificación según (4.6). En la literatura estadística encontramos núcleos diferentes a (4.9), denominado gaussiano, como el triangular, el del coseno o de Epanechnikov, entre otros. Hay que tener en cuenta que la estimación de las densidades, y por tanto la estrategia de clasificación, depende de la elección del núcleo K y del ancho de banda δ. Diversos trabajos vienen a convencernos de que la elección del núcleo es poco determinante. Sin embargo, la elección del ancho de banda sí lo es. No podemos hablar, desde luego, de un ancho de banda universal, sino que debe depender del problema considerado. La selección de un ancho de banda excesivamente grande tenderá a estimar la densidad demasiado plana, mientras que uno excesivamente pequeño la estimará de manera excisivamente abrupta.

Otro inconveniente a tener en cuenta es la denominada "maldición de la dimensión", que consiste en que el número de datos requerido para lograr una estimación satisfactoria de la densidad crece exponencialmente en relación con la dimensión considerada. Por lo tanto, cuando tengamos un amplio número de variables precisaremos de una cantidad ingente de datos para obtener una estimación fiable de la densidad. Eso explica el hecho de que sigamos haciendo hincapié aquí en el método tradicional para clasificar observaciones, denominado Análisis Discriminante Lineal (LDA), debido a Fisher.

4.2. Análisis Discriminate Lineal

En la Estadística Paramétrica es muy frecuente partir del supuesto de normalidad a la hora de formular un modelo cuyos estimadores y tests de hipótesis podrán tener distintos comportamiento ante casos reales en los que este supuesto no se cumpla. En el contexto del modelo lineal normal, al supuesto de normalidad se le añade el de igualdad de varianzas o de matrices de covarianzas. Eso es precisamente, lo que haremos a continuación.

Sabemos que para determinar la estrategia aleatoria óptima en un marco Bayesiano se precisa conocer la distribución *a priori* y las diferentes funciones de densidad. Si no somos capaces de aportar una estimación con garantías de las densidades, podemos optar por suponer que las r distribuciones en juego siguen modelos p-normales con idéntica matriz de covarianzas, es decir, $P_i = N_p(\mu_i, \Sigma)$, para $i = 1, \ldots, r$. En ese caso, la estrategia dependerá únicamente de la distribución *a priori* y de la distancia D^2_Σ definida en (1.38). En un problema de clasificación binaria con una distribución *a priori* dada por $q \in [0, 1]$, la estrategia aleatoria óptima

consistiría, según (4.2), en la distribución *a posteriori* dada por

$$P(I=1|Y=y)=\frac{q\cdot\exp\{-\frac{1}{2}\cdot D^2_\Sigma(y,\mu_1)\}}{q\cdot\exp\{-\frac{1}{2}\cdot D^2_\Sigma(y,\mu_1)\}+(1-q)\cdot\exp\{-\frac{1}{2}\cdot D^2_\Sigma(y,\mu_2)\}}, \tag{4.11}$$

de lo que se deduce que

$$P(I=1|Y=y)=\mathtt{L}\big(-(\beta_0+y'\underline{\beta})\big) \tag{4.12}$$

donde

$$\beta_0 = \log\frac{1-q}{q}+\frac{1}{2}\left[\mu_1'\Sigma^{-1}\mu_1-\mu_0'\Sigma^{-1}\mu_0\right], \tag{4.13}$$

$$\underline{\beta} = \Sigma^{-1}(\mu_0-\mu_1) \tag{4.14}$$

y $\mathtt{L}$ denota la denominada función logística, representada en la figura 4.6, que se define mediante

$$\mathtt{L}(x)=\frac{1}{1+\mathsf{e}^x},\qquad x\in\mathbb{R} \tag{4.15}$$

De esta forma, al condicionar sobre los valores numéricos de las observaciones, $\mathtt{Y}_{i1},\ldots,\mathtt{Y}_{in_i}$, $i=1,2$, en el modelo original Bayesiano, obtenemos el conocido como modelo de regresión logística, que es del tipo lineal generalizado. Por contra, si condicionamos respecto a las categorías de las observaciones tendremos dos muestras aleatorias simples e independientes de sendas distribuciones $N_p(\mu_i,\Sigma)$, lo cual supone un modelo lineal normal multivariante.

Tanto (4.11) como (4.12) expresan la estrategia en función de parámetros poblacionales que en la práctica serán desconocidos. En la regresión logística, los parámetros se estimarán siguiendo la técnica propia de los modelo lineales generalizados. Sin embargo, el método LDA se deriva de la estrategia no aleatoria minimax, es decir, la asociada a la estrategia aleatoria óptima (4.11) para $q=0.5$[1], o equivalentente (4.4), pero sustituyendo los parámetros probabilísticos μ_1, μ_2 y Σ por estimadores óptimos de los mismos, según el modelo lineal normal multivariante. Es decir, se considera como estimador de μ_i a la media muestral $\overline{\mathtt{Y}}_{i\cdot}$ de la categoría i-ésima; como estimador de Σ tomaremos $(n_1+n_2)^{-1}(S_1+S_2)$, siendo S_i la matriz de covarianzas muestral de la i-ésima categoría. En ese caso, el método asigna Y a la media más cercana según la distancia $D^2_{\hat\Sigma}$. En general, si se trata de un problema de clasificación respecto a r categorías, la estrategia minimax consiste en asignar Y a la categoría i-ésima cuando se verifica

$$(Y-\overline{\mathtt{y}}_{i\cdot})'\hat\Sigma^{-1}(Y-\overline{\mathtt{y}}_{i\cdot})\le(Y-\overline{\mathtt{y}}_{j\cdot})'\hat\Sigma^{-1}(Y-\overline{\mathtt{y}}_{j\cdot}),\qquad\forall j\neq i \tag{4.16}$$

Si contemplamos una distribución *a priori* diferente de la uniforme podemos estimarla mediante las proporciones de la muestra y corregir las desigualdades de (4.16) mediante sumandos que dependen de las proporciones por categorías. Es lo que podría denominarse estrategia LDA Bayesiana.

Cada desigualdad en (4.16) da lugar a la división del $\mathbb{R}^p$ en dos semiespacios cuya frontera es una subvariedad afín $(p-1)$-dimensional, de ahí que esta estrategia se denomine lineal (de Fisher) para diferenciarse de la cuadrática (de Fisher también), que veremos más adelante, en la cual $\mathbb{R}^p$ estará fragmentado por cuádricas.

Como ejemplo, utilizaremos el archivo irisdata de Fisher para intentar clasificar una flor entre las tres especies consideradas en función de sus cuatro medidas morfológicas. El programa

[1]Que es la que mejor se asimila a un marco Frecuentista.

SPSS diseña la estrategia LDA a partir de los datos ya asignados a categorías y es capaz de clasificar en función de la misma cualquier otro dato que aparezca desagrupado. También reclasifica según la estrategia los propios datos agrupados, como el caso que vemos a continuación. La reclasificación aporta una estimación de los riesgos de la estrategia, que son, según el cuadro 4.2, del 0 % para setosa, del 2 % para virginica y 4 % para vesicolor.

Número de caso	Grupo real	Grupo mayor			
		Grupo pronosticado	p	P(G=g | D=d)	Distancia de Mahalanobis
1	3	3	,503	1,000	1,376

Cuadro 4.1: Reclasificación según LDA

Resultados de la clasificación[a]

			Grupo de pertenencia pronosticado			
		species	setosa	vesicolor	virginica	Total
Original	Recuento	setosa	50	0	0	50
		vesicolor	0	48	2	50
		virginica	0	1	49	50
	%	setosa	100,0	,0	,0	100,0
		vesicolor	,0	96,0	4,0	100,0
		virginica	,0	2,0	98,0	100,0

a. Clasificados correctamente el 98,0% de los casos agrupados originales.

Cuadro 4.2: Estimaciones de los riesgos LDA

Según se indica en el cuadro 4.1, el dato uno se ha clasificado en el grupo 3 porque la media de éste minimiza la distancia de Mahalanobis. Tanto es así que el cociente $\hat{p}_3(y)/\sum_i \hat{p}_i(y) \simeq 1$, con las densidades estimadas sustituyendo las medias y matrices de covarianzas por sus estimadores. Por lo tanto, podemos considerarla una clasificación clara. De hecho, sabemos que es correcta.

La estrategia LDA de Fisher, definida en (4.16), posee buenas propiedades asintóticas. Concretamente, puede probarse que, en el caso binario ($r = 2$) los riesgos de la misma convergen asintóticamente al valor

$$\int_{\frac{1}{2}\theta}^{\infty} f(x)\, dx \tag{4.17}$$

siendo f la densidad de la distribución $N(0,1)$, y $\theta = D^2_{\Sigma}(\mu_1, \mu_2)$. Se trata del parámetro θ que aparece en (3.8) en relación con el contraste de la hipótesis inicial $H_0 : \mu_1 = \mu_2$, que se identifica con $\theta = 0$. Por lo tanto, si $\mu_1 = \mu_2$, la estrategia de Fisher se comportaría asintóticamente como un sorteo a cara o cruz. Sin embargo, a medida que las medias se alejan según la métrica de Mahalanobis, los riesgos asintóticos tienden a 0. En la práctica, que las distribuciones estén bien diferenciadas suele ser mucho más importante que el cumplimiento de los supuestos del modelo de cara a lograr una estrategia con riesgos bajos, que es lo que a la postre nos interesa. Eso es lo que ocurre con irisdata: no estamos en condiciones de asumir la normalidad ni la igualdad de matrices de covarianzas, pero las tres especies consideradas se diferencian claramente según

sus medidas morfológicas, de ahí el éxito de la estrategia de Fisher, que queda patente en el cuadro 4.2.

En definitiva, el manova de una vía y la estrategia de clasificación lineal de Fisher son métodos que parten del mismo modelo, aunque en el primer caso es el factor el que desempeña la función explicativa, mientras que en el segundo es el vector numérico. Un resultado poco significativo a la hora de comparar las medias no ofrece expectativas de éxito en la clasificación, justo al contrario que un resultado significativo. Por eso decimos que el manova y la clasificación son el anverso y el reverso de una misma moneda. De hecho, es el problema de clasificación el que da pleno sentido al estudio del manova y, dado que este último puede entenderse como una regresión multivariante respecto a las variables dummys, da sentido al estudio de los coeficientes de correlación canónica, pues el contraste de igualdad de medias puede expresarse en términos de los mismos según (3.24).

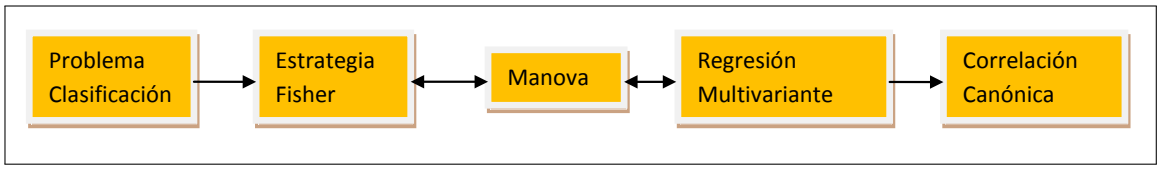


Figura 4.2: Esquema general

Una vez hemos entendido el problema de clasificación como un problema de relación entre un vector aleatorio p-dimensional y un factor con r categorías, cobra especial interés el método de selección de variables Lambda de Wilks, estudiado en el capítulo 3, pues permite desechar aquellas componentes del vector que no aportan información particular en el problema de clasificación.

4.2.1. LDA y ejes discriminantes

▪ *Ejercicio* 71. Como caso particular probar que, si $r = 2$, la estrategia (4.16) consiste en asignar Y a $\overline{\mathrm{y}}_1$ cuando

$$\left(Y - \frac{1}{2}(\overline{\mathrm{y}}_1 + \overline{\mathrm{y}}_2)\right)' S_c^{-1}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2) > 0 \tag{4.18}$$

lo cual sugiere una partición del espacio en función del eje discriminante definido en (3.11) y que se representa en la figura 3.2.

Este resultado establece una primera conexión entre los ejes discriminantes y el problema de clasificación, que desarrollaremos en el caso general $r \geq 2$. Las observaciones originales componen en principio un matriz $\mathtt{Y} \in \mathcal{M}_{n\times p}$, pero pueden expresarse también a través de la matriz $\mathtt{W}$, definida en (3.17) mediante $\mathtt{W} = \mathtt{Y}A$, cuyas columnas recogen las diferentes puntuaciones obtenidas al proyectar sobre los ejes discriminantes, determinados por las correspondientes columnas de A. Proyectar sobre los ejes discriminantes puede entenderse como un cambio de coordenadas. Podemos obrar de igual forma con cualquier vector aleatorio Y proyectándolo sobre los ejes discriminantes para obtener $W = A'Y$. Dado que la proyección es una aplicación lineal, la media aritmética de los datos proyectados coincide con la proyección de la media aritmética de los datos originales. De esta forma, la estrategia (4.16), expresada en términos de las puntuaciones discriminantes, asigna W a $\overline{\mathtt{W}}_{i\cdot}$ cuando, para $j = i$ se alcanza el siguiente mínimo:

$$\min_{1\leq j\leq r} \left[(A')^{-1}(W - \overline{\mathtt{W}}_{j\cdot})\right]' \hat{\Sigma}^{-1} \left[(A')^{-1(}W - \overline{\mathtt{W}}_{j\cdot})\right] \tag{4.19}$$

Dado que $\hat{\Sigma} = (n-r)^{-1}\mathsf{S}_3$ y en virtud (3.18), la estrategia de Fisher asigna W a la categoría de $\overline{\mathtt{W}}_{i\cdot}$ cuando en $j = i$ se alcanza $\min_j \|W - \overline{\mathtt{W}}_{j\cdot}\|^2$. Es decir, cuando expresamos los datos en términos de las puntuaciones discriminantes se trata de minimizar la distancia Euclídea. Por lo tanto, se trata de buscar el siguiente mínimo

$$\min_{1\le j\le r} \sum_{k=1}^{p} \left(W[k] - \overline{\mathtt{W}}_{j\cdot}[k]\right)^2 \tag{4.20}$$

Se sigue entonces de (3.19) que, si $k > b$, dado que $t_k = 0$, $\overline{\mathtt{W}}_{j\cdot}[k] = \overline{\mathtt{W}}_{\cdot\cdot}[k]$ para todo j, es decir, que el sumando k-ésimo en (4.20) será constante en j y, por lo tanto, no interviene en la búsqueda del mínimo en j. Por eso podemos ignorar las puntuaciones discriminantes asociadas a autovalores nulos, de manera que el problema queda reducido a minimizar distancias Euclídeas en dimensión b. Si el valor de b es bajo podemos pues visualizar el problema de clasificación mediante un gráfico b-dimensional. Por ejemplo, en el caso de irisdata, tenemos el diagrama de dispersión de la figura 4.3.

Para valores altos de b podemos visualizar igualmente el problema desechando las puntuaciones discriminantes asociadas a autovalores pequeños pues, razonando de manera análoga, los correspondientes sumandos en (3.19) serán casi constantes y, por lo tanto, tendrán escasa influencia en el problema de minimización. Por ejemplo, en la figura 4.4 representamos las tres primera puntuaciones discriminantes en un problema de clasificación respecto a 7 categorías, desechando pues la información aportada por los tres últimos ejes discriminantes. Para determinar si el eje discriminante k-ésimo puede ser despreciado podríamos en principio resolver un contraste de hipótesis del tipo $H_0(k) : \theta_k = 0$, según se ve en el apartado 3.3.1. No obstante, este método requiere de supuestos teóricos relativos a la distribución de los datos y, además, es muy conservador. Lo habitual es ponderar desde un punto de vista puramente muestral los autovalores $t_1, \ldots, t_b$ que, a su vez, se asocian según (3.24) a los coeficientes de correlación canónica $r_1, \ldots, r_b$.

En el caso de irisdata (figura 4.3), podemos apreciar que el peso de la discriminación recae casi exclusivamente en la primera puntuación discriminante, según sabíamos ya por el cuadro 3.5. En la figura 4.4 (izquierda) se aprecia cierta confusión entre algunas de las variedades de aceituna a partir de 17 variables numéricas medidas[2] al representar las dos primeras puntuaciones discriminantes. Sin embargo, la confusión se resuelve en parte al introducir la tercera puntuación, como se aprecia en la figura de la derecha.

4.2.2. Estrategia cuadrática de Fisher

Se trata de una generalización inmediata de la estrategia lineal. Se asume igualmente la hipótesis de p-normalidad pero no se asume la igualdad de las r matrices de covarianzas. En consecuencia, la estrategia consiste en modificar (4.16) reemplazando la estimación conjunta de la matriz Σ por las diferentes estimaciones S_i de cada una de las matrices Σ_i. Así pues, la estrategia consiste en asignar Y a la media $\overline{\mathtt{y}}_{i\cdot}$ cuando, para $j \neq i$, se verifica

$$(Y - \overline{\mathtt{y}}_{i\cdot})' S_i^{-1} (Y - \overline{\mathtt{y}}_{i\cdot}) \le (Y - \overline{\mathtt{y}}_{j\cdot})' S_j^{-1} (Y - \overline{\mathtt{y}}_{j\cdot}) \tag{4.21}$$

[2]Datos recogidos en el INTAEX de Badajoz.

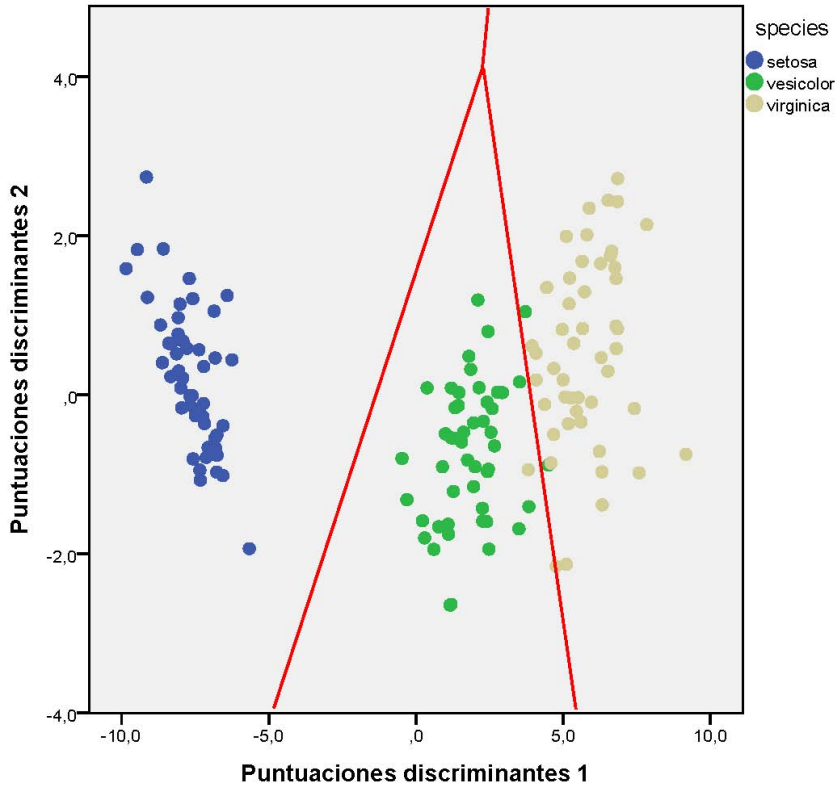


Figura 4.3: LDA en irisdata

Al contrario de lo que sucedía en la estrategia lineal, los términos cuadráticos no se anulan en la inecuación, de ahí el nombre. Para una mejor comprensión de la estrategia, podemos proyectar los datos sobre los primeros ejes discriminantes, aunque el gráfico no podrá interpretarse en los términos anteriores.

