Tema 2
Datos multivariantes

Aurea Grané
Departamento de Estadística
Universidad Carlos III de Madrid

2 Datos multivariantes

1. Matrices de datos
2. Medias, covarianzas y correlaciones
3. Variables compuestas
4. Teorema de la dimensión
5. Distancias

Introducción

El análisis multivariante es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar un número $p > 1$ de variables estadísticas sobre una muestra de n individuos.

Las variables obervables son **homogéneas** y **correlacionadas**, sin que ninguna predomine sobre las demás.

Generalmente la información multivariante es una matriz de datos. Aunque, a menudo, también puede ser una matriz de distancias (o similaridades), que miden el grado de discrepancia (o similitud) entre los individuos.

2.1 Matrices de datos

Supondremos que hemos observado p variables en un conjunto de n elementos o individuos. Cada una de estas p variables es una variable **univariante** y el conjunto de las p variables forma una **variable multivariante**.

La matriz de datos $\mathbf{X}$ es la representación de estas p variables medidas en los n individuos:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

La matriz $\mathbf{X}$ puede representarse de dos formas distintas: por filas y por columnas.

Representación por filas:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

donde $\mathbf{x}_i' \in \mathbb{R}^p$ $(i = 1, \dots, n)$ representa los valores observados para el individuo i-ésimo en las p variables.

Representación por columnas:

$$\mathbf{X} = (X_1, X_2, \dots, X_p),$$

donde $X_j \in \mathbb{R}^n$ $(j = 1, \dots, p)$ representa la variable univariante j-ésima medida sobre todos los individuos de la muestra.

2.2 Medias, covarianzas y correlaciones

Dada una matrix $\mathbf{X}$, $n \times p$, con datos cuantitativos, se define el **vector de medias** de $\mathbf{X}$ como el vector columna $p \times 1$

$$\overline{\mathbf{x}} = \left(\overline{X}_1, \overline{X}_2, \dots, \overline{X}_p\right)', \text{ donde } \overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$

Generalmente, el vector de medias se expresa como

$$\overline{\mathbf{x}} = \frac{1}{n}\,\mathbf{X}'\,\mathbf{1},$$

donde $\mathbf{1} = (1, 1, \dots, 1)'$ es un vector columna $n \times 1$.

La matriz cuyas columnas tienen media cero se denomina **matriz de datos centrados** y es el resultado de aplicar a cada columna de $\mathbf{X}$ una traslación igual a menos su media, es decir,

$$\mathbf{X}_0 = \mathbf{X} - \mathbf{1}\,\overline{\mathbf{x}}' = \mathbf{X} - \mathbf{1}\left(\frac{1}{n}\,\mathbf{1}'\,\mathbf{X}\right) = \mathbf{X} - \frac{1}{n}\mathbf{1}\,\mathbf{1}'\,\mathbf{X} = \mathbf{H}\mathbf{X},$$

donde $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\,\mathbf{1}'$ es la **matriz de centrado**, $\mathbf{I}$ es la matriz identidad de orden n.

Propiedades de la matriz de centrado

1. $\mathbf{H}' = \mathbf{H}$ (simétrica)
2. $\mathbf{H}^2 = \mathbf{H}$ (idempotente)
3. $\mathbf{H}\,\mathbf{1} = \vec{0}$ ($\mathbf{1}$ es un autovector de $\mathbf{H}$ de autovalor 0)
4. $\mathsf{rang}(\mathbf{H}) = n - 1$
5. Los autovalores de $\mathbf{H}$ son 0 y 1.

