Índice

Capítulo 1

Preliminares

1.1. Introducción

Hemos preferido comenzar el texto con un capítulo en el que se enmarcaran con precisión los temas que se van a estudiar en el resto del libro, además de servir de formalización de algunos elementos matemáticos indispensables en su desarrollo formal.

Como ya dijimos en el Prólogo, los apartados que comiencen con un asterisco pueden evitarse si no se desea un estudio formal de las cuestiones que allí se aborden.

1.2. Nombres nuevos para conocidos métodos clásicos

Los primeros capítulos del libro corresponden a lo que suele denominarse *Análisis Multivariante* porque nuestros datos serán observaciones de p variables aleatorias en los n individuos de la muestra, en lugar de observaciones de una sola variable aleatoria como ocurría en la mayoría de los métodos de *Análisis Univariante* estudiados en CB.

Por tanto, la *matriz de datos*, en donde aparecen recogidas las observaciones, es una matriz (es decir, una ordenación por filas y columnas) de la forma

$$\begin{array}{cc} & \text{Variables} \\ \text{Individuos} & \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \end{array}$$

En este tipo de análisis, al igual que ocurría en su homólogo *Análisis Univa-*

riante, caben dos formas posibles de estudio: el *Análisis Exploratorio de Datos* (véase CB-capítulo 14), en donde no se utilizan suposiciones ajenas a los datos, tales como modelos para las variables de donde se obtuvieron y en donde se deja que éstos *hablen por sí mismos*; el propósito de este tipo de análisis es el descubrir posibles patrones de comportamiento de los datos tales como simetrías, modelos probabilísticos, posibles grupos de datos homogéneos, etc. En él juega un papel especial el uso de gráficos. Los capítulos que siguen de Componentes Principales, Análisis de Correspondencias, Escalado Multidimensional y Análisis de Conglomerados, serán básicamente de este tipo.

La otra posible vía de estudio de los datos, tanto en el caso univariante como en el multivariante, se denomina *Análisis Confirmatorio de Datos*, en el que se utiliza de forma destacada el contraste de hipótesis como herramienta estadística para la confirmación o rechazo de hipótesis sobre el modelo supuesto. En este caso, la suposición de una distribución normal multivariante para los datos es esencial. La utilización de Métodos Robustos en estas situaciones resulta muy interesante.

Pues bien, el *Análisis Exploratorio de Datos Multivariantes* recibe hoy en día el nombre de *Data Mining*, traducido en ocasiones, de forma desafortunada según mi opinión, por *Minería de Datos*, en donde el propósito será, como dijimos más arriba, explorar los datos sin suposiciones adicionales, buscando patrones de comportamiento, clasificaciones en grupos de datos, etc. Dado el gran volumen de datos con el que se suele trabajar en los tiempos actuales, otra característica de este tipo de análisis es el uso intensivo del ordenador, especialmente en la obtención de gráficos.

Una de las razones de realizar un Análisis Multivariante de datos (tanto exploratorio como confirmatorio) en lugar de p Análisis Univariantes, es el determinar relaciones entre las p variables de donde se obtuvieron los datos.

Si para descubrir estas estructuras o grupos, cuántos grupos hay, cuáles individuos pertenecen a cada grupo, etc., no utilizamos información previa referente a otros grupos similares de sujetos, se suele hablar de *Estadística no Supervisada*. Con objeto de buscar respuesta a esas preguntas pueden utilizarse *ordenaciones*, con un *Análisis de Componentes Principales*, o un *Multidimensional Scaling*, o *clasificaciones* con un *Análisis Cluster*.

Alternativamente, podemos conocer previamente los grupos en los que clasificar los datos, utilizando métodos de *Estadística Supervisada*, tales como el *Análisis Discriminante* o los *Modelos Lineales*.

No obstante, los Métodos Estadísticos que estudiaremos en el libro lo serán de forma individual, ya que éstos no están diseñados habitualmente con un único propósito. Tan solo hemos pretendido enunciar aquí algunos de los nombres que suelen utilizarse hoy en día para asignar a grupos de Métodos Estadísticos y que pueden representar, en el mejor de los casos, el objetivo común para el que van a ser utilizados.

1.3. *Algunos elementos matemáticos básicos

Como dijimos más arriba, la matriz de datos está formada por las observaciones de las p variables en estudio en los n individuos de la muestra. Estas observaciones serán, por lo general, números reales, es decir, *escalares* aunque, como alguna variable puede ser del tipo cualitativo, como por ejemplo Color de los Ojos, en ocasiones los datos recogidos para esa variable y que forman la correspondiente columna de la matriz de datos, no serán escalares sino *valores* de la forma: Azul, Verde, Castaño, Azul, etc.