En el caso de la clasificación de aceitunas según sus variedades, la estrategia cuadrática de Fisher disminuye ligeramente los riesgos de la lineal, al menos según es estima mediante la reclasificación. No obstante, precisamos de métodos alternativos de clasificación que presenten diferencias más radicales respecto al LDA. Estudiaremos muy brevemente algunos de ellos en la siguiente sección. Las diferentes variantes del método de Fisher se ejecutan en SPSS a través del cuadro de diálogos 3.8.

4.3. Métodos alternativos

En esta sección expondremos esquemáticamente los tres métodos de clasificación alternativos al LDA más populares, posiblemente, al margen de aquéllos relacionados con la estimación de densidades.

4.3.1. Regresión logística

Hemos visto que el método LDA puede considerarse el reverso del manova, el test principal del modelo lineal. Sin embargo, la regresión logística binaria corresponde a un modelo lineal

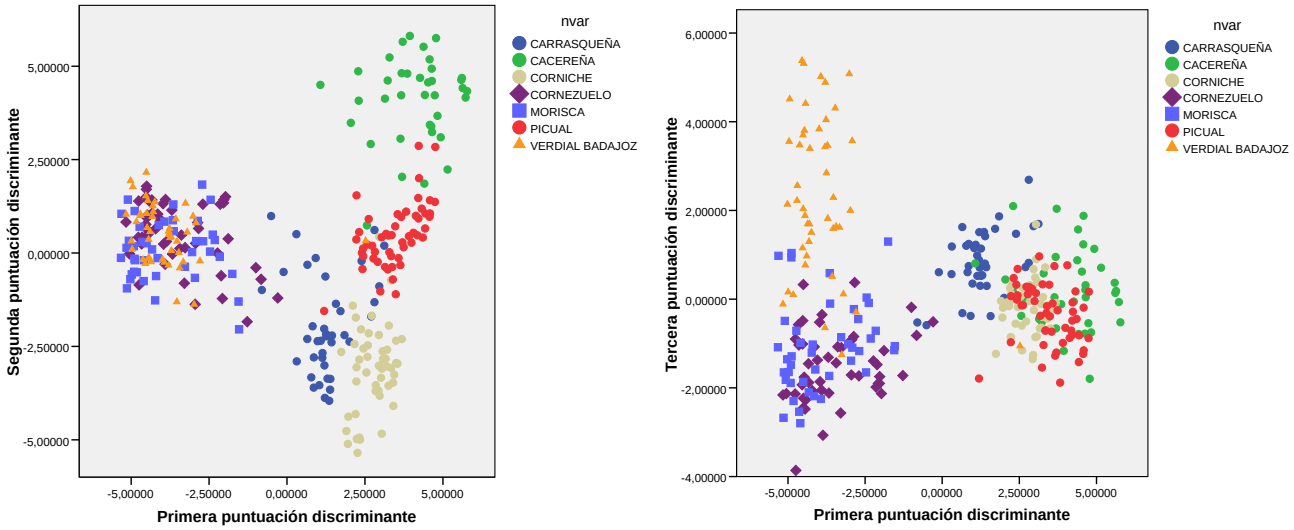


Figura 4.4: Variedades de aceituna: puntuaciones discriminantes 1, 2 y 3

generalizado. En el primero el parámetro principal se estima mediante (2.45)como solución a un sistema de ecuaciones lineales y, en el segundo, como a aproximación a un sistema de ecuaciones no lineales. Vistos con perspectivas, no deberíamos hablar de un modelo LDA y un modelo de reresión logística binaria, sino entenderlos como métodos diferentes para estimar la probabilidad *a posteriori* (4.11) partiendo de un mismo modelo origen. En todo caso, no es de extrañar pues que la regresión logística binaria ofrezca resultados muy similares a la estrategia LDA Bayesiana, tal y como se aprecia en la figura 4.5. De ahí que la regresión logística no sea, en propiedad, una alternativa al método de Fisher.

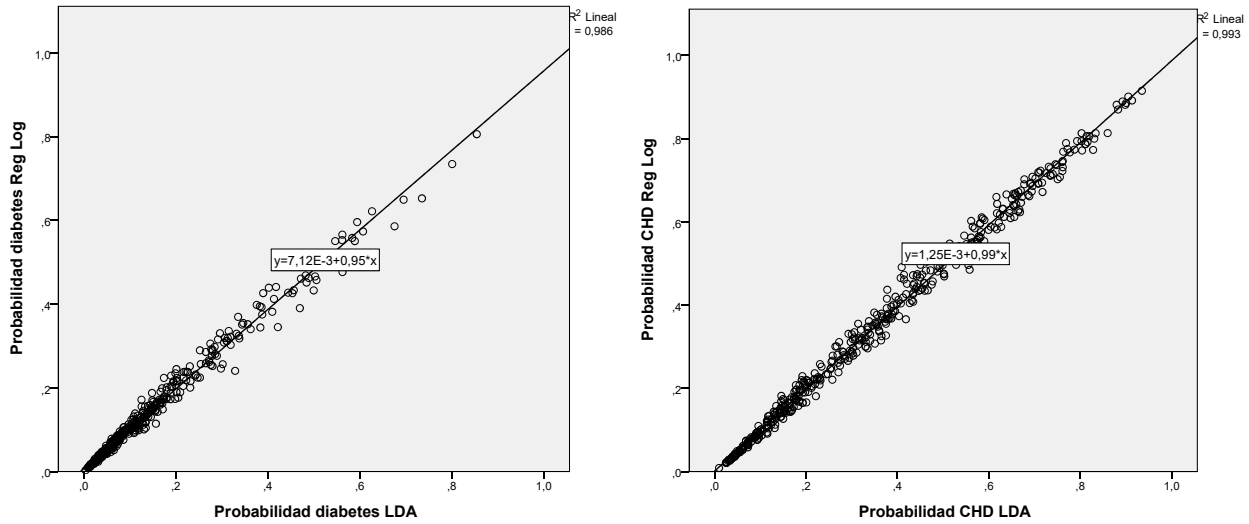


Figura 4.5: Comparativa entre las probabilides de dabetes tipo II (izquierda) y enfermedad coronaria (derecha) predichas mediante LDA Bayesiano y regresión logística, a partir de los datos de `Diabetes Data Study (Schoderling)` y `South African Heart Disease Study (Rousseauw)`, respectivamente.

Si el factor cualitativo distingue $r > 2$ categorías podemos aplicar el método de regresión logística multinomial. A grandes rasgos, consiste en una composición de $r - 1$ regresiones logísticas tomando una categoría como referencia. Cada una de estas regresiones permite estimar la probabilidad de que un dato concreto pertenezca a una categoría dada, dividida por la probabilidad de que pertenezca a la categoría de referencia. Si los $r - 1$ cocientes resultan ser inferiores a 1, el dato se asigna a la categoría de referencia; en caso contrario, se asigna a la que aporte un cociente máximo.

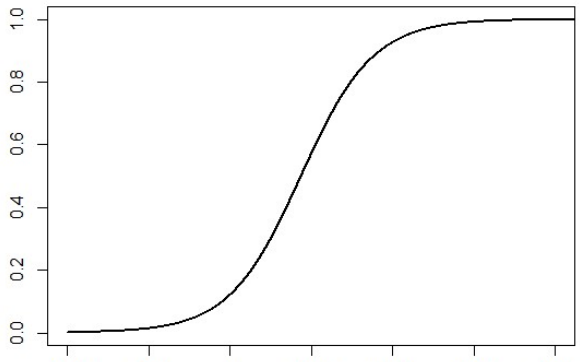


Figura 4.6: Función logística

A pesar de lo dicho anteriomente, en los programas estadísticos habituales se observan claras diferencias en las salidas ofrecidas según el método empleado (LDA o regresión logística binaria). La regresión logística aporta directamente estimaciones de Odds Ratios, un concepto de gran interés en Epidemiología. Además, el método viene tradicionalemente acompañado del test de Hosmer-Lemeshov, que contrasta la veracidad de la estimación de la distribución condicional de la respuesta Y dado el vector explicativo. Dicho test consiste en dividir la muestra en 10 subconjuntos en función de los deciles de la probabilidad estimada de $Y = 1$, y contrastar mediante un estadístico χ^2 con 9 grados de libertad si las proporciones reales de datos con $Y = 1$ en dichos subconjuntos se ajustan a lo que marcan las probabilidades medias de los mismos. Un resultado significativo en este test se traduce en una ineptitud del modelo de regresión logística para estimar la distribución condicional de la variable categórica y, en particular, para llevar a cabo la clasificación, lo cual nos conduciría a ensayar con estrategias alternativas de diferente naturaleza, como las que veremos a continuación. Realmente, nada impediría implementar este test en el método de Fisher, entre otros. Por otra parte, la abundancia de resultados no significativos en el test de Hosmer-Lemeshow, incluso con grandes muestras, revela la relativa intrascendencia en la elección del modelo estadístico si la compramos con el grado de correlación entre las variables observadas y la respuesta Y. Es más, aplicar este test sin tener en cuenta las condiciones de validez del test χ^2 implica una penalización de las correlaciones altas mediante resultados significativos. En definitiva, se trata de un test que debe interpretarse con mucho cuidado pues puede generar confusión.

4.3.2. Vecino más próximo

Dado un valor $k = 1, 2 \ldots$, el método del vecino más próximo para k (K-NN) es un procedimiento de clasificación no paramétrico pues no impone supuesto alguno sobre la distribución

de los datos, salvo que las variables medidas deben ser numéricas. El método se estudia con detalle en Hastie et al. (2008). Resumidamente, se trata de asignar una observación $Y \in \mathbb{R}^p$ a la categoría que tenga mayor presencia en el entorno de Y constituido por las k observaciones de la muestra más próximas. La cercanía se evalúa en principio en función de la métrica Euclídea en $\mathbb{R}^p$, aunque pueden considerarse alternativas. La estrategia depende en gran medida del valor de k seleccionado, de ahí que, como mínimo, se precise tantear con diferentes valores y seleccionar aquél cuya estrategia ofrezca menores riesgos estimados.

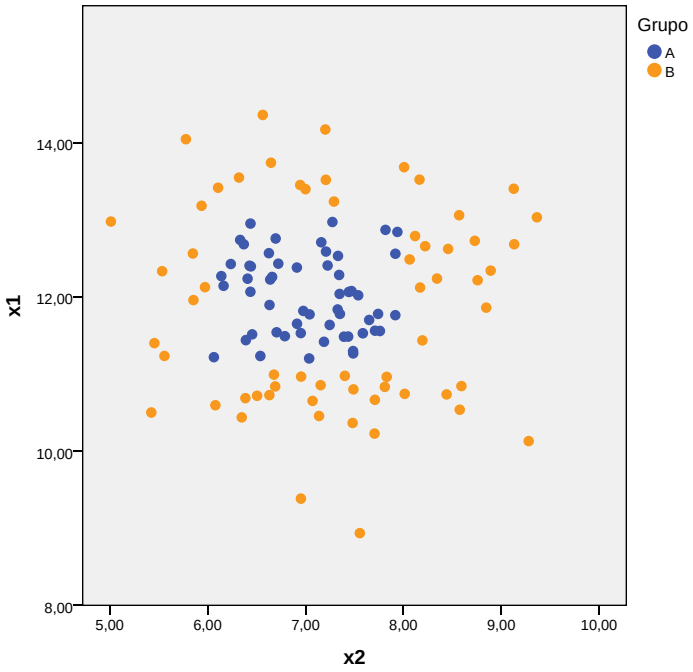


Figura 4.7: K-NN 94 %

En la figura 4.7 persentamos el diagrama de dispersión relativo a un vector con $p = 2$ variables en el cual distinguimos 2 categorías. Si pretendemos determinar una estrategia de asignación a categorías, los riesgos estimados por reclasificación de la muestra son del 43 % para LDA, 24 % para la alternativa cuadrática, 45 % para la regresión logística binaria y 6 % para K-NN con $k = 3$. En casos como éste se precisa pues una alternativa radicalmente diferente a LDA.

▪ *Ejercicio* 72. ¿Cómo se explica que la alternativa cuadrática de Fisher mejore sustancialmente la estrategia lineal en el ejemplo de la figura 4.7?

▪ *Ejercicio* 73. ¿Influye la escala en que se miden las componentes de Y en la clasificación según el método LDA? ¿Influye en la regresión logística? ¿Influye en el método K-NN? Caso de influir, ¿cómo podríamos compensar ese hecho?

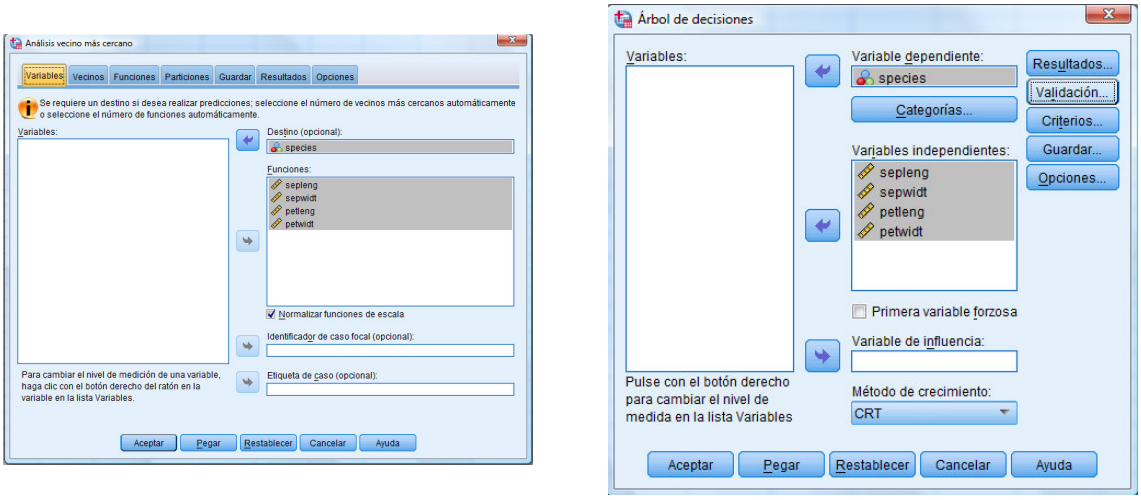


Figura 4.8: Vécino más próximio y árbol de decisión para irisdata

4.3.3. Árbol de decisión

Se trata de otro método no paramétrico (aunque lo más apropiado sería calificarlo, al igual que K-NN, como un procedimiento típico de Minería de Datos) que se puede resumir como una categorización óptima en la toma de decisión. Se fundamenta, como veremos a continuación, en una idea bastante sencilla. Sin embargo, ante la ausencia de supuestos estadísticos, la optimización se realiza mediante una gran cantidad de ensayos, por lo que no se concibe sin el uso de un ordenador. De ahí que métodos como éste o como el del vecino más próximo sean relativamente recientes. Se estudia con detalle en Hastie et al. (2008).

Predicción categórica de una variable numérica: El algoritmo que explicaremos a continuación fue descrito en Breiman et al. (1984), y se denomina CART o CRT[3].. Se diseña originalmente para una variable respuesta Y de tipo numérico y un vector p-dimensional X explicativo y se fundamenta intuitivamente en la propia definición de integral de Riemann como límite de sumas finitas de áreas de rectángulos. Más generalmente, tenemos en cuenta que cualquier función en L^2 puede aproximarse tanto como se quiera, según la métrica (1.4), mediante una suma de funciones constantes sobre intervalos, si el espacio origen es $\mathbb{R}$ o, en general, sobre rectángulos medibles, si es $\mathbb{R}^p$.

Dividir el espacio origen $\mathbb{R}^p$ en rectángulos medibles se traduce en la práctica en una categorización conjunta de las p variables explicativas numéricas, ideada para ajustarse lo mejor posible (en sentido mínimo cuadrático) a la variable numérica respuesta Y. El método CRT es de tipo binario pues el vector X será categorizado mediante dicotomizaciones o biparticiones sucesivas[4]. En la figura 4.9 podemos apreciar cómo funciona una dicotomización de este tipo para el caso $p = 1$. Ésta se entiende a su vez como una ramificación de un nodo único (nodo parental) en dos ramas o categorías (nodos filiales), a las cuales se les asigna un valor constante como aproximación a la variable respuesta Y.

Sabemos (ejercicio 3) que, si queremos aproximar en L^2 la variable Y mediante una constante, ésta debe ser su media respecto a la probabilidad P (en términos muestrales, su media aritmé-

[3]Classification and Regressión Trees

[4]Cosa que no ocurre con otros métodos como CHAID.

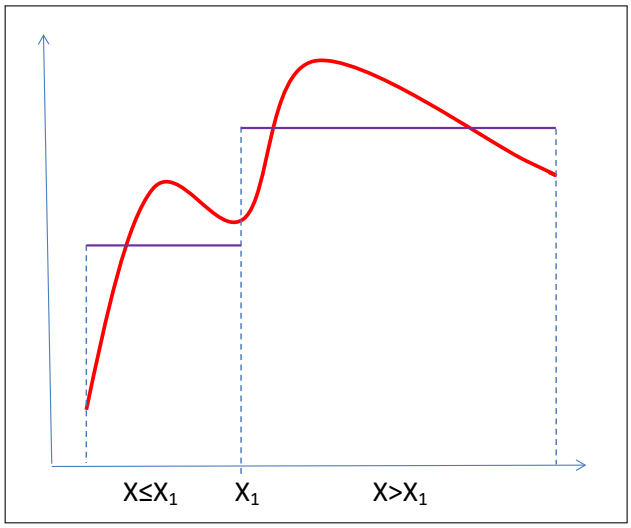


Figura 4.9: Dicotomización de una variable explicativa

tica), siendo su varianza la medida del error cuadrático. Podemos conseguir un mejor ajuste si consideramos dos rectángulos, es decir, un punto de corte $\mathtt{X}_1$. En tal caso consideraríamos como constantes a cada lado del mismo las respectivas medias. La mejoría en términos de L^2 sería la diferencia entre el error cuadrático, es decir, la varianza total por la medida del segmento completo, y la suma de las nuevas varianzas multiplicadas por las medidas de los respectivos subsegmentos delimitados por $\mathtt{X}_1$. El punto de corte óptimo $\mathtt{X}_1$ será aquél que maximice esa ganancia. Desde un punto de vista computacional, el problema se reduce a ensayar diferentes puntos de cortes a lo largo del eje X en busca de la mínima media ponderara entre las varianzas resultantes. Si tenemos p variables debemos ensayar a lo largo de cada uno de los ejes explicativos.