Ejercicio 2.1 *Demostrar las propiedades anteriores.*

Se define la **matriz de covarianzas** de $\mathbf{X}$ como

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\,\mathbf{H}\,\mathbf{X} = (s_{jk})_{p\times p}.$$

Observad que para cada par (j, k), $1 \le j, k \le p$ el elemento s_{jk} de $\mathbf{S}$ es la covarianza de las columnas j y k de la matriz $\mathbf{X}$, es decir,

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X}_j)(x_{ik} - \overline{X}_k).$$

En particular, para cada j, $j = 1, \dots, p$, el elemento s_{jj} de $\mathbf{S}$ es la varianza de la columna j de la matriz $\mathbf{X}$, es decir,

$$s_{jj} = s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X}_j)^2.$$

Cuando se desee obtener un estimador insesgado de las varianzas y covarianzas de la población, se utilizará la matriz

$$\tilde{\mathbf{S}} = \frac{n}{n-1}\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\,\mathbf{H}\,\mathbf{X}.$$

Se define la **matriz de correlaciones** de $\mathbf{X}$ como aquella matriz cuyos elementos son los coeficientes de correlación de las columnas de $\mathbf{X}$, es decir,

$$\mathbf{R} = (r_{jk})_{p\times p}, \text{ donde } r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}\, s_{kk}}}.$$

En notación matricial, la matriz $\mathbf{R}$ se obtiene como

$$\mathbf{R} = \mathbf{D}_s^{-1}\, \mathbf{S}\, \mathbf{D}_s^{-1},$$

donde $\mathbf{D}_s = \mathsf{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots, \sqrt{s_{pp}})'$.

Observad que $\mathbf{S}$ y $\mathbf{R}$ son matrices simétricas. Más adelante veremos que también son matrices semidefinidas positivas.

El vector $\overline{\mathbf{x}}$ es una medida de centralidad multivariante de los datos. La matriz $\mathbf{S}$ y, sobre todo, la matriz $\mathbf{R}$ son medidas matriciales de interdependencia lineal entre las variables.

Como medidas escalares de dispersión multivariante (o de variabilidad global) se definen la varianza generalizada como $\mathsf{det}(\mathbf{S})$ y la variación total como $\mathsf{tr}(\mathbf{S})$.

Como medida escalar de interdependencia lineal (o de dependencia global) se define $\eta^2 = 1 - \mathsf{det}(\mathbf{R})$, que verifica las propiedades:

1. $0 \leq \eta^2 \leq 1$,
2. $\eta^2 = 0 \Leftrightarrow$ las p variables están incorreladas,
3. $\eta^2 = 1 \Leftrightarrow$ existen relaciones lineales entre las variables.

Ejercicios computacionales

Ejercicio 2.2 *Dada una matriz de datos* $\mathbf{X}$*, escribir un programa en Matlab que calcule el vector de medias, la matriz de covarianzas y la matriz de correlaciones de* $\mathbf{X}$*.*

Ejercicio 2.3 *Dada una matriz de datos* $\mathbf{X}$*, escribir un programa en Matlab que calcule la varianza generalizada, la variación total y el coeficiente* η^2*.*

2.3 Variables compuestas

Algunos métodos del Análisis Multivariante consisten en obtener e interpretar combinaciones lineales adecuadas de las variables observables.

Se llama **variable compuesta** a toda combinación lineal de las variables observables. Por ejemplo, sea $\mathbf{a} = (a_1, a_2, \ldots, a_p)'$, entonces

$$Y = a_1\, X_1 + a_2\, X_2 + \ldots + a_p\, X_p = \mathbf{X}\, \mathbf{a},$$

es una variable compuesta.

Propiedades de las variables compuestas

Sean $Y = \mathbf{X}\mathbf{a}$, $Z = \mathbf{X}\mathbf{b}$ dos variables compuestas. Se verifica que:

1. La media de Y es $\overline{y} = \overline{\mathbf{x}}'\mathbf{a}$,
2. La varianza de Y es $s_y^2 = \mathbf{a}'\, \mathbf{S}\, \mathbf{a}$,
3. La covarianza entre Y y Z es $s_{yz} = \mathbf{a}'\, \mathbf{S}\, \mathbf{b} = \mathbf{b}'\, \mathbf{S}\, \mathbf{a} = s_{zy}$.

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

En la transparencia anterior hemos visto que una variable compuesta queda definida por un vector de coeficientes. Pero, de forma más general, una matriz $\mathbf{T}$ de tamaño $p \times q$ definirá q variables compuestas $Y_1, Y_2, \ldots, Y_q$. La expresión

$$\mathbf{Y} = \mathbf{X}\,\mathbf{T},$$

donde $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_q)$, da lugar a una matriz $n \times q$ que contiene los valores de q nuevas variables para los n individuos de la muestra.