No obstante, si queremos utilizar potentes Métodos Estadísticos, las columnas de la matriz de datos deberán estar formadas por números reales, de manera que podamos utilizar técnicas matemáticas estándares. En ese caso, deberemos cuantificar las variables de tipo cualitativo con valores de tipo indicador: 0, 1, etc.

Los escalares los representaremos como hasta ahora, pero a las matrices (como la matriz de datos) las representaremos con letras negritas. Así, hablaremos de la matriz **A**, o de la matriz **B**, etc.

Si **A** es la matriz

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \\ 2 & 4 \end{pmatrix}$$

en muchas ocasiones nos interesará trabajar con la matriz *traspuesta* de la anterior, que representaremos como $\mathbf{A^t}$ y que se define como la matriz en la que sus filas están formadas por las columnas de la dada; es decir, en la que hemos traspuesto las filas y columnas. Así, la matriz traspuesta de la matriz **A** es

$$\mathbf{A^t} = \begin{pmatrix} 1 & 0 & 2 \\ 2 & 3 & 4 \end{pmatrix}$$

ya que, por ejemplo, la que figuraba como primera fila, figura ahora como primera columna, la que figuraba como segunda columna es ahora la segunda fila, etc.

La *dimensión* de una matriz es el número de filas y de columnas por el que está formado (en ese orden). Así, la matriz **A** tiene dimensión 3×2 y la matriz $\mathbf{A^t}$ dimensión 2×3. Una matriz se dice *cuadrada* si ambos valores de su dimensión son iguales; es decir, una matriz 2×2 o una 3×3 son matrices cuadradas y una 2×3 no lo es. Si una matriz coincide con su traspuesta se dice que es *simétrica*.

Una matriz que aparece frecuentemente es la *matriz identidad*, **I**, formada por unos en la diagonal principal y ceros en el resto,

$$\mathbf{I} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

La inversa de una matriz cuadrada $\mathbf{A}$ se define como una matriz, a la que denominaremos $\mathbf{A}^{-1}$, tal que su producto por $\mathbf{A}$ es la matriz identidad.

Además de los escalares y las matrices, trabajaremos en este texto con *vectores*, que van a ser ordenaciones de datos (habitualmente de tipo numérico), concebidos como columnas. Al igual que con las matrices, representaremos los vectores con letras negritas (de hecho se puede pensar en un vector formado por r escalares como en una matriz $r \times 1$).

Si $\mathbf{v}$ es el vector

$$\mathbf{v} = \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix}$$

su traspuesto será el vector $\mathbf{v^t} = (3, 1, 3)$.

El producto de vectores y/o matrices tiene sentido sólo cuando el segundo valor de la dimensión del primer factor sea igual que el primer valor de la dimensión del segundo factor; el orden es relevante. Así, se puede (pre) multiplicar una matriz 3×2 por una matriz 2×2, pero no al revés.

El producto del vector $\mathbf{v^t}$ por el vector $\mathbf{w}$, ambos de longitud, digamos m, se define como

$$\mathbf{v^t w} = (v_1, v_2, ..., v_m) \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = v_1 \cdot w_1 + v_2 \cdot w_2 + \cdots v_m \cdot w_m = \sum_{i=1}^{m} v_i w_i.$$

La definición del producto de dos matrices y/o vectores $\mathbf{A}$ y $\mathbf{B}$ es (cuando se pueda definir el producto) una matriz (o un vector) tal que el elemento que ocupa el lugar (i, j) (es decir, el que ocupa la fila i-ésima y la columna j-ésima) es el resultado de multiplicar la fila i-ésima de la matriz $\mathbf{A}$ por la columna j-ésima de la matriz $\mathbf{B}$, consideradas ambas como vectores, de la misma manera que en el párrafo anterior.

La dimensión de la matriz (vector) resultante es el primer valor de la dimensión del primer factor $\times$ el segundo valor de la dimensión del segundo factor.

Así, $\mathbf{AB}$ será igual a

$$\mathbf{AB} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ -1 & 4 \end{pmatrix} = \begin{pmatrix} 1 \cdot 3 + 2 \cdot (-1) & 1 \cdot 1 + 2 \cdot 4 \\ 0 \cdot 3 + 3 \cdot (-1) & 0 \cdot 1 + 3 \cdot 4 \\ 2 \cdot 3 + 4 \cdot (-1) & 2 \cdot 1 + 4 \cdot 4 \end{pmatrix} = \begin{pmatrix} 1 & 9 \\ -3 & 12 \\ 2 & 18 \end{pmatrix}$$

y tendrá dimensión 3×2.