Una vez encontrado el primer punto de corte óptimo, debemos proceder de igual forma para obtener un segundo punto de corte $\mathtt{X}_2$ en alguna de las variables explicativas, lo cual supondrá un nuevo nivel de ramificación en algún nodo ya existente, y así sucesivamente. Las medias de los nodos terminales resultantes nos darán la aproximación categórica a Y. En la figura 4.10 se representa esquemáticamente cuatro iteraciones de ramificación en un problema con $p = 2$ variables (ejes) explicativas. Ello da lugar a cinco nodos terminales o categorías diferentes. La mayor proximidad entre colores simboliza un separación más tardía entre los respectivos nodos.

Dejando a un lado la idea que inspira el método, que es integral de Riemann, entenderemos en lo sucesivo el procedimiento como un algoritmo iterativo de dicotomización, que busca en cada paso la mínima media ponderada de las varianzas ponderada resultantes al reemplazar el valor medio del nodo parental por los valores medios de los nodos filiales. Este algoritmo puede aplicarse pues sin problema en presencia de variables explicativas de tipo cualitativo. Por lo tanto, no implica la asunción de supuesto alguno respecto al conjunto de p variables explicativas.

Sobreajuste: Dado que, en la práctica, no contamos con una variable $Y \in L^2$ sino con una muestra $\mathtt{Y}$ de tamaño n de la misma, bastaría considerar un árbol en el que cada elemento de la muestra constituyera un nodo terminal para conseguir un ajuste perfecto con error cuadrático nulo. Las situaciones de ese tipo se denominan sobreajustes y debe evitarse a toda costa.

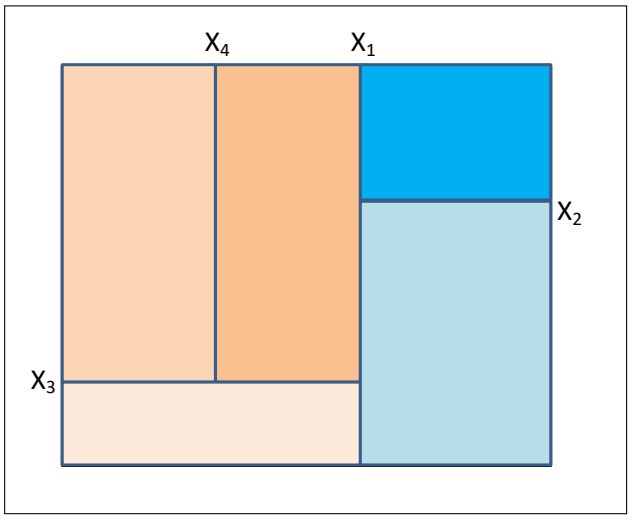


Figura 4.10: Esquema de árbol con cuatro iteraciones en la ramificación

Efectivamente, en el contexto de la Minería de Datos, no valoramos una estrategia $\mathcal{S}$ concreta sino el algoritmo que la genera. Es decir, nuestra intención es programar lo mejor posible un algoritmo que, dada una muestra ξ de tamaño n de (X, Y) y una observación independiente $\mathtt{X}_0$ de X, determine a partir de ξ una estrategia $\mathcal{S}_\xi$ para predecir el valor de Y correspondiente a $X = \mathtt{X}_0$. Así pues, una vez establecido el algoritmo, la predicción $\hat{Y} = \mathcal{S}_\xi(\mathtt{X}_0)$ se entiende como una función real de ξ y $\mathtt{X}_0$. Podemos entender intuitivamente que la varianza de $\hat{Y}|_{X=\mathtt{x}_0}$ crece según aumenta el grado de sobreajuste en el algoritmo programado, porque aquello que está diseñado para explicar demasiado bien una muestra concreta ξ puede diferir mucho de lo que está diseñado para explicar otra distinta ξ'. En ese sentido, el sobreajuste es el enemigo oculto en la Minería de Datos al igual que la violación de los supuestos del modelo lo era en la Inferencia Paramétrica.

Para evitar esta patología podemos optar por imponer una cantidad mínima de datos por nodo. De esta manera renunciamos a árboles complejos a menos que trabajemos con grandes muestras. También pueden limitarse los niveles de ramificación aplicando mecanismos automáticos de "poda" a partir de un árbol con error 0. La intensidad de dicha poda puede modularse a partir de cierto parámetro α cuyo valor mínimo es 0 (consultar los detalles en Hastie et al. (2008)).

Predicción categórica de una variable cualitativa: Pero en este capítulo estamos intentando aproximarnos al valor de una variables respuesta Y que no es numérica sino cualitativa, con r posibles categorías. El algoritmo del árbol de decisión puede aplicarse también en ese caso con una serie de modificaciones bastante naturales. Si Y es una variable cualitativa el algoritmo consiste en asignar a cada nodo filial la categoría de Y predominante en el mismo (en lugar de la media). La dicotomización se realiza en cada paso de manera que se logre la máxima reducción en una medida de error de clasificación expresada en función de cierto índice. Existen diferentes formas de medir dicho error [5]. En este manual y en cada nodo nos decantamos por el denominado índice de Gini que se define como sigue: si en cierto nodo $\mathcal{N}$ se denota por $\hat{p}_{\mathcal{N}}(i)$ la proporción de datos pertenecientes de la categoría i de Y en $\mathcal{N}$, se define

[5] Hastie et al. (2008)

la medida de Gini de $\mathcal{N}$ mediante

$$G(\mathcal{N}) = \sum_{i=1}^{r} \hat{p}_{\mathcal{N}}(i)\big(1 - \hat{p}_{\mathcal{N}}(i)\big) \tag{4.22}$$

Nótese que, si todos los datos del nodo $\mathcal{N}$ pertenecieran a una cierta categoría, los r sumandos de $G(\mathcal{N})$ serían nulos. Lo mismo ocurriría si ninguno perteneciera a i. Es más, para cada i, $\hat{p}_{\mathcal{N}}(i)$ es la media de la función indicador que asigna 1 a los datos en i y 0, al resto, siendo $\hat{p}_{\mathcal{N}}(i)\big(1 - \hat{p}_{\mathcal{N}}(i)\big)$ su varianza. Así pues, el índice de Gini puede interpretarse como una suma de varianzas o, más concretamente, como una varianza en el sentido más amplio definido en (1.9), lo cual refuerza la analogía con el caso numérico. Así pues, la ganancia en términos de Gini al pasar de un nodo parental a dos filiales se define como la diferencia ente la medida de Gini (varianza) inicial y la media ponderada (en función de las proporciones de datos) de las varianzas filiales. A lo largo de las diferentes iteraciones podemos ir contabilizando, para cada variable, la ganancia que supondría en la misma una dicotomización óptima según el índice de Gini (aunque en la práctica sólo se generará una ramificación en la variable que logre la máxima reducción). Este contador acumulado da lugar a la denominada función de importancia.

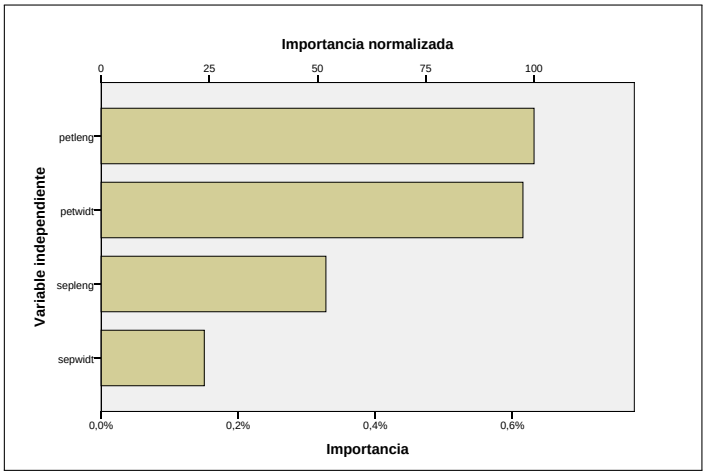


Figura 4.11: Función de importancia para iris data

4.3.4. Validación de estrategias:

Como ya hemos comentado, los métodos propios de la Minería de Datos, como K-NN y el árbol de decisión, no se basan en suposiciones sobre las distribuciones de los datos, sino que elaboran una explicación *ad hoc* de la clasificación observada en la propia muestra. Luego, si estimáramos los riesgos inherentes a la estrategia reclasificando directamente las observaciones de la muestra que hemos utilizado para diseñarla, estaríamos incentivando claramente el sobreajuste, lo cual conduciría a un ingenuo optimismo sobre la validez del método. En este contexto resulta más apropiado estimar los riesgos a partir de datos cuya categoría de procedencia se conozca, pero que no se hayan utilizado para diseñar la estrategia. Por ejemplo, podemos dividir la muestra en dos partes, no necesariamente iguales: una para diseñar la estrategia (muestra de entrenamiento) y otra para validarla (muestra de validación). Es decir, los datos que se utilizan para estimar los riesgos de la estrategia no se han usado en el diseño de la misma. Si

finalmente estos riesgos se consideran apropiados, podemos reconstruir la estrategia contando con la muestra completa.

Por ejemplo, en la figura 4.12 podemos apreciar el árbol de decisión que ofrece el programa SPSS para determinar la especie a la que pertenece un lirio a partir de sus medidas de pétalo y sépalo. A partir de (aproximadamente) la mitad la muestra y mediante dos iteraciones (ya que en el algoritmo hemos impuesto un tamaño mínimo de los nodos de 10) se ha diseñado un árbol de decisión con tres nodos terminales. Para este árbol se estiman, mediante el resto de la muestra, los riesgos de clasificación errónea, que son concretamente del del 0 % para setosa, 9 % para vesicolor y del 15 % para virgínica. En la figura 4.11 se expresa la función de importancia asociada al árbol.

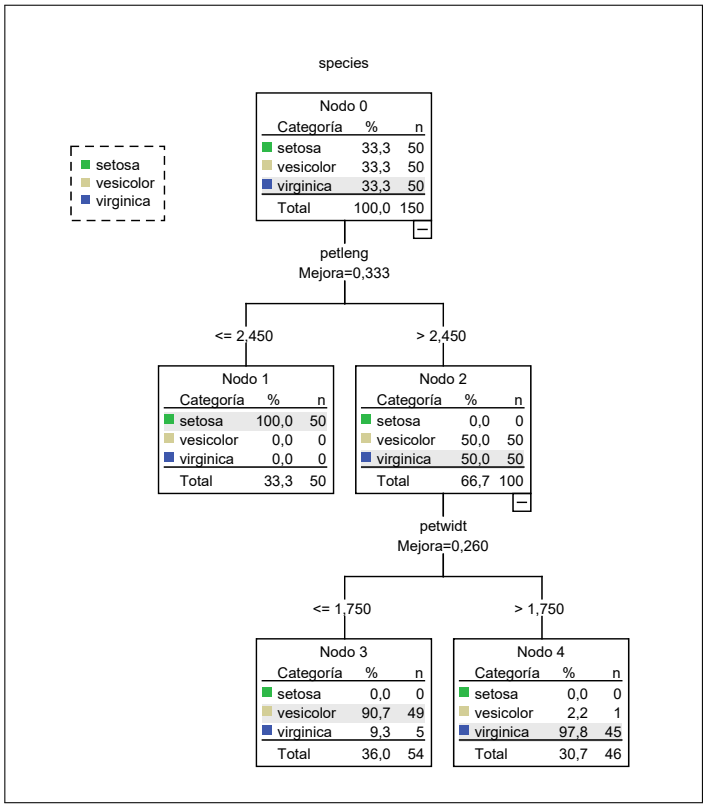


Figura 4.12: Árbol de decisión para irisdata

Técnicas derivadas del árbol de decisión: El árbol de decisión 4.12 se ejecuta en SPSS a través del cuadro de diálogos 4.8. Existen otros métodos más sofisticados basados en los mismos principios que son los preferidos por numerosos autores (ver Hastie et al. (2008)) para afrontar un problema de clasificación: *Random Forest* y *Adaboost*, ejecutables en R mediante las librerías `randomForest` y `Adabag`, repectivamente. El primero consiste básicamente en diseñar diferentes árboles simples a partir de submuestras escogidas al azar (bootstrap) y con un determinado número de variables escogidas también al azar. Cada árbol simple consiste en una única partición binaria escogiendo la variable y el punto óptimo; en el árbol complejo final se clasifica la observación en la categoría mayoritaria entre todos los árboles simples diseñados. Por su parte, el método adaboost itera un árbol binario incrementando el peso de los datos mal clasificados; cada árbol binario se pondera a su vez en función de su capacidad predictiva, de

manera que se establece como estrategia de clasificación la suma ponderada de los diferentes árboles binarios.

4.3.5. Indicaciones sobre Redes Neuronales

Por último, mencionamos otro conjunto de técnicas de regresión y clasificación alternativa a las anteriores denominada genéricamente por Redes Neuronales, que consisten en explicar la variable respuesta en función de una o varias capas sucesivas de variables ocultas que se calculan a partir de las observadas, combinando funciones lineales con otras no lineales muy específicas.

El modelo más simple, con una única capa, consiste en explicar un vector respuesta $P = (P_1, \ldots, P_k)'$ a partir de las variables observadas $Y = (Y_1, \ldots, Y_p')$ como $P = g \circ T \circ Z(Y)$, donde $g = (g_1, \ldots, g_k)'$, $Z = (Z_1, \ldots, Z_M)'$ y $T = (T_1, \ldots, T_k)'$. Las opciones más habituales para un problema de clasificación respecto a $k+1$ categorías son las siguientes:

$$Z_m(Y) = \mathtt{L}(\alpha_{0m} + \alpha_{1m}' Y), \quad m = 1, \ldots, M \tag{4.23}$$

$$T_k(Z) = \beta_{0k} + \beta_{1k}' Z, \quad k = 1, \ldots, K \tag{4.24}$$

$$g_k(T_k) = \frac{e^{T_k}}{1 + e^{T_k}}, \quad k = 1, \ldots, K \tag{4.25}$$

En ese caso, $P_k(Y)$ se entiende como la probabilidad estimada de que la observación Y pertenezca a la categoría k-ésima del factor respuesta. Nótese que, si $K = M = 1$ y, además, $g = \mathtt{Id}$ (que es la opción que se suele considerar por defecto), el método consistiría e una regresión logística binaria, luego podríamos considerar las redes neuronales como una generalización en sentido muy amplio de dicha técnica.

Los detalles sobre la estimación de estos parámetros podemos encontrarlos en la sección 11.4 de Hastie et al. (2008). Se trata posiblemente del método de clasificación más flexible y se aplica con frecuencia al reconocimiento de imágenes. Por contra, genera un efecto denominado de "caja negra" que impide entender la influencia de las variables observadas en la respuesta. En la figura 4.13 podemos apreciar la red neuronal utilizada para pronosticar la probabilidad de enfermedad coronaria a partir de los datos de `South African Heart Disease Study`.

Si efectuamos una comparativa entre los métodos estudiados (LDA, regresión logística, vecino más próximo, árbol de decisión y redes neuronales) a través de los datos de `Southafrica Heart Disease`, observamos en el cuadro 4.3 muy altas correlaciones entre las probabilidades pronosticadas de LDA Bayesiano, regresión logística binaria y red neuronal (muy especialmente entre los dos primeros), como cabría esperar por lo razonado anteriormente. Además, los porcentajes de clasificación correcta para estos métodos son muy similares (en torno al 85 % para los sanos y al 55 % para los enfermos coronarios); para el método LDA minimax los porcentajes de clasificación correcta son del 70 % para los sanos y 73 % para los enfermos; los otros dos métodos (KNN y árbol) correlacionan menos entre sí y con los demás. Presentan porcentajes de clasificación correcta en torno al 95 % para los sanos y al 25 % para los enfermos, calculadas a partir de las muestras de validación.

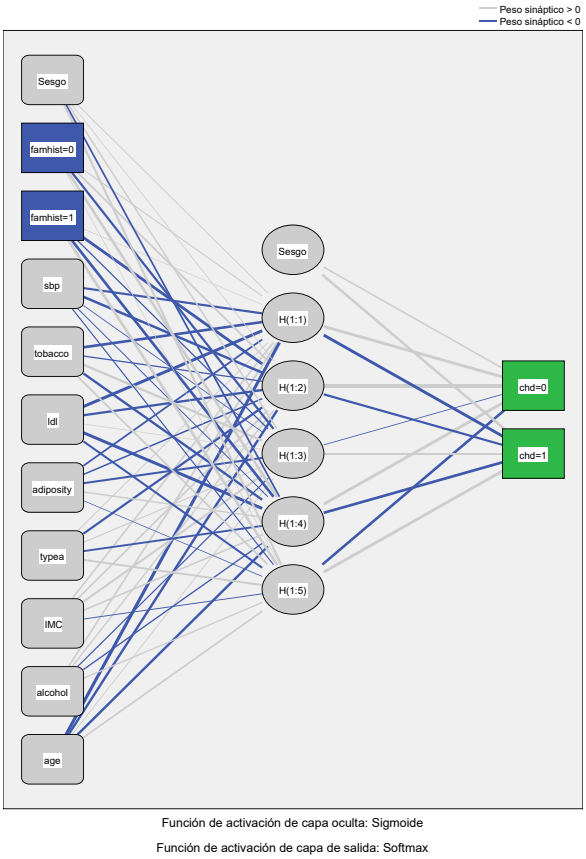


Figura 4.13: Probabilidad de enfermedad coronaria según red neuronal

	Reg Log	KNN	Árbol	Red Neuro.
LDA	0.997	0.631	0.735	0.981
Log		0.639	0.738	0.980
KNN			0.524	0.622
Árbol				0.725

Cuadro 4.3: Correlaciones entre probabilidades pronosticadas de enfermedad coronaria

Capítulo 5

Reducción dimensional

En este capítulo, el único exclusivamente multivariante, se recogen una serie de técnicas que tienen por objetivo simplificar un conjunto de datos multidimensional, aunque centraremos mayormente nuestro interés en el análisis de componentes principales. El denominador común de estas técnicas es que, en última instancia, se fundamentan en el teorema 1.5.1. Hay que destacar que un estudio de este tipo se enmarca en una fase inductiva, pues el producto final no es la aplicación de un test de hipótesis sino un gráfico en dimensiones reducidas que permita una primera visión global de nuestra muestra, a partir de la cual podamos formular distintas hipótesis. Por ello se ha optado por una redacción de la mayor parte del capítulo en lenguaje muestral.

Ya hemos comentado que el núcleo del capítulo lo constituye el análisis de componentes principales, aunque esta técnica puede considerarse un opción particular del denominado análisis factorial. No parece existir unanimidad de criterios a la hora de catalogar y distinguir los diferentes métodos. En este volumen se ha optado por presentarlos como un única técnica de reducción dimensional de carácter dual, que se denomina análisis de componentes principales desde el punto de vista de la observaciones (filas) y análisis factorial desde el punto de vista de la variables (columnas). Aunque existen variedades del análisis factorial al margen del análisis de componentes principales, nos hemos limitado a introducirlas brevemente. Así mismo, el análisis de correspondencias, que en cierta forma puede entenderse como una extensión del análisis de componentes principales, se presenta de forma muy escueta, y más aún su generalización, que es el análisis de correspondencias múltiple. Por último, se ha incluido una sección en la que se ilustra brevemente la utilidad del uso de componentes principales en un problema de regresión lineal múltiple.

