ANALISIS DE DATOS MULTIDIMENSIONALES

CONCEPTOS BASICOS

VECTOR ALEATORIO.

Un vector eleatorio es un vector cuyas coordenadas son variables aleatorias:

$$\underline{X} = \begin{bmatrix} x_1 \\ x_2 \\ : \\ x_p \end{bmatrix} \qquad (1)$$

CONCEPTOS MUESTRALES

Supóngase que se toma una muestra de tamaño n de una población definida por el vector aleatorio p-dimensional (1). Esta muestra consiste en n observaciones p-dimensionales, que se organizan en forma de una matriz denominada "Matriz de datos"

Matriz de Datos

Es una matriz cuyas filas son observaciones de una población definida por un vector aleatorio. Por lo tanto cada columna de la matriz de datos corresponde a observaciones de una variable o caracterírtica, y cada fila corresponde a observaciones o casos multivariantes o multidimensionales. Una matriz de datos es una muestra de una población multivante.

$$\begin{array}{c} \text{variables o características} \\ X = \begin{bmatrix} x_{i1} & x_{i2} & .. & x_{ip} \\ x_{2i} & x_{22} & .. & x_{2p} \\ .. & .. & & \\ .. & .. & & \\ .. & .. & & \\ .. & .. & & \\ x_{n1} & x_{n2} & .. & x_{np} \end{bmatrix} \end{array} \text{ observaciones o casos}$$

Vector de Promedios o de medias muestrales

Está formado por los promedios de los valores de las columnas, es decir, los promedios muestrales de cada variable, de una matriz de datos:

$$\underline{\overline{X}} = \begin{bmatrix} \frac{1}{n}\sum x_{i1} \\ \frac{1}{n}\sum x_{i2} \\ .. \\ .. \\ \frac{1}{n}\sum x_{ip} \end{bmatrix} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ .. \\ .. \\ \overline{x}_p \end{bmatrix} = \frac{1}{n} X' \underline{1}$$

en que $\underline{1} = \begin{bmatrix} 1 \\ 1 \\ .. \\ .. \\ 1 \end{bmatrix}$

Matriz de varianza covarianza muestral.

Está formada por las varianzas y covarianzas muestrales, calculadas en base a una matriz de datos.

$$S = \begin{bmatrix} s_{11} & s_{12} & .. & s_{1p} \\ s_{21} & s_{22} & .. & s_{2p} \\ .. & .. & & .. \\ s_{p1} & s_{p2} & .. & s_{pp} \end{bmatrix}$$

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) \qquad \text{si } i \text{ es distinto de } j,$$
$$\text{y} \qquad s_{ii} = \frac{1}{n}\sum_{k=1}^{n} (x_{ki} - \overline{x}_i)^2 \qquad \text{en la diagonal,}$$

$\overline{x}_i$ y $\overline{x}_j$ son los promedios de las columnas (variables) i y j, respectivamente. La matriz de varianzas-covarianzas es cuadrada y simétrica.

En forma matricial, en términos de la matriz de datos X, se puede escribir como

$$S = \frac{1}{n} X' \left(I - \frac{I}{n} 11' \right) X = \frac{1}{n} X' H X$$

en que H es la matriz de centrado

$$H = I - \frac{I}{n} 11'$$

Se puede ver que esta es una matriz simétrica e idempotente ($H \times H = H$).

Matriz de correlaciones muestral.

Contiene unos en la diagonal y las respectivas correlaciones muestrales fuera de ella.

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

en que r_{ij} es la correlación muestral entre las variables (columnas) i y j, $r_{ij} = \frac{s_{ij}}{\sqrt{s_i}\sqrt{s_j}}$, la covarianza entre las variables i y j dividida por sus respectivas desviaciones standard.

En forma matricial se puede escribir

$$R = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

en que $D^{-\frac{1}{2}}$ es la matriz diagonal que contiene los inversos de las desviaciones estádar.

$$D^{-\frac{1}{2}} = \begin{bmatrix} \sqrt{s_{11}} & 0 & .. & 0 \\ 0 & \sqrt{s_{22}} & .. & 0 \\ .. & .. & & .. \\ 0 & 0 & .. & \sqrt{s_{pp}} \end{bmatrix}$$

VALORES Y VECTORES PROPIOS DE UNA MATRIZ

Sea M una matriz cuadrada $p \times p$. Un *vector propio* $\underline{v}$ de M, asociado al *valor propio* λ son tales que

$$M\underline{v} = \lambda\underline{v}$$

Los valores propios pueden ser números complejos.

El número de valores propios es igual a la dimensión de la matriz, aunque pueden haber valores propios iguales. En tales casos se dice que el valor propio tiene multiplicidad 2, 3, etc., según el número de veces que se repite el mismo valor. Hay un vector propio distinto asociado a cada valor propio. Si un valor propio tiene multiplicidad r mayor que uno, entonces hay un r valores propios distintos (no paralelos) asociados a ese valor propio.