1.4. Algunos elementos básicos de los vectores aleatorios

Decir que observamos p variables aleatorias unidimensionales $X_1, X_2, ..., X_p$ es lo mismo que decir que observamos el *vector aleatorio* $\mathbf{X^t} = (X_1, X_2, ..., X_p)$. Y al igual que las variables aleatorias unidimensionales tenían su *media* y su *varianza*, las variables aleatorias multidimensionales, o vectores aleatorios, tienen asociados el *vector de medias*, definido como el vector de las medias de las variables que forman el vector aleatorio,

$$\mathbf{m^t} = (E[X_1], ..., E[X_p]) = (\mu_1, ..., \mu_p)$$

y la *matriz de varianzas-covarianzas* (o simplemente matriz de covarianzas), que está formada por las covarianzas entre las variables del vector aleatorio, en donde la covarianza entre las variables X_i y X_j se define (CB-sección 4.3) como

$$Cov(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = \sigma_{ij}$$

siendo la última igualdad, simplemente, una notación abreviada. Si $i = j$ aparece la varianza de la variable

$$\sigma_{ii} = E\left[(X_i - \mu_i)^2\right] = Var(X_i) = \sigma_i^2$$

Por tanto, la matriz de covarianzas será

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

en donde suele ser $n > p$.

Una vez observadas la p variables en los n individuos de la muestra, y obtenida así la matriz de datos, el estimador natural del vector de medias poblacional $\mathbf{m}$ es el *vector de medias muestrales*

$$\overline{\mathbf{x}}^{\mathbf{t}} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_p)$$

en donde $\overline{x}_i$ es la media de los datos correspondientes a la variable i-ésima; es decir, la media aritmética de los datos de la columna i-ésima de la matriz de datos,

$$\overline{x}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

La matriz de varianzas-covarianzas poblacional $\mathbf{\Sigma}$ se estima mediante la *matriz de covarianzas muestral*

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x_i}-\overline{\mathbf{x}}\right)\left(\mathbf{x_i}-\overline{\mathbf{x}}\right)^{\mathbf{t}}$$

en donde $\mathbf{x_i}$ es la i-ésima fila de la matriz de datos considerada como vector (es decir, como columna) aleatorio

$$\mathbf{x_i} = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix}$$

Por tanto, la matriz de varianzas-covarianzas muestral $\mathbf{S}$ será la matriz de dimensión $p \times p$

$$\begin{pmatrix}
\displaystyle\sum_{i=1}^{n}\frac{(X_{i1}-\overline{x}_1)^2}{n-1} & \displaystyle\sum_{i=1}^{n}\frac{(X_{i1}-\overline{x}_1)(X_{i2}-\overline{x}_2)}{n-1} & \cdots & \displaystyle\sum_{i=1}^{n}\frac{(X_{i1}-\overline{x}_1)(X_{ip}-\overline{x}_p)}{n-1} \\
\displaystyle\sum_{i=1}^{n}\frac{(X_{i2}-\overline{x}_2)(X_{i1}-\overline{x}_1)}{n-1} & \displaystyle\sum_{i=1}^{n}\frac{(X_{i2}-\overline{x}_2)^2}{n-1} & \cdots & \displaystyle\sum_{i=1}^{n}\frac{(X_{i2}-\overline{x}_2)(X_{ip}-\overline{x}_p)}{n-1} \\
\vdots & \vdots & \vdots & \vdots \\
\displaystyle\sum_{i=1}^{n}\frac{(X_{ip}-\overline{x}_p)(X_{i1}-\overline{x}_1)}{n-1} & \displaystyle\sum_{i=1}^{n}\frac{(X_{ip}-\overline{x}_p)(X_{i2}-\overline{x}_2)}{n-1} & \cdots & \displaystyle\sum_{i=1}^{n}\frac{(X_{ip}-\overline{x}_p)^2}{n-1}
\end{pmatrix}$$

1.5. La distribución normal multivariante

Una suposición que habitualmente es necesario realizar, en los capítulos en los que efectuamos Análisis Confirmatorio, es que la variable aleatoria en observación p-dimensional, $\mathbf{X} = (X_1, ..., X_p)^t$ se distribuye según una *distribución normal multivariante.*

Diremos que $\mathbf{X}$ sigue una *distribución normal multivariante* con vector de medias $\mathbf{m} = (\mu_1, ..., \mu_p)^t$ y matriz de covarianzas $\mathbf{\Sigma}$ si su función de densidad es

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{|\mathbf{\Sigma}|}} e^{-\frac{1}{2}\{(\mathbf{x}-\mathbf{m})^t \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{m})\}}.$$

Una cuestión central en el segundo volumen, es la utilización de Métodos Robustos, para los cuales no es imprescindible tal suposición.