Desde el punto de vista teórico, debemos entender en esencia que el teorema 1.5.1 viene a proponer un cambio de la base vectorial inicial, dada por las variables medidas, a una base ortonormal constituida por los autovectores de la matriz de covarianzas. De esta transformación obtenemos nuevas variables incorreladas entre sí que se denominarán componentes principales, lo cual puede permitir una comprensión más fácil de nuestros datos.

5.1. Componentes principales

Antes de abordar un estudio de carácter muestral y con un propósito meramente didáctico, empezaremos con una definición probabilística de las componentes principales.

Consideremos un vector aleatorio p-dimensional X de media nula y matriz de covarianzas Σ,

la cual descompone según el teorema 1.5.1 en $\Gamma\Delta\Gamma'$, con $\Gamma = (\gamma_1 \dots \gamma_p)$ y $\Delta = \mathrm{diag}(\delta_1, \dots, \delta_p)$. Para cada $j = 1, \dots, p$ se define la j-ésima componente principal mediante

$$U_j = \gamma_j' X \tag{5.1}$$

De esta forma, las p componentes principales ordenadas componen un vector aleatorio p-dimensional U que se obtiene como transformación lineal (rotación en sentido amplio) del vector original X, concretamente mediante $U = \Gamma' X$, y que verifica, por lo tanto, $\mathtt{Cov}(U) = \Delta$. De ello se deduce que las componentes principales son incorreladas entre sí. También se deduce del teorema 1.5.1 que, de todas las posibles proyecciones de X sobre algún eje de $\mathbb{R}^p$, U_1 es la que presenta una máxima varianza. U_1 es concretamente la proyección sobre el eje $\langle \gamma_1 \rangle$ y su varianza es δ_1.

- *Ejercicio* 74. Probar que, dada una proyección de U sobre un determinado eje $\langle \gamma \rangle$, se verifica que es incorrelada con U_1 si, y sólo si, $\gamma \perp \gamma_1$.
- *Ejercicio* 75. Probar que, de todas las posibles proyecciones de X sobre algún eje de $\mathbb{R}^p$ incorreladas con U_1, la que presenta una varianza máxima es U_2.

Así se van obteniendo sucesivamente las componentes principales hasta U_p. Así pues, "pasar a componentes principales" se entiende como considerar, en lugar del vector original X, otro vector transformado o "rotado" U cuyas componentes son incorreladas y están ordenadas en función de sus varianzas. ¿Qué ventajas comportan ambas cualidades? Por el momento nos limitaremos a aportar una explicación intuitiva, si bien el el resto del capítulo presentaremos un razonamiento más formal: primeramente, las ventajas que se derivan de la propiedad probabilística o estadística de incorrelación entre las componentes principales (lo cual bajo el supuesto de p-normalidad equivaldría a independencia) son las mismas que se derivan de la propiedad vectorial de la ortogonalidad. Más específicamente, dado que nos proporciona una clara descomposición de la varianza total, permite conocer con precisión que porcentaje de la misma perdemos si eliminamos una componente principal concreta. Segundo, y en lo que respecta precisamente a la eliminación de componentes principales, serán las últimas, es decir, las de menor varianza, las candidatas a ser suprimidas. Estamos afirmando pues que la importancia de las diferentes componentes principales viene dada por la magnitud de su varianza y que, por lo tanto, el orden en el que se obtienen establece una jerarquía real desde el punto de vista de la reducción dimensional. Intentaremos justificar tal afirmación desde un punto de vista geométrico en la siguiente sección.

5.2. Justificación geométrica

En esta sección se trata el problema de simplificar la dimensión de un conjunto de datos, tanto desde el punto de vista de las observaciones, es decir, de las filas de la matriz (1.1), como de las variables, es decir, de las columnas. Debemos tener presente la notación introducida en el capítulo preliminar y, en especial, la distancia (1.36) definida en el conjunto de las matrices de dimensión $n \times p$. Por otra parte, dado $k \leq p$, se denota por $\mathcal{H}_k$ el conjunto de las subvariedades afines de dimensión k en $\mathbb{R}^p$. Dada una matriz $\mathtt{X} \in \mathcal{M}_{n\times p}$, con matriz de covarianzas muestral S, consideraremos su descomposición tipo (1.40) en función de sus autovalores $d_1, \dots, d_p$ y sus respectivos autovectores, $g_1, \dots, g_p$, pero distinguiendo entre los k primeros y los $p-k$ restantes mediante

$$S = (G_1|G_2) \left(\begin{array}{c|c} D_1 & 0 \\ \hline 0 & D_2 \end{array} \right) \left(\begin{array}{c} G_1' \\ \hline G_2' \end{array} \right) \tag{5.2}$$

El análisis de componentes principales (PCA) se basa en el siguiente resultado, que es consecuencia del lema 1.5.2:

Teorema 5.2.1. Dados $0 \leq k \leq p$ y $\mathtt{X} \in \mathcal{M}_{n\times p}$ con matriz de covarianzas S que descompone según (5.2), se verifica

$$\text{mín}\left\{\mathtt{d}^2_{n,p}(\mathtt{X},\mathtt{X}^k) : \ \mathtt{X}^k_i \in H \ \forall i \leq n \text{ para algún } H \in \mathcal{H}_k \ \right\} = \mathtt{tr}(D_2), \tag{5.3}$$

y se alcanza con $\mathtt{X}^k_i = \overline{\mathtt{x}} + P_{\langle G_1\rangle}(\mathtt{X}_i - \overline{\mathtt{x}}) \ \forall i \leq n$.

La figura 5.1 ilustra el teorema anterior para un problema bidimensional.

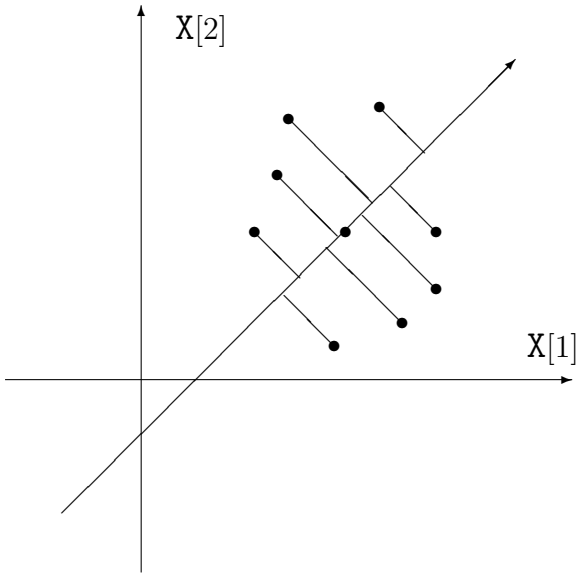


Figura 5.1: Proyección de observaciones

En el caso trivial $k = 0$, el teorema afirma que el vector de $\mathbb{R}^p$ constante por el que debemos reemplazar las observaciones $\mathtt{X}_i$ con el menor error cuadrático posible es la media aritmética $\overline{\mathtt{x}}$, siendo la varianza total muestral, definida en (1.37), la medida de dicho error.

▪ *Ejercicio* 76. Probar que $s^2_T = \sum_{j=1}^p d_j$

A medida que aumenta el valor de k, la distancia respecto a la simplificación $\mathtt{X}^k$ disminuye. Expresándolo de manera inversa diríamos que al reemplazar las observaciones originales se explica sólo una parte de la varianza total, concretamente $\sum_{j=1}^k d_i$. Esta cantidad es máxima respecto al conjuno de subvariedades afines k-dimensionales y se alcanza proyectando sobre el subespacio generado por los k primeros autovectores de S, salvo traslaciones. La proporción de varianza total explicada por los mismos es pues

$$\frac{\sum_{j=1}^k d_i}{\mathtt{tr}(S)} \tag{5.4}$$

Que esta proporción sea próxima a 1 para un valor de k pequeño se traduce en que las n observaciones reales se encuentran muy próximas a una subvariedad afín de baja dimensión, lo

cual equivale a un fuerte grado de correlación lineal (afín) entre las variables medidas para los datos observados. Ésa es la cualidad que permite obtener una reducción dimensional profunda mediante la técnica PCA, cosa que parece más factible bajo el supuesto de p-normalidad. En todo caso, el método puede generalizarse mediante el uso de las denominadas curvas y superficies principales, que pueden aproximarse mediante iteraciones descritas en Hastie et al. (2008), si bien es en Gifi (1990) donde se le dedica más espacio a este tipo de problema.

En lo que resta supondremos sin pérdida de generalidad (ya veremos por qué) que $\overline{\mathtt{x}} = 0$. En ese caso y para cada $j = 1, \ldots, p$, se denota $\mathtt{U}[j] = \mathtt{X} \cdot g_j$, es decir, el vector de $\mathbb{R}^n$ que recoge las proyecciones de cada observación $\mathtt{X}_i$, $i = 1, \ldots, n$, sobre el eje $\langle g_j \rangle$ determinado por el j-ésimo autovector de S. Dicho eje se denomina j-ésimo eje principal y $\mathtt{U}[j]$ se denomina j-ésima componente principal. Las p componentes principales ordenadas constituyen una nueva matriz $\mathtt{U} \in \mathcal{M}_{n\times p}$ definida mediante

$$\mathtt{U} = \mathtt{X}G \tag{5.5}$$

que expresa las coordenadas de $\mathtt{X}$ respecto de la base ortonormal canónica de autovectores G. Dado $k \leq p$, se denota por $\mathcal{U}$ la matriz $n \times k$ compuesta por las k componentes principales, cuyas filas y columnas se denotarán con el mismo criterio que en (1.1). Por otra parte, se denota $\mathtt{E} = (\mathtt{U}[k+1], \ldots, \mathtt{U}[p])G_2' \in \mathcal{M}_{n\times p}$. En ese caso, se sigue de (5.5) que

$$\mathtt{X} = \mathcal{U}G_1' + \mathtt{E} \tag{5.6}$$

siendo $\mathcal{U}G_1'$ la matriz en $M_{n\times p}$ que permite alcanzar las distancia mínima a $\mathtt{X}$ en el teorema 5.2.1.

▪ *Ejercicio* 77. Probar que las componentes principales son incorreladas entre sí. Probar que el primer eje principal es aquél sobre el que hay que proyectar las observaciones para obtener una máxima varianza, que vale d_1.

▪ *Ejercicio* 78. Probar que el segundo eje principal es aquél sobre el que hay que proyectar las observaciones para obtener la máxima varianza de entre todas la variables incorreladas con la primera componente principal, que vale d_2, y así sucesivamente. Probar que el último eje principal es aquél sobre el que hay que proyectar para obtener una mínima varianza.

▪ *Ejercicio* 79. ¿Cómo se interpreta $d_p = 0$? ¿Cómo se interpreta $|\Sigma| = 0$ para un vector aleatorio con distribución $N_p(\mu, \Sigma)$?

▪ *Ejercicio* 80. Probar que la matriz de covarianzas de $\mathcal{U}$ es $S_{\mathcal{U}} = D_1$.

Así pues, los ejes principales resuelven el problema de maximización de la varianza, mientras que los ejes discriminantes, estudiados en los capítulos 3 y 4, solucionan el problema de maximización relativo a la discriminación entre categorías, que a su vez puede entenderse como una maximización de correlaciones lineales. Esto puede llevar a una cierta confusión en problemas multivariantes con presencia de un factor cualitativo (como, por ejemplo, el caso del irisdata de Fisher). La figura 5.2 trata de clarificar la relación que se da en tales casos entre los distintos tipos de ejes, teniendo en cuenta que, según la notación utilizada en 3.13, la varianza total s^2 de una proyección dada, que también podemos denotar mediante SCT, descompone según la figura 2.5 mediante $SCT = SCH + SCE$.

Al contrario que en el caso de los ejes discriminantes, el problema de maximización de la varianza es sensible ante cambios de escala en las variables medidas. En la figura 5.3 se ilustra el efecto sobre el cálculo de los ejes principales de un cambio de escala en la variable del eje OY.

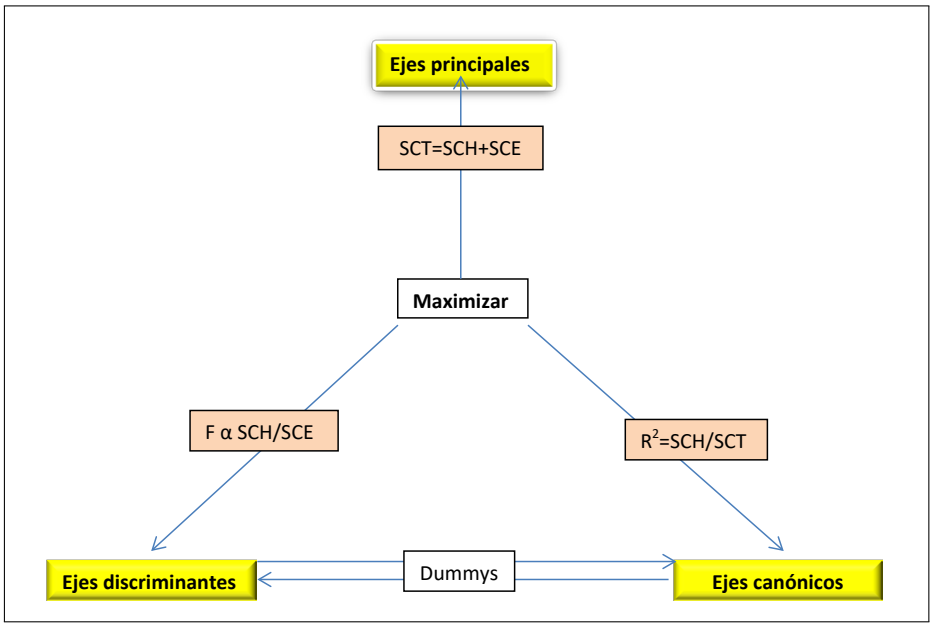


Figura 5.2: Ejes principales, canónicos y discriminantes

Sin embargo, podemos lograr artificialmente un método de reducción dimensional invariante ante cambios de escala si trabajamos con las variables tipificadas, que es lo que consideraremos por defecto. Ello equivale a trabajar en todo momento con la matriz de correlaciones muestral R en lugar de la matriz de covarianzas original S. Dado que $\mathtt{tr}(\mathsf{R}) = p$, (5.4) es igual a

$$\overline{d}(k) := \frac{\sum_{j=1}^{k} d_i}{p} \tag{5.7}$$

Al trabajar con variables tipificadas podemos suponer que la media es nula, como indicábamos antes.

5.3. Representación de observaciones y variables

Definimos la matriz $\mathtt{F} = \mathcal{U} D_1^{-1/2}$ de dimensiones $n \times k$ que se compone de los siguientes elementos[1]

$$\mathtt{F} = \begin{pmatrix} \mathtt{F}_1[1] & \dots & \mathtt{F}_1[k] \\ \vdots & & \vdots \\ \mathtt{F}_n[1] & \dots & \mathtt{F}_n[k] \end{pmatrix} \tag{5.8}$$

Para cada i, la transposición de la fila i-ésima de $\mathtt{F}$ es un vector de $\mathbb{R}^k$ que se denota[2] por $\overrightarrow{\mathtt{F}}_i$ y decimos que está compuesto por las k puntuaciones factoriales de la observación $\mathtt{X}_i$. Los diferentes vectores columnas de $\mathtt{F}$, que se denotan por $\overrightarrow{\mathtt{F}}[j] \in \mathbb{R}^n$, para $j = 1, \dots, p$, se

[1] El paso de la matriz $\mathcal{U}$ a $\mathtt{F}$ responde, como veremos más adelante, a la intención de facilitar la representación de las variables, aunque ello vaya en detrimento de la representación de observaciones.

[2] Con carácter excepcional distinguiremos en esta sección los vectores de los escalares mediante el uso de flechas en la notación.

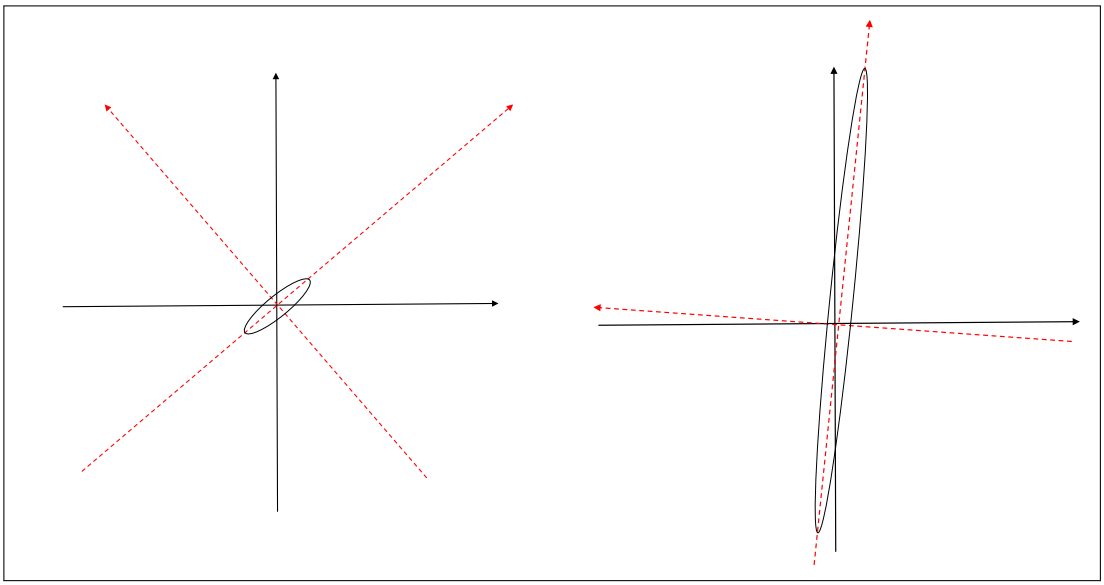


Figura 5.3: Efecto de un cambio de escala en los ejes principales

denominan ejes factoriales y, salvo una constante (dado que todos tienen norma Euclídea $\sqrt{n}$), constituyen un sistema ortonormal[3].