Propiedades:

1) Si $\underline{v}$ es un vector propio de una matriz M, cualquier múltiplo de $\underline{v}$ tambien es vector propio. Basta ver que si $M\underline{v} = \lambda\underline{v}$ entonces para cualquier contante k se cumple que $M(k\underline{v}) = \lambda(k\underline{v})$, entonces $k\underline{v}$ es otro vector propio asociado al mismo valor propio. Se suele trabajar con vectores

propios normalizados, es decir, tales que su norma es 1. De estos hay dos asociados a cada valor propio.

2) Sea M una matriz simétrica. Entonces sus valores propios son reales y sus vectores propios son ortogonales.

3) *Descomposición espectral.* Sea M una matriz simétrica. Sea Γ una matriz cuyas columnas son los vectores propios normalizados, y sea $\Lambda = diag\{\lambda_1, \lambda_2 ... \lambda_p\}$, los λ_i sus valores propios, ordenados de acuerdo al orden de los vectores propios en Γ. Entonces Γ es una matriz ortogonal ($\Gamma'\Gamma = \Gamma\Gamma' = I$) y M se puede expresar como

$$M = \Gamma\Lambda\Gamma'$$

o bien se puede escribir

$$\Gamma' M \Gamma = \Lambda \qquad \text{o} \qquad M\Gamma = \Gamma\Lambda$$

4) El determinante de una matriz se puede expresar en términos de sus valores propios como

$$\det M = \prod_{i=1}^{p} \lambda_i$$

5) La traza de una matriz se puede expresar en términos de sus valores propios como

$$trM = \sum_{i=1}^{p} \lambda_i$$

Matrices definidas y semidefinidas positivas

Una matriz M *pxp* es *semidefinida positiva* si para todo vector p-dimensional $\underline{x}$,

$$\underline{x}' M \underline{x} \geq 0$$

Una matriz M *pxp* es *definida positiva* si para todo vector p-dimensional $\underline{x} \neq \underline{0}$,

$$\underline{x}' M \underline{x} > 0$$

En forma análoga, se pueden definir matrices semidefinidas negativas y definidas negativas.

Propiedades:

1) Una matriz M es semidefinida positiva si y sólo si todos sus valores propios son no-negativos.

2) Una matriz M es definida positiva si y sólo si todos sus valores propios son positivos.

3) Una matriz es definida positiva si y sólo si es no-singular (invertible)

Ejemplo 1:

$$M = \begin{bmatrix} 10 & 3 \\ 4 & 6 \end{bmatrix}$$

Sus valores propios son las soluciones λ al sistema de ecuaciones

$$|M - \lambda I| = 0$$

En este caso el sistema es

$$\begin{vmatrix} 10-\lambda & 3 \\ 4 & 6-\lambda \end{vmatrix} = (10-\lambda)(6-\lambda) - 12 = \lambda^2 - 16\lambda + 48 = 0$$

de donde los valores propios son : $\lambda_1 = 12$ y $\lambda_2 = 4$

Los vectores propios se obtienen reemplazando λ por su respectivo valor en la ecuación

$$M\underline{v} = \lambda\underline{v}$$

Se obtiene asociado a λ_1 el vector propio $\underline{V}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ o cualquier múltiplo de él y a λ_2 el vector propio $\underline{V}_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ o cualquier múltiplo de él.

Los vectores propios normalizados son:

$$\underline{V}_1^* = \begin{bmatrix} 3/\sqrt{13} \\ 2/13\sqrt{} \end{bmatrix}$$

$$\underline{V}_2^* = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}$$

Ejemplo 2 La matriz

$$\begin{bmatrix} 17 & -4 & -7 \\ -4 & 14 & -4 \\ -7 & -4 & 17 \end{bmatrix}$$

tiene valores propios $\lambda_1 = 2$, $\lambda_2 = \lambda_3 = 3$.

Sus respectivos vectores propios normalizados son

$$\underline{v}_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad \underline{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \qquad \underline{v}_3 = \frac{1}{\sqrt{6}} \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix}$$

PROBABILIDAD ASOCIADA A VECTORES ALEATORIOS

Sea Si $\underline{X}$ un *vector aleatorio*. Se define la *función de distribución de probabilidad* (o función de distribución de probabilidad acumulada) como la función

$$F : R^p \longrightarrow R$$
$$F(x_1, x_2, .., x_p) = F(\underline{x}) = \Pr(X_1 \leq x_1, X_2 \leq x_2, .., X_p \leq x_p)$$

Esta última expresión e puede escribir como $\Pr(\underline{X} \leq \underline{x})$.