Las columnas de $\mathbf{Y} = \mathbf{X}\,\mathbf{T}$ se llaman también variables transformadas. En realidad, $\mathbf{Y}$ es una transformación lineal de la matriz de datos $\mathbf{X}$.

Además, si $\mathbf{T}$ es una matriz ortogonal (es decir $\mathbf{T}\,\mathbf{T}' = \mathbf{T}'\,\mathbf{T} = \mathbf{I}$), entonces $\mathbf{Y}$ es el resultado de aplicar una rotación o una reflexión respecto de un hiperplano. Mediante las nuevas coordenadas $\mathbf{Y}$ los individuos se encuentran repesentados en un sistema de ejes ortogonales.

Propiedades

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

Ejercicio 2.4 *Demostrar las propiedades anteriores.*

2.4 Teorema de la dimensión

La matriz $\mathbf{S}$ es semidefinida positiva, puesto que, $\forall \mathbf{a} \in \mathbb{R}^p$,

$$\mathbf{a}'\,\mathbf{S}\,\mathbf{a} = \frac{1}{n}\mathbf{a}'\,\mathbf{X}'\,\mathbf{H}\,\mathbf{X}\,\mathbf{a} = \frac{1}{n}\mathbf{a}'\,\mathbf{X}'\,\mathbf{H}\,\mathbf{H}\,\mathbf{X}\,\mathbf{a} = \mathbf{b}'\,\mathbf{b} = \|\mathbf{b}\|^2 \geq 0,$$

donde $\mathbf{b} = \frac{1}{\sqrt{n}}\mathbf{H}\,\mathbf{X}\,\mathbf{a}$.

El teorema de la dimensión dice que el rango de la matriz $\mathbf{S}$ determina la dimensión del espacio vectorial generado por las variables observables, es decir, que el número de variables linealmente independientes es igual al rango de $\mathbf{S}$.

TEOREMA 2.1 *Si $r = \mathsf{rang}(\mathbf{S}) \leq p$, entonces hay r variables linealmente independientes y las otras $p - r$ son combinación lineal de estas r variables.*

Demostración: Sea $\mathbf{X}_0 = \mathbf{H}\,\mathbf{X}$ la matriz de datos centrados de tamaño $n \times p$. Observemos que la matriz de covarianzas de $\mathbf{X}$ puede escribirse en función de $\mathbf{X}_0$ como

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\,\mathbf{H}\,\mathbf{X} = \frac{1}{n}\mathbf{X}'\,\mathbf{H}\,\mathbf{H}\,\mathbf{X} = \frac{1}{n}\mathbf{X}_0'\mathbf{X}_0,$$

donde hemos usado que $\mathbf{H}^2 = \mathbf{H}$ y $\mathbf{H}' = \mathbf{H}$.

Utilizando una de las propiedades del rango (propiedad 5), sabemos que $\mathsf{rang}(\mathbf{S}) = \mathsf{rang}(\mathbf{X}_0)$. Por tanto, si $\mathsf{rang}(\mathbf{X}_0) = r \leq p$ significa que existen r variables X_j's linealmente independientes y que el resto $p - r$ son combinación lineal de estas variables. □

Corolario 2.1 *Si todas las variables tienen varianza no nula y* $r = \mathsf{rang}(\mathbf{R}) \leq p$*, entonces hay* r *variables linealmente independientes y las otras* $p - r$ *son combinación lineal de estas* r *variables.*

Demostración: Puesto que $\mathbf{R} = \mathbf{D}_s^{-1}\,\mathbf{S}\,\mathbf{D}_s^{-1}$, donde $\mathbf{D}_s = \mathsf{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots, \sqrt{s_{pp}})'$, entonces la matriz de covarianzas puede escribirse como

$$\mathbf{S} = \mathbf{D}_s\,\mathbf{R}\,\mathbf{D}_s.$$

Finalmente, utilizando otra propiedad del rango (propiedad 1 del rango de matrices cuadradas), se tiene que $\mathsf{rang}(\mathbf{R}) = \mathsf{rang}(\mathbf{S})$. □

2.5 Distancias

Algunos métodos del Análisis Multvariante están basados en criterios geométricos y en la noción de distancia entre individuos y entre poblaciones.