Por otra parte, para que se siga verificando la ecuación (5.6), transformamos de manera inversa G_1 definiendo $\Lambda = G_1 D_1^{1/2} \in \mathcal{M}_{p\times k}$. La matriz Λ se expresará así

$$\Lambda = \begin{pmatrix} \lambda_1[1] & \dots & \lambda_k[1] \\ \vdots & & \vdots \\ \lambda_1[p] & \dots & \lambda_k[p] \end{pmatrix} \tag{5.9}$$

Para cada $l = 1, \dots, p$, la transposición de la fila l-ésima de Λ es un vector de $\mathbb{R}^k$ que se denota por $\overrightarrow{\lambda}[l]$ y se denomina vector de las k componentes de la variable $\overrightarrow{\mathtt{X}}[l]$. Los diferentes vectores columnas de Λ se denotan por $\overrightarrow{\lambda}_j \in \mathbb{R}^p$, para $j = 1, \dots, k$, y determinan los k primeros ejes principales. Por lo tanto, constituyen un sistema ortogonal en $\mathbb{R}^p$. La ecuación (5.6) se expresa en estos nuevos términos mediante

$$\mathtt{X} = \mathtt{F}\Lambda' + \mathtt{E} \tag{5.10}$$

O, lo que es lo mismo,

$$\begin{pmatrix} \mathtt{X}_1[1] & \dots & \mathtt{X}_1[p] \\ \vdots & & \vdots \\ \mathtt{X}_n[1] & \dots & \mathtt{X}_n[p] \end{pmatrix} = \begin{pmatrix} \mathtt{F}_1[1] & \dots & \mathtt{F}_1[k] \\ \vdots & & \vdots \\ \mathtt{F}_n[1] & \dots & \mathtt{F}_n[k] \end{pmatrix} \cdot \begin{pmatrix} \lambda_1[1] & \dots & \lambda_1[p] \\ \vdots & & \vdots \\ \lambda_k[1] & \dots & \lambda_k[p] \end{pmatrix} + \mathtt{E} \tag{5.11}$$

- *Ejercicio* 81. Probar que

$$n^{-1}\mathtt{F}'\mathtt{F} = \mathtt{Id}, \qquad \mathtt{F}'\mathtt{E} = 0 \tag{5.12}$$

[3]Lo cual facilita la representación de las variables a partir de sus coordenadas respecto a este sistema.

5.3.1. Representación de observaciones

Si reproducimos la igualdad (5.10) desde el punto de vista de las filas (observaciones) tenemos que, para cada $i = 1, \ldots, n$,

$$\boxed{\overrightarrow{\mathtt{X}}_i = \sum_{j=1}^{k} \mathtt{F}_i[j] \cdot \overrightarrow{\lambda}_j + \overrightarrow{\mathtt{E}}_i} \tag{5.13}$$

Además, se sigue del teorema 5.2.1 que

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \| \overrightarrow{\mathtt{E}}_i \|^2_{\mathbb{R}^p} &= \mathtt{d}^2_{n,p}(\mathtt{X}, \mathtt{F}\Lambda') && (5.14)\\ &= \mathtt{d}^2_{n,p}(\mathtt{X}, \mathcal{U}G_1') && (5.15)\\ &= \sum_{j=k+1}^{p} d_j && (5.16) \end{aligned}$$

En consecuencia, al reemplazar las observaciones originales por los puntos expresados mediante las puntuaciones factoriales cometemos unos error global cuya cuya media cuadrática es $\sum_{j=k+1}^{p} d_j$. Si este término es pequeño podemos considerar que una representación de las observaciones en dimensión k, como lo que tenemos en la figura 5.5, se aproxima a la realidad.

5.3.2. Representación de variables

Si reproducimos la igualdad (5.10) desde el punto de vista de las columnas (variables) tenemos que, para cada $l = 1, \ldots, p$,

$$\boxed{\overrightarrow{\mathtt{X}}[l] = \sum_{j=1}^{k} \lambda_j[l] \cdot \overrightarrow{\mathtt{F}}[j] + \overrightarrow{\mathtt{E}}[l]} \tag{5.17}$$

Por otra parte, dado que estamos trabajando con los datos tipificados, se verifica que $\mathsf{R} = n^{-1}\mathtt{X}'\mathtt{X}$. Se deduce nuevamente de (5.10) junto con (5.12) que

$$\mathsf{R} = \Lambda\Lambda' + n^{-1}\mathtt{E}'\mathtt{E} \tag{5.18}$$

En particular, el coeficiente de correlación entre las variables l-ésima y s-ésima es

$$r_{ls} = \langle \overrightarrow{\lambda}[l], \overrightarrow{\lambda}[s] \rangle + n^{-1} \langle \overrightarrow{\mathtt{E}}[l], \overrightarrow{\mathtt{E}}[s] \rangle \tag{5.19}$$

Luego, podemos reproducir la correlación entre las variables originales multiplicando sus componentes en dimensión k, salvo un error determinado por el segundo sumando que acotaremos a continuación: en primer lugar, se deduce de la desigualdad de Cauchy-Schwarz que

$$n^{-1} \langle \overrightarrow{\mathtt{E}}[l], \overrightarrow{\mathtt{E}}[s] \rangle \leq \sqrt{n^{-1} \| \overrightarrow{\mathtt{E}}[l] \|^2_{\mathbb{R}^n} \cdot n^{-1} \| \overrightarrow{\mathtt{E}}[s] \|^2_{\mathbb{R}^n}} \tag{5.20}$$

Esta desigualdad puede expresarse también en función de las componentes de la variable teniendo en cuenta la igualdad (5.18) aplicada a los elementos de la diagonal. Efectivamente, para todo $l = 1, \ldots, p$,

$$n^{-1} \| \overrightarrow{\mathtt{E}}[l] \|^2_{\mathbb{R}^n} = 1 - \| \overrightarrow{\lambda}[l] \|^2_{\mathbb{R}^k} \tag{5.21}$$

Si se denota el término $\| \overrightarrow{\lambda}[l] \|^2_{\mathbb{R}^k}$ por h_l^2 (que estará comprendido en todo caso entre 0 y 1), se deduce entonces de (5.19), (5.20) y (5.21) que

$$\left| r_{ls} - \langle \overrightarrow{\lambda}[l], \overrightarrow{\lambda}[s] \rangle \right| \leq \sqrt{(1-h_l^2)(1-h_s^2)} \tag{5.22}$$

En lo sucesivo, los términos $h_1^2, \ldots, h_p^2$ se denominarán comunalidades de las respectivas variables. La desigualdad (5.22) se interpreta pues como sigue: si tanto la comunalidad de la variables l-ésima como de la de la s-ésima son próximas a 1, podremos reproducir la correlación entre ambas variables, salvo un pequeño error, mediante el producto en $\mathbb{R}^k$ de sus vectores de componentes, como los que se representan, por ejemplo, en la figura 5.5. En tal caso, si dichos vectores están aproximadamente en la misma dirección cabrá pensar en una fuerte correlación lineal, que será positiva si están en el mismo sentido y negativa en caso contrario; por contra, si están direcciones aproximadamente perpendiculares, cabrá pensar en una escasa correlación lineal.

- *Ejercicio* 82. Probar que, si se denota $\overline{h^2} = \frac{1}{p}\sum_{j=1}^p h_j^2$ (es decir, la media aritmética de las comunalidades), se verifica

$$\overline{d}(k) = \overline{h^2} \tag{5.23}$$

Por lo tanto, la media aritmética de las comunalidades equivale a la proporción de varianza total explicada por las k primeras componentes principales. Luego, si $\sum_{j=k+1}^p d_j$ es próximo a 0 tendremos un buen comportamiento global de las comunalidades y, por lo tanto, el análisis de los vectores componentes en $\mathbb{R}^k$ nos dará una idea bastante aproximada de la correlación real entre las variables estudiadas.

- *Ejercicio* 83. Probar que $\lambda_j[l]$ es el coeficiente de correlación lineal entre $\overrightarrow{\mathtt{X}}[l]$ y $\overrightarrow{\mathtt{U}}[j]$. Por lo tanto, la matriz Λ recoge las correlaciones entre las variables originales y las componentes principales.

Podemos apreciar en las ecuaciones (5.13) y (5.17) que los papeles que desempeñan las matrices $\mathtt{F}$ y Λ se permutan tal y como se indica en el cuadro 5.1 según representemos las observaciones o las variables.

	k ejes	k coordenadas
n observaciones	Λ	$\mathtt{F}$
p variables	$\mathtt{F}$	Λ

Cuadro 5.1: Dualidad observaciones-variables

También hemos de advertir que la representación de las variables mediante sus componentes constituye tan sólo una simplificación de la información que nos aporta ya de por sí la matriz de correlaciones R. Concretamente, consideramos en (5.18) la aproximación $\mathsf{R} \simeq \Lambda\Lambda'$, donde la diferencia entre ambas se denota por Ψ, denominándose varianzas especificas los elementos de su diagonal, ψ_{jj}, $j = 1, \ldots, p$. De esta forma, la igualdad (5.21) puede expresarse así

$$1 = h_j^2 + \psi_{jj}, \quad 1 \leq j \leq p \tag{5.24}$$

5.3.3. Representación conjunta de observaciones y variables

Dado que tanto las observaciones como las variables pueden identificarse de manera aproximada con vectores de $\mathbb{R}^k$, según (5.13) y (5.17), respectivamente, podemos representarlas

conjuntamente mediante un gráfico k-dimensional que debemos interpretar según la siguiente igualdad, que se deduce indistintamente de cualquiera de las dos igualdades mencionadas. Concretamente, se verifica para cada $i = 1, \ldots, n$ y para cada $l = 1, \ldots, p$:

$$\boxed{\mathtt{X}_i[l] = \langle \overrightarrow{\mathtt{F}}_i, \overrightarrow{\lambda}[l] \rangle + \mathtt{E}_i[l]} \tag{5.25}$$

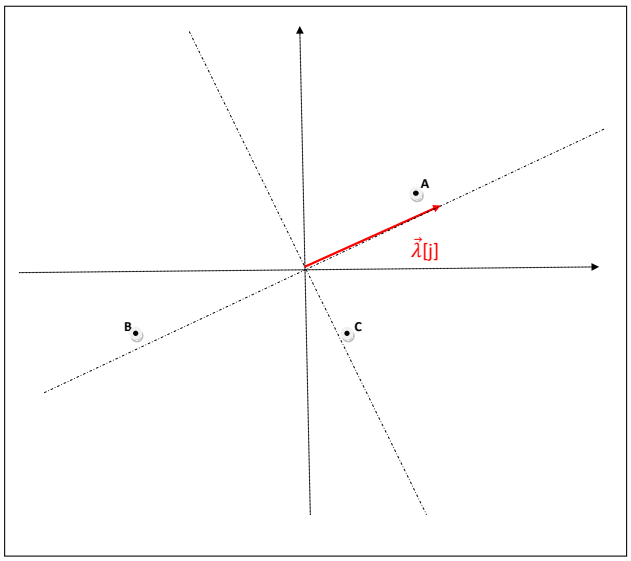


Figura 5.4: Relación observaciones-variables

En la figura 5.4 se representan conjuntamente en dimensión $k = 2$ una variable aleatoria $\overrightarrow{\mathtt{X}}[j]$ medida en n individuos, identificada con el vector $\overrightarrow{\lambda}[j]$, y tres observaciones multidimensionales $\overrightarrow{A}$, $\overrightarrow{B}$ y $\overrightarrow{C}$. En este caso, la observación $\overrightarrow{A}$ se caracteriza por un valor de la variable j-ésima por encima de la media, la observación $\overrightarrow{B}$, por un valor por debajo de la media y $\overrightarrow{C}$, por un valor en torno a la media.

Por ejemplo, en un estudio (Bravo et al. (2011)) realizado en el CENSYRA de Badajoz sobre $p = 8$ variables que caracterizan la motilidad de los espermatozoides en carneros, a partir de una muestra de $n = 383$ observaciones (ver tabla 1.1), se obtuvieron los resultados que se muestran en los cuadros 5.2 y 5.3. Del cuadro 5.2 se deduce que podemos explicar un $\overline{h^2} = 82\,\%$ de la varianza total proyectando sobre los dos primeros ejes principales, es decir, calculando $\overrightarrow{\mathtt{F}}[1]$ y $\overrightarrow{\mathtt{F}}[2]$. Según el cuadro 5.3, hay variables como `vcl` o `vap` que quedan explicadas casi perfectamente de esta forma, mientras que `bcf` y `wob` quedan deficientemente representadas. En la parte derecha se recoge la matriz de componentes que se representará en la figura 5.5 junto con las puntuaciones factoriales, y que permite simplificar la matriz de correlaciones R. Se trata de un tipo de gráfico que suele denominarse biplot que se ha obtenido en este caso a partir del comando princomp de R. En dicha figura se aprecia claramente cómo las variables `vcl`, `vsl` y `vap` correlacionan fuerte y positivamente entre sí, y correlacionan débilmente con el resto; por otra parte, las variables `wob`, `lin` y `str` correlacionan fuerte y positivamente entre sí y negativamente con `alh` y `bcf`, si bien `bcf` y `wob` no quedan satisfactoriamente representada por este gráfico, como hemos dicho anteriormente. Además, podemos apreciar qué espermatozoides presentan

valores altos, medios o bajos para las distintas variables: por ejemplo, los espermatozides 164 y 115 se caracterizan por sus altos valores en `lin` y `str`, lo cual se corresponde con bajos valores para `alh`; sin embargo, sus valores en `vcl`, `vap` y `vsl` pueden considerarse medios. Se ha superpuesto en el gráfico la circunferencia unidad para que podamos apreciar qué variables presentan una comunalidad próxima a 1.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,834	47,929	47,929	3,834	47,929	47,929
2	2,743	34,283	82,213	2,743	34,283	82,213
3	,860	10,747	92,960			
4	,366	4,578	97,538			
5	,161	2,014	99,552			
6	,033	,410	99,962			
7	,002	,030	99,992			
8	,001	,008	100,000			

Método de extracción: Análisis de Componentes principales.

Cuadro 5.2: Autovalores de R

	Inicial	Extracción
vcl	1,000	,991
vsl	1,000	,971
vap	1,000	,993
LIN%	1,000	,943
STR%	1,000	,654
WOB%	1,000	,704
alh	1,000	,881
bcf	1,000	,440

	Componente	
	1	2
vcl	-,010	,995
vsl	,632	,756
vap	,455	,887
LIN%	,956	-,169
STR%	,678	-,441
WOB%	,839	,029
alh	-,862	,372
bcf	-,639	,180

Cuadro 5.3: Comunalidades y matriz de componentes

5.3.4. Rotación de ejes

Dado que nuestra interpretación de un gráfico como el de la figura 5.5 viene dada exclusivamente en términos del producto escalar en $\mathbb{R}^k$, permanecerá invariante ante cualquier rotación que apliquemos solidariamente a las puntuaciones factoriales y los vectores de componentes de las variables. En particular, podemos intentar aplicar una rotación que aproxime lo mejor posible los diferentes vectores $\overrightarrow{\lambda}[j]$ a cualquiera de los k ejes de coordenadas. Para conseguir tal propósito existen diversos métodos iterativos como varimax o equamax, que se describen con mayor detalle en Rencher (1995). Si finalmente se consigue el objetivo, el análisis de las variables será más sencillo, pues aquéllas cuyos vectores estén en la dirección de un mismo eje de coordenadas presentarán una fuerte correlación lineal entre ellas, y guardarán poca correlación con las que estén sobre un eje diferente. El gráfico 5.5 ya lleva incorporada una rotación, lo cual se aprecia en el hecho de que las variables ya se agrupan aproximadamente en las direcciones de los ejes OX y OY.

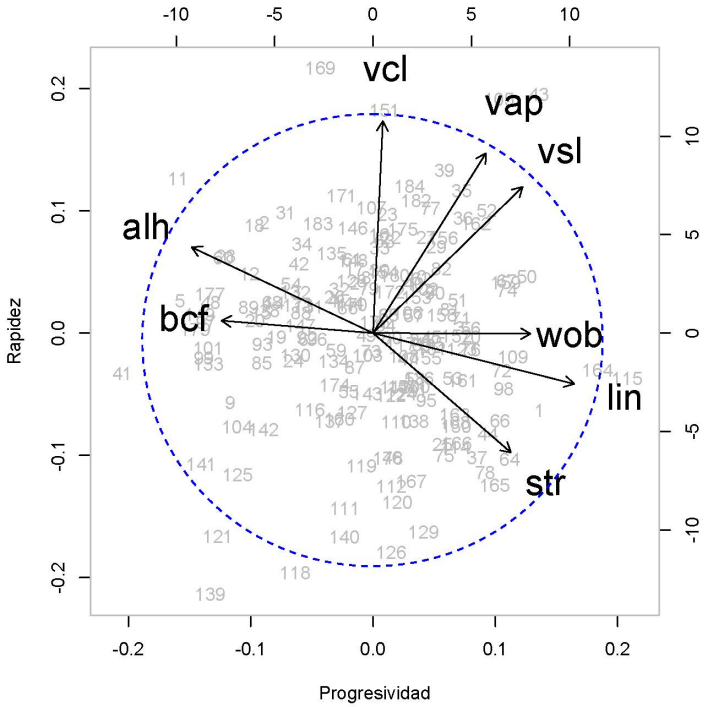


Figura 5.5: Biplot con las puntuaciones factoriales y matriz de componentes

Esto resulta especialmente útil cuando $k > 2$, porque en ese caso no podemos representar satisfactoriamente las variables. Además, el análisis numérico de las correlaciones a través de la matriz de componentes puede resultar en principio complicado. No obstante, tras una rotación adecuada, el análisis de dicha matriz resulta más sencillo porque deja más claro con qué eje se identifica cada variable (aquél en el que se obtenga una coordenada próxima a ± 1).

En el cuadro de diálogos 5.6 se indica a grandes rasgos cómo ejecutar un analisis de componentes principales (PCA) con SPSS. La ejecución con R se realiza a través del comando `princomp`.

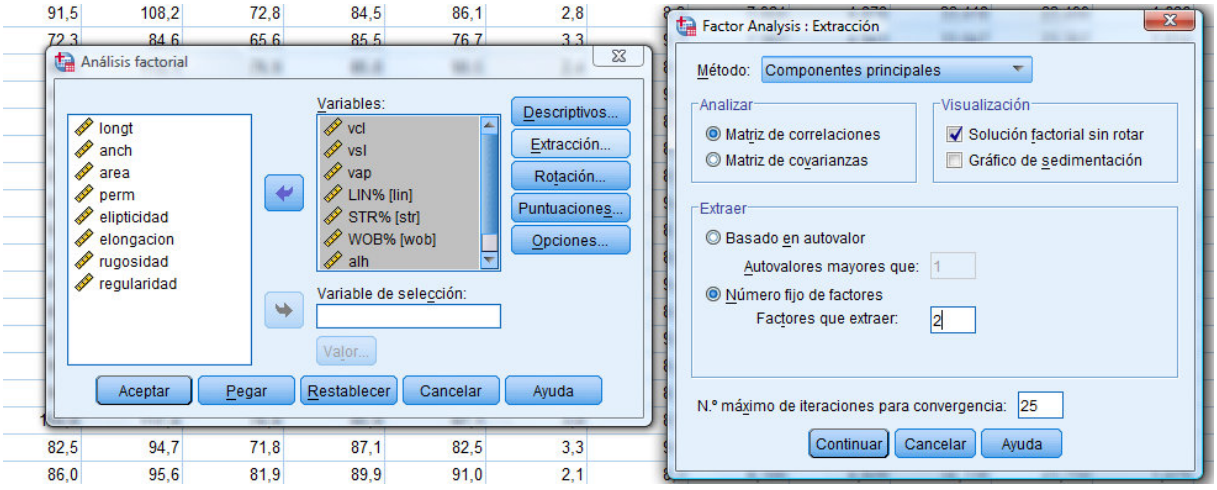


Figura 5.6: Cuadro de diálogos PCA

5.4. Análisis factorial

De la figura 5.5 se desprende la existencia de dos conglomerados de variables si utilizamos como criterio de afinidad la correlación lineal: por un lado tenemos `vcl`, `vsp` y `vap`, que se agrupan en torno al eje OY tras una rotación varimax y, por otro, el resto, que se agrupan en torno al eje OX (aunque en el segundo grupo distinguimos `lin`, `str` y `wob` que correlacionan positivamente de `alh` y `bcf`, que correlacionan negativamente con las anteriores). Desde un punto de vista formal podríamos definir factor como una clase de equivalencia en el conjunto de variables si consideramos una relación basada en la correlación lineal. Desde un punto de vista práctico, es tarea del investigador experimental dar sentido a dichos factores. En el ejemplo que nos ocupa el primer factor (eje OY) se identificaría aproximadamente con el concepto biológico de `rapidez espermática`, mientras que el segundo (eje OX) se identificaría de manera aproximada con el de `progresividad`. Queremos decir a grandes rasgos que esos son en esencia los dos factores a tener en cuenta en un espermatozoide de este tipo. Ambos factores han sido ya incorporados al propio biplot manipulando la leyenda de los ejes de coordenadas.

Nada nos garantiza en principio una agrupación clara de nuestras variables en torno a un número reducido de factores, aunque existen parámetros que pueden darnos algunas pistas antes de proceder con el análisis, como el coeficiente de Kaiser-Meyer-Olkin, definido en (5.26)

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2} \tag{5.26}$$

Se basa en la idea de que, de darse una buena agrupación de muchas variables en pocos factores, deberían observarse fuertes correlaciones simples pero bajas correlaciones parciales. Concretamente, la suma de coeficientes de correlación simple al cuadrado entre cada posible par de variables diferentes debe ser grande en relación con la suma de los coeficientes de correlación parcial al cuadrado entre cada par, conocidas el resto, que se denota por $\sum_{i \neq j} a_{ij}^2$ en (5.26). En la práctica, valores de KMO inferiores a 0.6 suelen considerase un mal indicio de cara a una clara simplificación del problema de correlación.

▪ *Ejercicio* 84. Razonar por qué un valor de KMO próximo a 1 se asocia a una reducción profunda en el análisis factorial.

En otros contextos se utilizan también métodos como oblimin[4] que aplican una transformación lineal, no necesariamente ortogonal, alos diferentes vectores $\overrightarrow{\lambda}[j]$, de manera que las puntuaciones factoriales transformadas dejan de ser necesariamente incorreladas. Este tipo de transformación permite una agrupación más clara de las variables en factores aunque éstos no sean incorrelados y pierdan, por lo tanto, su sentido original.

5.4.1. Modelos con factores latentes

Se entiende en general el análisis factorial (AF) como la búsqueda de los factores subyacentes en las variables consideradas. Desde ese punto de vista, el PCA podría entenderse como una de las posibles técnicas de extracción de factores, aunque esta visión es fuente de viejas polémicas. Muchos especialistas prefieren distinguir claramente entre el PCA y el AF, distinguiendo a su vez entre un AF exploratorio, orientado a la búsqueda de dichos factores, y otro confirmatorio, que pretende validar el modelo obtenido.

[4] Rencher (1995).

Los otros métodos más conocidos del análisis factorial (o los métodos del análisis factorial propiamente dicho, según se vea) nos los estudiaremos aquí, aunque adelantamos que son el de máxima verosimilitud y el del eje principal. A diferencia de lo explicado hasta el momento en este capítulo, estamos entendiendo los factores como variables reales pero latentes (no observadas) a partir de las cuales pueden obtenerse las variables observadas mediante una ecuación lineal, salvo errores incorrelados. Estos métodos parten de restrictivas suposiciones formales que, además, no pueden cuestionarse, pues los factores no son observables[5] Todos los métodos tienen como objetivo común una descomposición de la matriz de correlaciones tipo (5.18), es decir

$$\mathsf{R} = \Lambda\Lambda' + \Psi \tag{5.27}$$

lo cual puede conseguirse mediante un PCA o bien mediante otras técnicas, algunas de ellas, como la del eje principal, estrechamente relacionadas al mismo. Las puntuaciones factoriales pueden obtenerse mediante una regresión lineal donde los coeficientes se estiman como en (2.47) pero sustituyendo S_{YF} por la estimación de Λ. En este contexto tienen sentido técnicas como la rotación oblicua introducido anteriormente. Para más detalle consultar Rencher (1995) o Uriel y Aldás (2005).

Escalamiento multidimensional Se trata de otra forma de abordar el mismo problema de representación tratado hasta ahora en el capítulo, pero intentando explicar en una baja dimensión, en lugar de la matriz de observaciones $\mathtt{X}$, una matriz de distancias o, más generalmente, *disimilaridades* entre datos o variables; también podemos estar interesados en explicar sus *similaridades*. En ambos casos, el método parte de una matriz de distancias o bien similaridades que podría ser calculada a partir de las propias observaciones, lo cual puede acabar siendo equivalente a un análisis factorial. Su interés estriba posiblemente en que ofrece un método para representar variables de muy distinta naturaleza (de escala, ordinal, cualitativa), puesto que lo que importa en el análisis son las distancias o similaridades que pueden definirse con una importante componente subjetiva. Para más detalles consultar Hair et al. (1999) y Hastie et al. (2008).

5.5. Indicaciones sobre Análisis de Correspondencias

Mientras el objetivo del PCA es representar en un espacio de baja dimensión una matriz $\mathtt{X}$ correspondiente a la medición de varias variables numéricas, el del análisis de correspondencias es, así mismo, representar en un espacio sencillo una matriz $\mathtt{O}$ asociadas a la medición de varias variables categóricas.

Cuando contamos únicamente con dos variables cualitativas con r y s categorías, respectivamente, la denominada tabla de contingencia o tabla de frecuencias cruzadas $\mathtt{O}$ puede identificarse con $r-1$ vectores filas en $\mathbb{R}^{s-1}$ o, equivalentemente, con $s-1$ vectores columnas en $\mathbb{R}^{r-1}$. Si r y s son grandes, puede ser interesante representar las filas o las columnas (o ambas) en un espacio $\mathbb{R}^k$ de dimensión reducida, de manera que se conserve una buena parte de, en lugar de la varianza total, de la distancia χ^2. Nótese que la varianza total expresa la distancia en el sentido (1.5) entre la matriz de datos aleatorios $\mathtt{X}$ y su proyección sobre el subespacio $\langle 1_n \rangle$, que identificamos con el determinismo. Es decir, $\overline{\mathtt{X}}$ es la matriz de columnas constantes

[5] Lo cual induce, como el lector intuirá, a numerosas polémicas al respecto y a diferentes ópticas a la hora de abordar este tipo de problemas.

más próxima a `X` según las distancia (1.5). De similar forma el valor χ^2 puede interpretarse en términos de distancia entre la matriz valores observados `O` y la de valores esperados `E`.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{5.28}$$

Nótese que se entiende (5.28) como una distancia entre la tabla de contingencia observada `O` y su proyección `E` sobre el subespacio que identificamos con la independencia entre las variables cualitativas. Para ser precisos, la matriz de valores esperados `E` es, de entre aquellas cuyas proporciones condicionadas equivalen a las proporciones marginales, la más próxima a la matriz de valores observados según cierta distancia, que en tal caso vale χ^2. Los detalles pueden estudiarse en Greenacre (1984) y también se desarrollan en el Manual de Análisis Multivariante[6], se basa pues en una generalización del teorema 5.2.1, y tiene como producto final un gráfico denominado biplot, como el de la figura 5.7, que se interpreta de manera parecida al caso numérico.

Cuando tenemos más de dos variables categóricas podemos optar por varios métodos: el primero consiste en agrupar las diferentes variables categóricas en dos (con gran cantidad de categorías) con la intención de aplicar un análisis de correspondencias simple. Por ejemplo, en un estudio[7] de relación entre la especie germinada (se distinguen 11 categorías de leguminosas) y tres variables que caracterizan las condiciones del terreno, con 4, 3 y 4 categorías anidadas, respectivamente, se optó por agrupar las tres variables del terreno en una única variable cualitativa denominada grupo que distingue entre 28 categorías diferentes. El biplot de la izquierda en la figura 5.7, que recoge el 63 % de la distancia χ^2 [8], ilustra las asociaciones entre las especies y las diferentes condiciones del terreno. El cuadro de diálogos 5.8 ilustra cómo se ejecuta esta técnica con SPSS.

El segundo método consiste en una generalización de la técnica anterior denominada análisis de correspondencias múltiples. Esta técnica, que también se estudia con detalle en Greenacre (1984), se basa en la aplicación del análisis simple a una supermatriz denominada de índices. Su mayor inconveniente radica en la gran pérdida en la explicación de la distancia χ^2 (o de la inercia) que suele conllevar en estos casos la representación gráfica en baja dimensión. Téngase en cuenta que en estas circunstancias podemos llegar a barajar dimensiones por encima de 1000, por lo que resultaría poco verosímil explicar más de un 25 % de la inercia en un simple plano. No obstante, a pesar de su tendencia al simplismo, ofrece una visión global del problema de correlación que puede resultar muy útil como referencia inicial. Por ejemplo, en la parte derecha de la figura 5.7 se esquematiza en un biplot la relación conjunta entre 11 variables cualitativas, algunas de ellas con muchas categorías, relacionadas con un estudio sobre duración da bajas laborales[9]

5.6. Multicolinealidad y PCA

La técnica PCA se utiliza en ocasiones para resolver el problema de multicolinealidad en una regresión lineal múltiple. Esta afirmación hay que matizarla porque la multicolinealidad

[6] Montanero (2008)
[7] Pérez-Fernández et al. (2006).
[8] Se suele trabajar realmente con la *inercia*, que es χ^2/n.
[9] González-Ramírez et al. (2017).

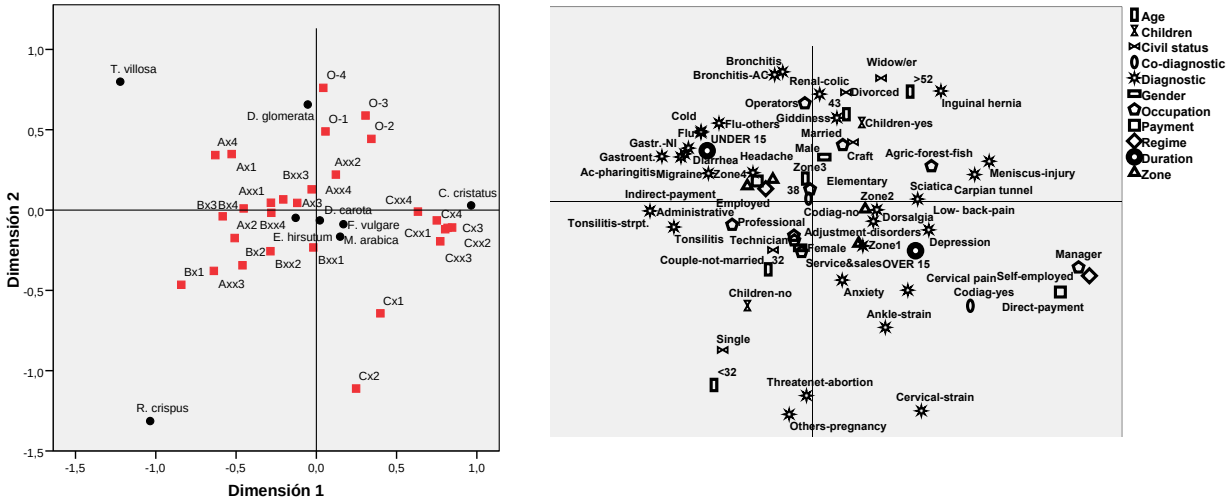


Figura 5.7: Biplots de Análisis de Correspondencias Simple (izda.) y Múltiple (dcha.)

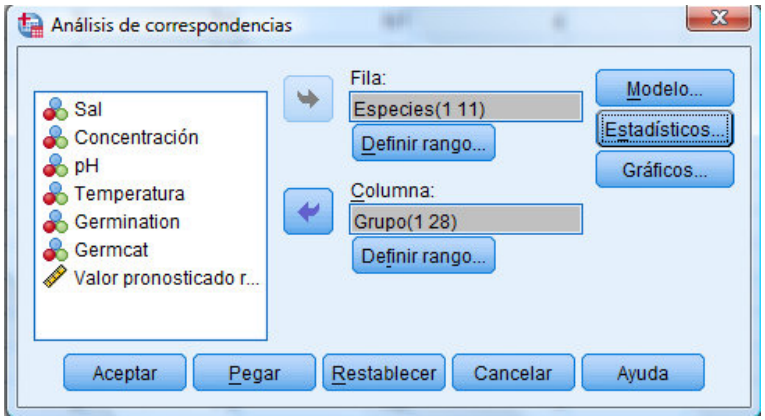


Figura 5.8: Cuadro de diálogos Análisis de Correspondencias Simple

no tiene por qué constituir un problema ni el PCA llega jamás a resolverlo. En todo caso, nos situamos en el contexto del ejemplo 4 en la página 33, suponiendo que las variables explicativas $\mathtt{Z}[1]\ldots,\mathtt{Z}[q]$ estén tipificadas. En tal caso, se dice que hay multicolinealidad cuando existe un alto grado de correlación lineal entre las variables explicativas, lo cual no supondrá perjuicio alguno a la hora de efectuar predicciones. La multicolinealidad da lugar a un incremento en la la varianza de los estimadores de la ecuación. Concretamente, se verifica para todo $j = 1,\ldots,q$:

$$\mathtt{var}\left[\hat{\beta}_j\right] = \sigma^2 \cdot \frac{1}{\mathtt{n}} \cdot \frac{1}{s^2_{\mathtt{Z}[j]}} \cdot \frac{1}{1-R^2_{j*}} \tag{5.29}$$

donde R_{j*} denota el coeficiente de correlación múltiple de $\mathtt{Z}[j]$ respecto al resto de variables explicativas. Como podemos observar, dado que las variables explicativas están tipificadas, la varianza del estimador depende de la varianza del modelo σ^2, del tamaño de la muestra n y del grado de correlación respecto al resto de variables explicativas según el FIV, es decir, $(1-R^2_{j*})^{-1}$. En virtud de (2.37), queda claro que una fuerte dependencia lineal de $\mathtt{Z}[j]$ respecto al resto de variables explicativas dificulta la significación en el contrate parcial $H_0 : \beta_j = 0$.

Hablando en términos más intuitivos, las redundancias lineales entre las variables explicativas se traducen en una menor confianza en los signos de sus coeficientes estimados, como ya se comentó en el capítulo 2.

En la figura 5.9 se ilustran los efectos de la multicolinealidad (derecha) sobre la varianza de los estimadores de regresión a partir de variables tipificadas, en contraposición con una situación en las que las variables explicativas son incorreladas (izquierda). En ambos casos se calculan los coeficientes de regresión para dos muestras diferentes, $\mathtt{Y}^1$ e $\mathtt{Y}^2$.

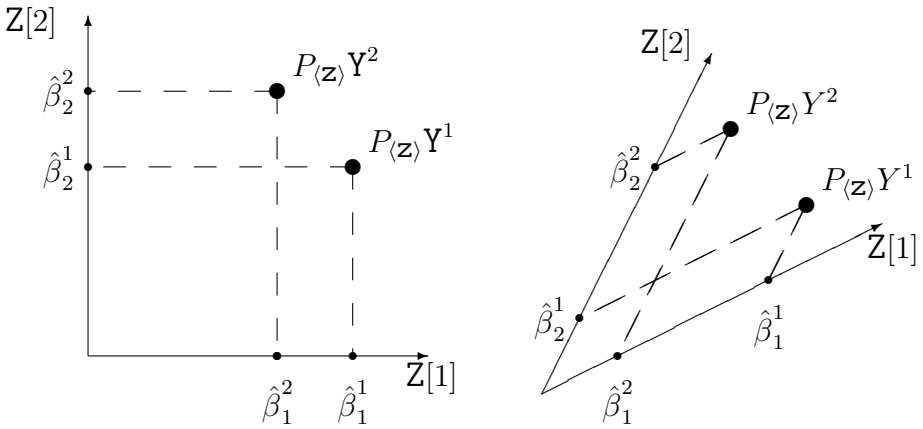


Figura 5.9: Interpretación geométrica en $\mathbb{R}^n$ de la multicolinealidad

▪ *Ejercicio* 85. Probar (5.29).

▪ *Ejercicio* 86. Simular una muestra de tamaño $n = 50$ de una par de variables $Z[1]$ y $Z[2]$, ambas con media 0 y desviación típica 1, y con coeficiente de correlación $\rho = 0.9$. A continuación, simular una variable Y que se obtenga mediante la ecuación $Y = 2Z[1] - Z[2] + \mathcal{E}$, con $\mathcal{E} \sim N(0, \sigma^2)$ y $\sigma^2 = 1$. Proceder a estimar los coeficientes de regresión β_1 y β_2 mediante una regresión lineal. Simular de nuevo Y en las mismas condiciones y estimar de nuevo los coeficientes, comparando los resultados. Repetir el procedimiento para $n = 100$; repetirlo igualmente para $\sigma^2 = 0.5$.

Que los estimadores de los coeficientes de regresión estén sometidos a una fuerte variabilidad sólo tiene trascendencia a la hora de optimizar el modelo, ya que da lugar a resultados no significativos en los tests parciales. El hecho de que una variable permanezca o salga del modelo en función del resultado del test parcial es pues muy cuestionable por el efecto de la multicolinealidad. Para compensarlo se han ideado los diferentes algoritmos de selección de variables que permiten retener algunas variables en detrimento de otras fuertemente correlacionadas con las primeras. No obstante, la selección puede depender de pequeños detalles de la muestra que podrían llegar a ser arbitrarios desde un punto de vista estadístico. Esta posible arbitrariedad no supone un problema si nuestra intención es predecir con la mayor precisión posible la variable respuesta y mediante la menor cantidad posible de variables explicativas. Tan sólo puede tener trascendencia si nuestro objetivo es determinar la influencia real de cada una de las variables explicativas en la respuesta.

En tal caso, podemos optar por ejecutar la regresión respecto a las componentes principales de las variables explicativas porque, al ser éstas incorreladas, los tests parciales no pueden verse contaminados por la multicolinealidad.