Consideremos la matriz de datos $\mathbf{X}$ en su representación por filas. Sean $\mathbf{S}$ su matriz de covarianzas, $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ y $\mathbf{x}_j' = (x_{j1}, x_{j2}, \ldots, x_{jp})$, respectivamente, las filas i y j de $\mathbf{X}$. Las definiciones más importantes de distancia entre dos individuos son:

1. Distancia euclídea (al cuadrado)
$$d_E^2(i,j) = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2.$$
2. Distancia de K.Pearson (al cuadrado)
$$d_P^2(i,j) = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{D}^{-1}(\mathbf{x}_i - \mathbf{x}_j) = \sum_{k=1}^{p}\frac{(x_{ik} - x_{jk})^2}{s_{kk}},$$
donde $\mathbf{D} = \mathsf{diag}(s_{11}, s_{22}, \ldots, s_{pp})$.

3. Distancia de Mahalanobis (al cuadrado)
$$d_M^2(i,j) = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j).$$

Observación 1 *La distancia* d_E *supone implícitamente que las variables son incorreladas. Además* d_E *no es invariante frente a cambios de escala (cambios en las unidades de medida de las variables).*

Consideremos el cambio de escala $y = \alpha\, x$*, donde* $\alpha \in \mathbb{R}$*,* $\alpha \neq 1$*. Ahora las puntuaciones de los individuos* i *y* j *son* $\mathbf{y}_i = \alpha\,\mathbf{x}_i$ *e* $\mathbf{y}_j = \alpha\,\mathbf{x}_j$*, y la distancia euclídea es*

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

Observación 2 *La distancia* d_P *supone que ls variables son incorreladas, pero es invariante frente a cambios de escala.*

Como anteriormente, consideremos el cambio de escala $y = \alpha\, x$*, donde* $\alpha \in \mathbb{R}$*,* $\alpha \neq 1$*, que da lugar a* $\mathbf{y}_i = \alpha\,\mathbf{x}_i$ *e* $\mathbf{y}_j = \alpha\,\mathbf{x}_j$*. Observemos que las varianzas de las* p *variables están afectadas por el cambio de escala, siendo ahora* $\alpha^2 s_{11}, \alpha^2 s_{22}, \ldots, \alpha^2 s_{pp}$*. De manera que la distancia de Pearson es*

$d_P^2(i,j) = (\mathbf{y}_i - \mathbf{y}_j)'(\alpha^2\mathbf{D})^{-1}(\mathbf{y}_i - \mathbf{y}_j) =$
$\alpha^2(\mathbf{x}_i - \mathbf{x}_j)'\frac{1}{\alpha^2}\mathbf{D}^{-1}(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{D}^{-1}(\mathbf{x}_i - \mathbf{x}_j)$.

Observación 3 *La distancia d_M tiene en cuenta las correlaciones entre las variables y es invariante frente a transformaciones lineales de las variables (en particular, es invariante frente a cambios de escala).*

Observación 4 *La distancia d_E es un caso particular de la distancia d_M cuando $\mathbf{S} = \mathbf{I}$. La distancia d_P es un caso particular de la distancia d_M cuando $\mathbf{S} = \mathsf{diag}(s_{11}, s_{22}, \ldots, s_{pp})$.*

Observación 5 *La ditancia d_M es la más adecuada porque tiene en cuenta las relaciones entre las variables, es decir, no presupone que sean incorreladas ni que tengan varianza unidad. En cambio, utilizar d_E significa suponer que las variables están incorreladas y tienen varianza unidad. Utlizar d_P implica suponer que las variables están incorreladas, pero con varianzas distintas (y distintas a la unidad, generalmente).*

Ejercicios computacionales

Ejercicio 2.5 *Escribir un programa en Matlab que calcule la distancia de Mahalanobis entre las filas de una matriz de datos* $\mathbf{X}$.