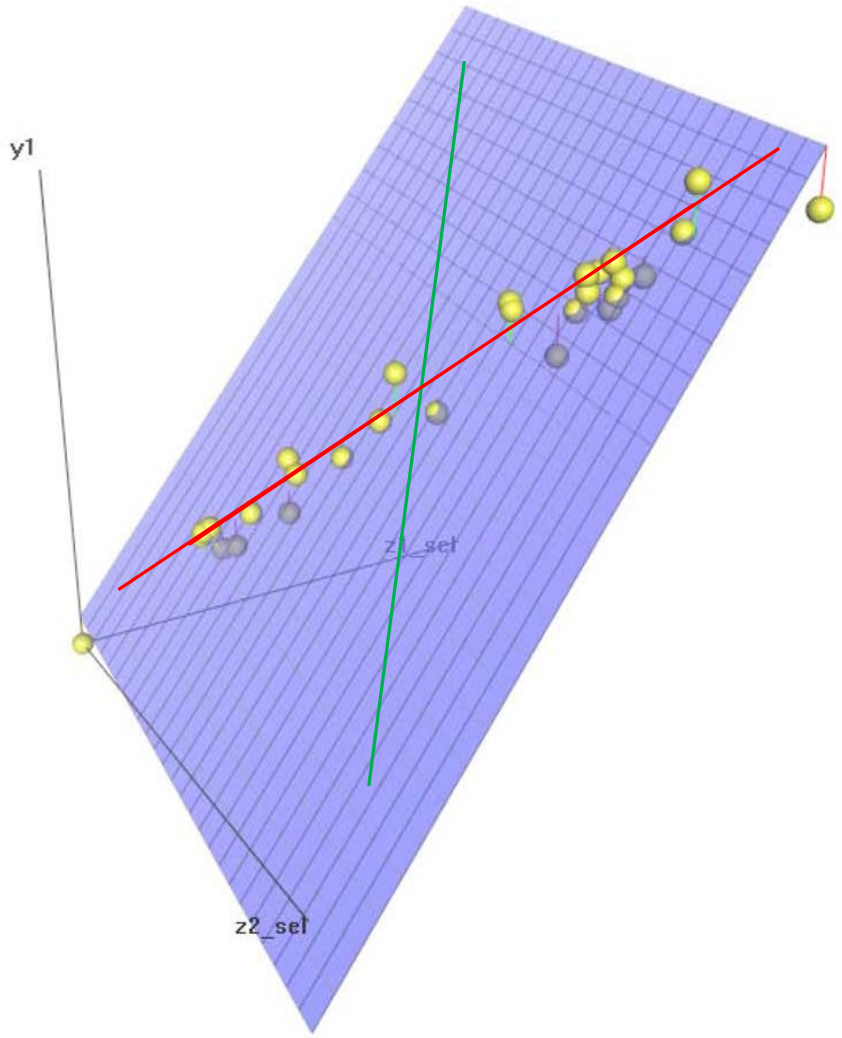


Figura 5.10: Interpretación geométrica en $\mathbb{R}^{q+1}$ de la multicolinealidad

▪ *Ejercicio* 87. Probar que el coeficiente de correlación múltiple R^2 de $\mathtt{Y}$ respecto a las variables explicativas $\mathtt{Z}$ permanece invariante si reemplazamos estas últimas por sus componentes principales $\mathtt{U}$. Probar que R^2 puede obtenerse como la suma de los coeficientes de correlación simple al cuadrado entre $\mathtt{Y}$ y cada una de las componentes principales (cosa que no sucede en general con las variables explicativas originales).

▪ *Ejercicio* 88. ¿Por qué una fuerte variabilidad de los estimadores se asocia a resultados no significativos en los tests parciales?

Una vez estimado el vector $\hat{\eta}$ con los coeficientes de regresión respecto de $\mathtt{U}$, debemos deshacer el cambio teniendo en cuenta (5.5), obteniendo así la estimación

$$\hat{\hat{\beta}} = G\hat{\eta}$$

Si hemos eliminado las últimas componentes principales en los tests parciales, esta nueva estimación de β estará sometida a tantas restricciones lineales como componentes eliminadas, y será sesgada pero con menor varianza que el EIMV $\hat{\beta}$. En las condiciones de la simulación propuesta en el ejercicio 86, el primer eje principal es $\langle (1,1)' \rangle$. Luego, si se desecha la segunda componente principal, la ecuación estimada consistirá en multiplicar $\mathtt{Z}[1]$ y $\mathtt{Z}[2]$ por un mismo coeficiente.

Desde un punto de vista práctico, distinguimos pues dos posibles circunstancias: que se eliminen componentes principales en la regresión lineal, lo cual conduce a considerar una ecuación más estable que puede entenderse como una especie compromiso entre las distintas variables correlacionadas, como en el ejemplo comentado anteriormente; o bien que no se eliminen componentes principales, lo cual debe entenderse como que la muestra consta de información suficiente para determinar qué variables poseen influencia real en la respuesta, en cuyo caso debemos acatar el resultado que aporten los tests parciales.

En la figura 5.10 representa una ecuación (plano) de regresión construida a partir de dos variables explicativas fuertemente correlacionadas. En ese caso, el plano estimado se apoya fundamentalmente, por así decirlo, en el primer eje principal, siendo muy sensible a pequeños cambios en el segundo (linea verde), ante los cuales bascularía fuertemente. Precisamente, el contraste parcial de la segunda componente principal determina si la posición respecto al segundo eje es estable (significativa), es decir, si la segunda componente principal recoge información suficiente respecto a $\mathtt{Y}$ como para otorgar fiabilidad a la ecuación obtenida. En caso contrario, se ignora y la ecuación se reduce a la recta que determina la primera componente (roja).

Una opción intermedia puede ser la denominada regresión Ridge que, en última instancia, se traduce (ver Hastie et al. (2008) sec. 3.4.1) en una ponderación del peso de las diferentes componentes principales en función de la varianza de las mismas en una estimación sesgada de la media. Concretamente, se define el estimador Ridge de $\underline{\beta}$ para $\lambda \geq 0$ mediante

$$\underline{\hat{\beta}}_{\lambda}^{\text{ridge}} = \operatorname{argmin}\{\|\mathtt{Y} - \overline{\mathtt{Y}} - \mathtt{Z}b\|^2 + \lambda \cdot \|b\|^2 : \ b \in \mathbb{R}^q\} \tag{5.30}$$

El caso $\lambda = 0$ proporciona el EIMV $\underline{\hat{\beta}}$. Puede probarse a través del cálculo de derivadas parciales que el estimador Ridge de $\underline{\beta}$ verifica, en términos de las puntuaciones factoriales $F[1], \ldots, F[q]$ de Z (tipificada),

$$\mathtt{Z}\underline{\hat{\beta}}_{\lambda}^{\text{ridge}} = \sum_{j=1}^{q} \mathtt{F}[j] \frac{d_j}{d_j + \lambda} \mathtt{F}[j]' \mathtt{Y} \tag{5.31}$$

Es decir, dado que uno de los efectos de la multicolinealidad puede ser la inflación en el tamaño del estimador $\hat{\beta}$, puede compensarse imponiendo una cota al mismo mediante una penalización (no eliminación) de las componentes principales con baja varianza.

Capítulo 6

Análisis de conglomerados

Recibe el nombre de análisis de conglomerados o análisis clúster un conjunto de técnicas destinadas a agrupar observaciones por similitud. En otras palabras, son métodos de identificación de variables categóricas latentes. Cada observación consistirá en p valores numéricos correspondientes a la medición de sendas variables y, por lo tanto, constituirán puntos de $\mathbb{R}^p$. Ésa es la razón por la que esta técnica haya sido considerada tradicionalmente como parte de la Estadística Multivariante, aunque actualmente tiende a catalogarse como Minería de Datos, de ahí que le dediquemos poco espacio. Para obtener una información más detallada se consultar Hastie et al. (2008), Mardia et al. (1979) y Hair et al. (1999).

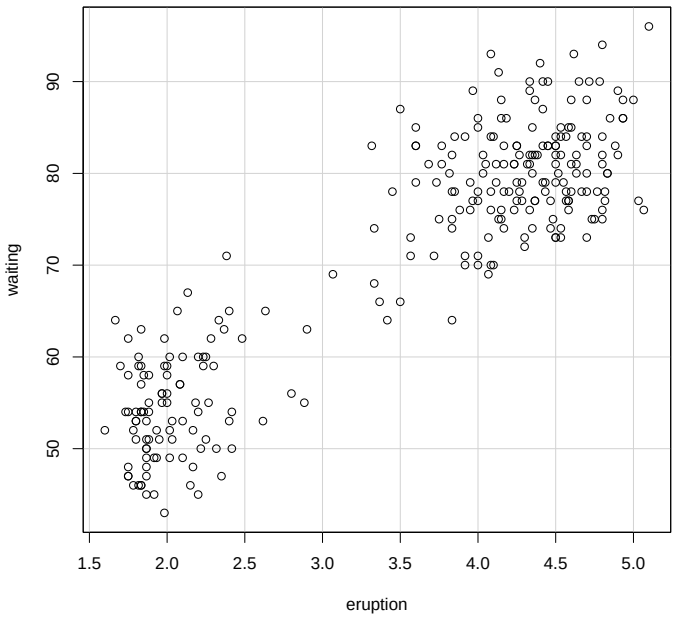


Figura 6.1: Datos geyser Old Faithful

En las dos primeras secciones abordaremos un breve estudio de los dos métodos tradicionales del análisis de conglomerados: el de k-medias y el jerárquico. En la tercera sección introduciremos escuetamente el algoritmo de agrupación EM, basado en un modelo de mezclas. Este tipo de técnica va más allá de la mera agrupación de observaciones pues tiene el ambicioso objeto de determinar de manera sencilla y precisa la distribución probabilística que las explica.

En todo caso, para hablar de similitud entre observaciones es preciso definirla previamente en el espacio $\mathbb{R}^p$. La opción más utilizada es considerar la distancia Euclídea, aunque puede optarse por medidas de similitud alternativas, algunas bastantes parecidas y otras, como la correlación entre datos que definiremos más tarde, no tanto. Si optamos por cualquiera de las dos medidas antes mencionadas debemos tener presente que no son invariantes ante cambios de escala en cualquiera de las variables consideradas, lo cual afecta de manera decisiva a la agrupación. De ahí que, en la mayoría de las situaciones, el análisis de conglomerados deba ir precedido de la tipificación de los datos. No sería el caso si optásemos, por ejemplo, por la distancia de Mahalanobis, dada la matriz de covarianzas muestral, que es sí invariante.

6.1. Método de k-medias

También conocido como quick-cluster, se utiliza para agrupar los datos en un número k de conglomerados determinado a priori. La elección de k puede basarse en argumentos formales, como los que se mencionan en la tercera sección, o bien en argumentos gráficos y, por lo tanto, intuitivos, como los que se desprenden de la figura 6.1, correspondientes a datos del geyser Olf Fatithful, de Yellowstone, donde parecen apreciarse con cierta claridad dos grandes conglomerados.

La técnica consiste en aglomerar todos los datos en torno a k puntos (que se denominan semillas) en función de la proximidad a éstos, según la distancia considerada (SPSS sólo contempla la distancia Euclídea). En ocasiones, estas semillas son establecidas de antemano en función de conocimientos previos, en cuyo caso el método es trivial. Si queremos formar k conglomerados pero no contamos con semillas, puede procederse de la siguiente forma: se seleccionan k datos, bien aleatoriamente o bien los k primeros, que serán las semillas iniciales. Los datos restantes se irán aglomerando en torno a ellos. No obstante, si la semilla más cercana a un dato dista del mismo más que que la semilla más cercana a ésta, dicho dato reemplaza como semilla a la más cercana y usurpa en lo sucesivo, por así decirlo, su conglomerado. Al final del proceso, se reconstruyen las semillas como centroides de los conglomerados finales y el procedimiento se repite sucesivamente hasta conseguir suficiente estabilidad en los centroides finales.

6.2. Método jerárquico

El método consiste en evolucionar en tantos pasos como datos tengamos desde un estado inicial, en la que cada individuo constituye un conglomerado unitario, hasta un estado final, en el que todos los individuos forman parte del mismo conglomerado global. En cada paso se unirán entre sí los dos conglomerados más similares para formar en lo sucesivo un conglomerado común indisoluble. El proceso de formación de los conglomerados queda registrado, de manera que se puede analizar el estado más interesante, que será aquél en el que queden patentes grandes diferencias entre los conglomerados y pequeñas diferencias dentro de los conglomerados. Eso querrá decir que en todos los pasos anteriores se unieron conglomerados próximos, pero en el inmediatamente posterior se unen dos conglomerados distantes, lo cual puede detectarse gráficamente mediante el dendrograma, que deberá interpretarse subjetivamente.

Así pues, el procedimiento jerárquico proporciona una método con una fuerte componente subjetiva que permite decidir el número k de conglomerados que subyacen en nuestros da-

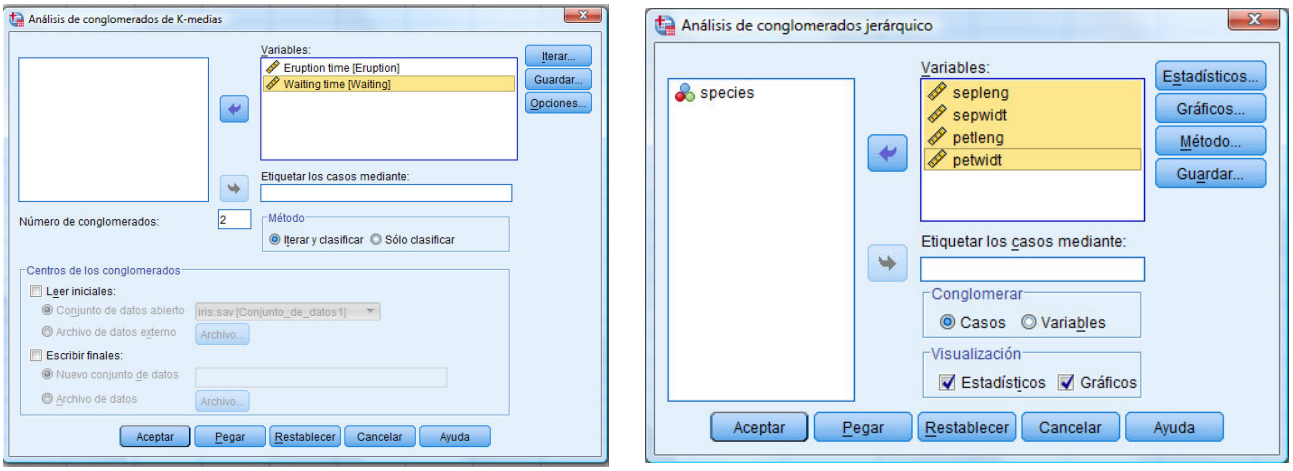


Figura 6.2: clúster k-medias para Faithful y jerarquizado para irisdata

tos. No obstante, la interpretación de dendogramas y, por lo tanto, la elección de k, sólo es aconsejable cuando se manejan muestras pequeñas o moderadas. Si la muestra es grande, el análisis jerárquico sigue siendo una herramienta igualmente válida para agrupar en un número k preestablecido de conglomerados.

Hemos dicho anteriormente que cada paso consistirá en la fusión de los dos conglomerados más similares entre sí. Obviamente, la similitud se determinará en virtud de la medida que hayamos escogido. En ese sentido y tal y como adelantábamos en la introducción, suele optarse por una de las siguientes opciones, radicalmente distintas, que son la distancia Euclídea y la correlación entre datos. Esta última se define como sigue: si $\mathtt{X}_i', \mathtt{X}_k' \in \mathcal{M}_{1\times p}$ son dos datos (filas) de la muestra con p componentes, se define la correlación entre ambos como la correlación entre sus transposiciones $\mathtt{X}_i, \mathtt{X}_k \in \mathcal{M}_{p\times 1}$. De esta forma, si los puntos obedecen a una misma tendencia entre las variables tendrán una similitud próxima a 1. Esta medida de similitud, que no es una distancia, se utiliza a veces para agrupar en conglomerados cuando se aprecian diferentes tendencias entre los datos. No obstante, este tipo de agrupación puede llevarse a cabo de una forma más sofisticada mediante el algoritmo EM que veremos a continuación.

En todo caso, debemos tener presente que cualquier medida de similitud entre datos $\mathtt{m}$ se aplica a cada par de puntos en $\mathbb{R}^p$, mientras que los conglomerados son conjuntos (unitarios o no). Por ello, queda aún pendiente determinar una medida $\tilde{\mathtt{m}}$ de similitud entre conjuntos partiendo de la medida $\mathtt{m}$ de similitud entre puntos seleccionada. En ese sentido contamos con varias opciones. El SPSS utiliza por defecto la vinculación inter-grupos, que consiste en definir la medida entre dos conglomerados A y B mediante

$$\tilde{\mathtt{m}}(A,B) = [\mathsf{card}(A\times B)]^{-1} \sum_{a\in A,\ b\in B} \mathtt{m}(a,b) \tag{6.1}$$

En la figura 6.3 presentamos el dendrogama correspondiente a 25 flores de irisdata aglomeradas en función de sus cuatro medidas morfológicas. Se han utilizados las opciones que SPSS ofrece por defecto: distancia Euclídea y vinculación intergrupos.

En un análisis de este tipo hay que tener muy presente que los datos extremos constituyen conglomerados unitarios hasta fases muy avanzadas del análisis, como es el caso de la flor 66, lo cual supone uno de los mayores inconvenientes del método jerárquico. Haciendo caso omiso de la misma, se perfilan, de manera subjetiva, entre dos y tres conglomerados de datos. Si optamos

por dos, podremos comprobar que el más pequeño está compuesto exclusivamente por flores tipo setosa, mientras que el más grande está compuesto por flores tipo vesicolor y virgínica.

De lo dicho hasta ahora se deduce que el mayor problema a la hora de formar conglomerados es determinar de una manera mínimamente objetiva el número k de clústers. No obstante, podemos optar por diferentes combinaciones entre ambas técnicas para lograr una solución aproximada: por ejemplo, podemos seleccionar a partir de la muestra original una pequeña muestra piloto y determinar k a partir del dendrograma de la segunda. También puede invertirse el orden agrupando primeramente respecto a un número elevado m de semillas, que da lugar a m centroides finales. Éstos se someten entonces a un análisis jerárquico, de manera que los grupos correspondientes a centroides próximos se unirán dando lugar a un número menor de conglomerados homogéneos.

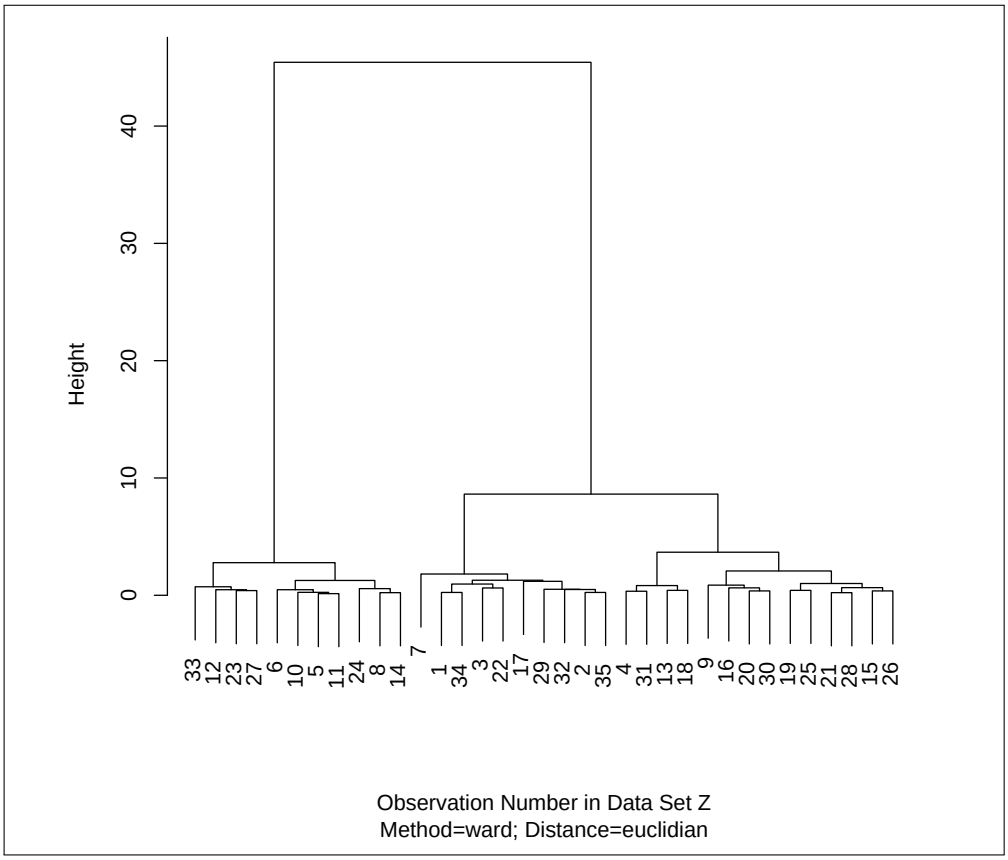


Figura 6.3: Dendrograma para irisdata

6.3. Algoritmo EM

En la sección anterior destacamos lo conflictivo que resulta determinar el número k de conglomerados a configurar a partir de la observación de la muestra. Aunque ya propusimos dos métodos (bastante subjetivos) para tal fin, existen diversos procedimientos semiautomáticos basados en principios bastante intuitivos, como el método gráfico del codo y el de Calinsky-Harabasz. El método Bayesiano que describimos a continuación está basado en un modelo de mezclas, es decir, en la aproximación de distribucionesde probabilidad p-dimensional mediante

combinaciones convexas de k distribuciones p-normales. Las condiciones de partidas son muy similares a las del modelo de regresión logística, con la salvedad de que la variable cualitativa I no tiene que ser necesariamente binaria y, además, es latente, es decir, no es observable. En definitiva, sobre un cierto espacio de probabilidad Bayesiano contamos con una variable aleatoria no observada I con valores en $\{1,\ldots,k\}$, siendo k en principio desconocido, y un vector aleatorio observado Y con valores en $\mathbb{R}^p$. De la primera suponemos que sigue una distribución multinomial con probabilidades *a priori* $\{q_1,\ldots,q_k\}$ y, de la segunda, que la distribución condicional de Y supuesto que $I=j$ es p-normal con media μ_j y matriz de covarianzas Σ_j, para $j=1,\ldots,k$. Se denotará por p_j la correspondiente función de densidad.

Obviamente, cuanto mayor sea el número de componentes k que integren la mezcla y menos restricciones lineales impongamos a las respectivas matrices de covarianzas, es decir, cuanto mayor sea el número de componentes d del parámetro θ del modelo, mayor será el logaritmo de la función de verosimilitud, $\mathcal{L}(\theta;\mathtt{Y})$. De ahí que, para evitar sobreajustes, se utilice el criterio de información bayesiano (BIC, ver Schwarz (1978)) a la hora de valorar la aptitud del modelo a seleccionar.

$$BIC = \mathcal{L}(\theta;\mathtt{Y}) - \frac{d}{2}\log n \tag{6.2}$$

El algoritmo Bayesiano denominado EM (esperanza-maximización), que se describe con mayor detalle en Hastie et al. (2008), tiene un carácter general, pues proponer una alternativa a la maximización del logaritmo de la versosimilitud en circunstancias como la mezcla de normales, en las que la distribución (marginal) de las variables observadas conducen a un logaritmo de sumas, mientras que la distribución (producto) de las variables observadas y las latentes condicionada a las primeras conduce a una suma de logaritmos, mucho más manejable desde el punto de vista del Cálculo Diferencial.

Así pues, dado un valor de k determinado, unas restricciones lineales concretas sobre las matrices de covarianzas y una estimación inicial $\hat{\theta}_0$ del parámetro, que procuraremos que sea lo más razonable posible, el algoritmo EM permite mejorarla en términos de la verosimilitud de las observaciones $\mathtt{Y}=(\mathtt{Y}_1,\ldots,\mathtt{Y}_n)'$. Se basa en lo siguiente: dado $\mathtt{Y}\in\mathcal{M}_{n\times p}$, si $\mathcal{L}$, $\tilde{\mathcal{L}}$ y $\mathcal{L}_{\mathtt{Y}}$ denotan respectivamente los logaritmos de las funciones de verosimilitud de las distribuciones marginales, producto y condicional dado $\mathtt{Y}$ (*a posteriori*), se verifica para todo θ que

$$\begin{aligned}\mathcal{L}(\theta;\mathtt{Y}) &= E_{\hat{\theta}_0}\big[\mathcal{L}(\theta;\mathtt{Y})|\mathtt{Y}\big] &(6.3)\\ &= E_{\hat{\theta}_0}\big[\tilde{\mathcal{L}}(\theta;\mathtt{Y},I_1,\ldots,I_n)|\mathtt{Y}\big] - E_{\hat{\theta}_0}\left[\mathcal{L}_{\mathtt{Y}}\left(\theta;I_1,\ldots,I_n\right)\right] &(6.4)\end{aligned}$$

El método esperanza-maximización consiste en reemplazar $\hat{\theta}_0$ por el parámetro $\hat{\theta}_1$ que maximice la esperanza condicional dado $\mathtt{Y}$ expresada en primer término en (6.4). Es decir, se trata de encontrar mediante las técnicas de Calculo Diferencial habituales los valores de (q_j,μ_j,Σ_j), con $j=1,\ldots,k$, que maximicen la siguiente expresión

$$\sum_{i=1}^{n}\sum_{j=1}^{k} P_{\hat{\theta}_0}[I_i=j|\mathtt{Y}_i]\cdot[\log q_j + \log p_j(\mu_j,\Sigma_j;\mathtt{Y}_i)] \tag{6.5}$$

Las probabilidades a posteriori $P_{\hat{\theta}_0}[I_i=j|\mathtt{Y}_i]$ se calculan mediante la regla de Bayes según se indica en (6.7). Dado que, en virtud de la desigualdad de Jensen, el término que figura restando en (6.4) alcanza su máximo en $\theta=\hat{\theta}_0$, se deduce que

$$\mathcal{L}(\hat{\theta}_1;\mathtt{Y}) \geq \mathcal{L}(\hat{\theta}_0;\mathtt{Y}) \tag{6.6}$$

Si el método se aplica de manera iterativa se puede ir mejorando la estimación hasta alcanzar un grado de estabilidad prefijado, llegando pues a una solución particular para el valor de k y las restricciones determinadas. La solución general será aquella que maximize el BIC de entre todas las soluciones particulares para diferentes valores de k y niveles de restricción respecto a la matriz de covarianzas. En definitiva, los pasos del algoritmo EM-clúster son los siguientes:

1. Se fijan k y las restricciones particulares sobre las matrices de covarianzas.

2. En la fase $s = 0$ se efectúa una estimación inicial $\hat{\theta}_0$ mediante los parámetros muestrales $\hat{q}_{j0}$, $\hat{\mu}_{j0}$ y $\hat{\Sigma}_{j0}$, $j = 1, \ldots, k$, obtenidos con un algoritmo de k-medias. La forma de estimar las matrices de covarianzas, tanto en este paso como los sucesivos, dependerá de las restricciones lineales que les impongamos.

3. En la fase s-ésima, con $s \geq 0$, se calcula mediante la regla de Bayes las probabilidades *a posteriori* para $\hat{\theta}_s$ (de esta forma se obtiene la esperanza condicional (6.5)):

$$P_{\hat{\theta}_s}(I = j|\mathtt{Y}_i) = \frac{\hat{q}_{js} \cdot p_j(\hat{\mu}_{js}, \hat{\Sigma}_{js}; \mathtt{Y}_i)}{\sum_{m=1}^{k} \hat{q}_{ms} \cdot p_m(\hat{\mu}_{ms}, \hat{\Sigma}_{ms}; \mathtt{Y}_i)} \tag{6.7}$$

4. Se calcula el parámetro $\hat{\theta}_{s+1}$ mejorado cuyas componentes maximizan (6.5) mediante[1]

$$\hat{q}_{j,s+1} = \frac{1}{n} \sum_{i=1}^{n} P_{\hat{\theta}_s}(I = j|\mathtt{Y}_i) \tag{6.8}$$

$$\hat{\mu}_{j,s+1} = \frac{1}{n\hat{q}_{j,s+1}} \sum_{i=1}^{n} P_{\hat{\theta}_s}(I = j|\mathtt{Y}_i) \cdot \mathtt{Y}_i \tag{6.9}$$

$$\hat{\Sigma}_{j,s+1} = \frac{1}{n\hat{q}_{j,s+1}} \sum_{i=1}^{n} P_{\hat{\theta}_s}(I = j|\mathtt{Y}_i) \cdot (\mathtt{Y}_i - \hat{\mu}_{j,s+1})(\mathtt{Y}_i - \hat{\mu}_{j,s+1})' \tag{6.10}$$

5. $\hat{\theta}_{s+1}$ reemplaza a $\hat{\theta}_s$ y vuelven a aplicarse los pasos 3 y 4 sucesivamente hasta alcanzar suficiente estabilidad en $\mathcal{L}(\hat{\theta}, \mathtt{Y})$.

6. Este procedimiento se lleva a cabo con diferentes valores de k y diversos grados de restricción para la matriz de covarianzas, seleccionando el modelo que maximize el BIC. En dicho modelo, cada observación $\mathtt{Y}_i$ se asigna al clúster (índice) que maximiza la probabilidad *a posteriori* (6.7) en la última iteración.

Este método puede ejecutase haciendo uso del paquete `mclust` del programa R, que considera 14 tipo de restricciones sobre las matrices de covarianzas. Si, por ejemplo, lo aplicamos a los datos de Old Faithful, el método proporciona un valor máximo del BIC para $k = 3$ componentes, con matrices de covarianzas asociadas a elipses con el mismo volumen, excentricidad y orientación (EEE). En la parte izquierda de la figura 6.4 se muestran la comparativa entre los 126 diferentes modelos considerados; en la parte derecha, se muestran los diferentes clústers con las tres distribuciones 2-normales superpuestas; por último, en la figura 6.5 se muestra la densidad estimada como mezcla de dichas distribuciones. En ocasiones como ésta y a la vista

[1] En este caso mostramos las estimaciones de las matrices de covarianzas que corresponderían al modelo sin restricciones.

del gráfico, puede procederse a agrupar clusters (verde y azul) cuya separación no resulte natural. Es decir, puede ser necesario combinar un algoritmo de agrupación de modelo de mezclas (EM-clúster) con otro de agrupación modal aplicado a las medias obtenidas[2].

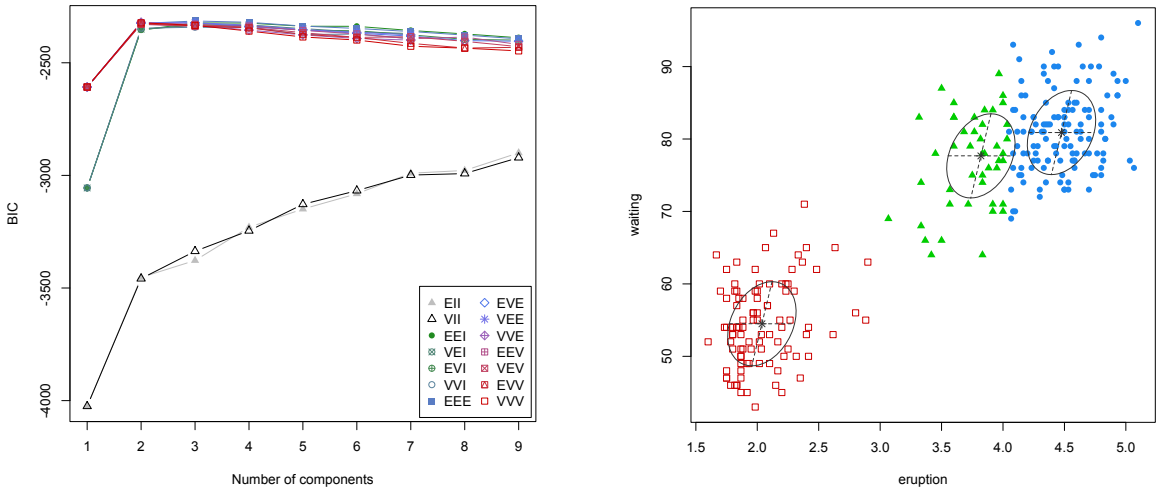


Figura 6.4: Comparativa de BIC y EM-clúster par Old Faithful

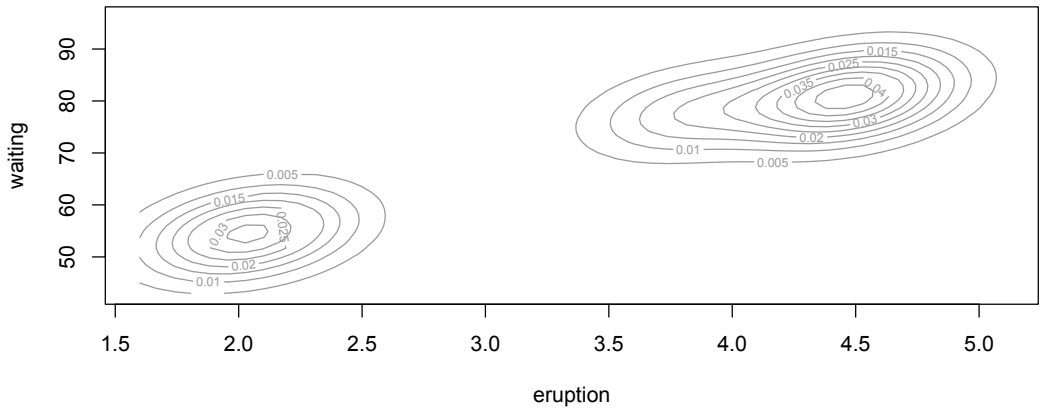


Figura 6.5: Densidad estimada para Old Faithful

6.4. Análisis de conglomerados bietápico

En este última sección apuntaremos algunas indicaciones acerca de otra técnica automática de selección de conglomerados implementada en el programa SPSS (de hecho, para entender o, mejor dicho, intuir su funcionamiento, hay que remitirse a día de hoy al propio manual del programa). Se trata de un método similar aunque más sofisticado que el propuesto al final de la sección6.2. Supone que todas las variables son independientes, siendo normales tipificadas las

[2]Ver Chacón, J.E., (2019). *Mixture model modal clustering.* Advances in Data Analysis and Classification.

numéricas y multinomiales las cualitativas (no obstante, el procedimiento resulta ser bastante robusto ante la violación de este supuesto).

En un primer paso se generan por similaridad un cierto número de conglomerados que no puede superar un máximo establecido. Los conglomerados se van formando conforme se leen los datos en el orden en el que aparecen enel archivo (lo cual puede resultar trascendental). La medida de similaridad por defecto es el logaritmo de la verosimilitud, calculada a partir de las p variables numéricas. De los conglomerados formados sólo interesará, en la siguiente fase, sus centroides (medias en dimensión p). El siguiente paso es un procedimiento de aglomeración de dichos centroides hasta llegar a un conglomerado (centroide) único. En esta fase se consideran las proporciones de las categorías correspondientes a las variables cualitativas a la hora de calcular el logaritmo de la verosimilitud y los centroides. El árbol resultante debe podarse para obtener una solución óptima en función del BIC y teniendo además en cuenta unas cotas máximas de complejidad a niveles horizontal y vertical predeterminadas.

Este método posee dos desventajas, al menos aparentes: su complejidad y los fuertes supuestos sobre los que descansa teóricamente. Por contra, tiene como ventaja principal, a parte de su automatismo, que contempla el uso de variables cualitativas.

Bibliografía

Anderson, T.W. (1958), *"An Introduction to Multivariate Statistical Analysis",* Wiley.

Arnold, S.F. (1981), *"The Theory of Linear Models and Multivariate Analysis",* Wiley.

Bilodeau, M. y Brenner, D. (1999), *"Theory of Multivariate Statistics",* Springer.

Bravo, J.A., Montanero, J., Calero, R., Roy, T.J., *"Identification of sperm moyility characteristics in ejaculates from Ile de France rams",* Animal Reproduction Science, vol 129, 22-29 (2011).

Breiman, L., Friedman, J., Stone, C.J. y Olsen, R.A. (1984) *"Classification and Regression Trees",* Taylor & Francis.

Dillon, W.R. y Goldstein, M. (1984), *"Multivariate Analysis. Methods and Aplications",* Wiley.

Dobson, A.J. (1990), *"An Introduction to Generalized Linear Models",* Chapman & Hall.

Flury, B. (1997), *"A First Course in Multivariate Statistics",* Springer.

Gifi, A. (1990), *"Nonlinear Multivariante Analysis",* Wiley.

González-Ramirez, C., Montanero-Fernández, J., Peral-Pacheco, D., *"A multifactorial study on duration of temporary disabilities in Spain",* Archives of Environmental & Occupational Health, vol 72(6), 328-335 (2017).

Greenacre, M.J. (1984), *"Theory and Applications of Correspondence Analysis",* Academic Press.

Hair, J.F., Anderson, R.E., Tatham, R.L., y Black, C.B. (1999), *"Análisis Multivariante",* Prentice Hall.

Hastie, T., Tibshirani, R. y Friedman, J. (2008), *" The Elements of Statistical Learning",* Springer.

Mardia, K.V., Kent, J.T. y Bibby, J.M. (1979), *" Multivariate Analysis",* Academic Press.

Montanero, J. (2008), *"Manual 56: Modelos Lineales",* Servicio de Publicaciones UEx. `http://hdl.handle.net/10662/2443`

Montanero, J. (2008), *"Manual 59: Análisis Multivariante",* Servicio de Publicaciones UEx. `http://hdl.handle.net/10662/2444`

Nogales, A.G. (1998), *"Estadística Matemática",* Servicio de publicaciones UEx.

Peña, D. (2010), *"Regresión y Diseño de Experimentos",* Alianza editorial.

Pérez-Fernández, M.A., Calvo Magro, E., Montanero-Fernández, J., Oyola-Velasco, J.A. *"Seed Germination in response to chemicals: Effect of nitrogen and pH in media",* Journal of Envi-

ronmental Biology, vol 27(1), 13-20 (2006).

Rencher, A.C. (1995), *"Methods of Multivariate Analysis",* Wiley.

Rodríguez-Mansilla, J., González López-Arza, M.V,. Varela-Donoso, E., Montanero-Fernández, J., González-Sánchez, B., Garrido-Ardila, E.M., *"The effects of ear acupressure massage therapy and no therapy on symptons of dementia: a randomized controlled trial",* Clinical Rehabilitation, vol 29(7) 683-693 (2015).

Silverman, B. W. (1986), *"Density Estimation for Statistics and Data Analysis",* Chapman & Hall.

Schwarz, G. (1978), *"Estimating de dimension of a model",* The Annals os Statistics, 6(2), 461-464.

Uriel, E. y Aldás, J. (2005), *"Análisis Multivariante Aplicado",* Thomson.

Índice alfabético