Un vector aleatorio $\underline{X}$ es *continuo* si existe una función $f : R^p \longrightarrow R$ tal que

$$F(\underline{x}) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} .. \int_{-\infty}^{x_p} f(t_1, t_2, ..t_p) dt_1 dt_2 ..dt_p = \int_{-\infty}^{\underline{x}} f(\underline{t}) d\underline{t}$$

f se denomina *función de densidad* de $\underline{X}$

Sea $\underline{X}$ un vector aleatorio continuo, sea f su función de densidad, y sea D un subconjunto de R^p. Entonces la probabilidad de D está dada por

$$\Pr(\underline{X} \in D) = \int_D f(\underline{t}) d\underline{t}$$

Un vector aleatorio $\underline{X}$ es *discreto* si su probabilidad está concentrada en un número finito o numerable de puntos de R^p, $S = \{\underline{x}_s \,/\, s = 1, 2, ..\}$

El conjunto S se denomina *soporte* del vector aleatorio.

Sea $\underline{X}$ un vector aleatorio discreto y sea $f(\underline{x})$ la función que le asigna probabilidad a cada punto de R^p. Se denomina *función de probabilidad* de $\underline{X}$. Entonces si D es un subconjunto de R^p. La probabilidad de D está dada por

$$\Pr(\underline{X} \in D) = \sum_{t_j \in D} f(\underline{t_j})$$

MOMENTOS

Vector de medias o de esperanzas

Sea $\underline{X}$ un vector aleatorio. Se define el *vector esperado* o *vector de medias* como el vector de los valores esperados de las coordenadas de $\underline{X}$.

$$\underline{\mu} = E(\underline{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ : \\ E(X_p) \end{bmatrix}$$

Tiene la propiedad de linealidad $E(a + bX) = a + bE(\underline{X})$, en que a y b son escalares constantes.

Generalizando el concepto anterior, si $g(\underline{x})$ es una función vectorial $\underline{g} : R^p \longrightarrow R^q$

$$\underline{g}(\underline{X}) = \begin{bmatrix} g_1(\underline{X}) \\ g_2\underline{X}) \\ : \\ g_q(\underline{X}) \end{bmatrix}$$

entonces el vector esperado de $g(\underline{X})$ es el vector q-dimensional

$$E(g(\underline{X})) = \begin{bmatrix} E(g_1(\underline{X})) \\ E(g_2(\underline{X})) \\ : \\ E(g_q(\underline{X})) \end{bmatrix}$$

Se puede generalizar al caso en que $g(\underline{X})$ entrega como imagen una matriz qxr. El siguiente es un caso de esta generalización.

Matriz de varianzas-covarianzas

Se define la matriz de varianzas-covarianzas $\Sigma = Var(\underline{X})$ del vector aleatorio $\underline{X}$ como la matriz pxp simétrica que tiene en su i-esimo elemento de la diagonal a la varianza de la coordenada x_i y en la posición (i, j) a la covarianza entre las coordenadas x_i y x_j .

$$\Sigma = Var(\underline{X}) = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & ... & cov(x_1, x_p) \\ cov(x_2, x_1) & var(x_2) & & ... \\ ... & & ... & ... \\ cov(x_p, x_1) & ... & ... & var(x_p) \end{bmatrix}$$

Propiedades:

Si $A_{n\times p}$ y $B_{p\times m}$ son matrices de constantes, entonces

$$Var(AX) = AVar(X)A' \qquad \text{matriz} n \times n$$

$$Var(XB) = B'Var(X)B \qquad \text{matriz} m \times m$$

Matrix de correlaciones

La matriz de correlaciones está formada por una diagonal de unos, y por las correlaciones respectivas, fuera de la diagonal.

$$R = Corr(\underline{X}) = \begin{bmatrix} 1 & corr(x_1, x_2) & ... & corr(x_1, x_p) \\ corr(x_2, x_1) & 1 & & ... \\ ... & & ... & ... \\ corr(x_p, x_1) & ... & ... & 1 \end{bmatrix}$$

Matrices de covarianzas.

Sean $\underline{X}_{p\times 1}$ e $\underline{Y}_{q\times 1}$dos matrices de datos. Se define la *matriz de covarianza*s de $\underline{X}$ e $\underline{Y}$ como la matriz $p \times q$ que contiene

todas las covarianzas entre pares de elementos de $\underline{X}$ y de Y.

$$Cov(\underline{X}, \underline{Y}') = \begin{bmatrix} cov(x_1, y_1) & cov(x_1, y_2) & .. & cov(x_1, y_q) \\ cov(x_2, y_1) & cov(x_2, y_2) & .. & cov(x_2, y_q) \\ .. & .. & .. & .. \\ cov(x_p, y_1) & cov(x_p, y_2) & .. & cov(x_p, y_q) \end{bmatrix}$$

Propiedades:

Si $A_{n\times p}$ y $B_{m\times p}$ son matrices de constantes, entonces

$$Cov(AX, (BY)')) = ACov(X, Y)B' \qquad \text{matriz}$$

$n \times m$

PROBABILIDADES MARGINALES Y CONDICIONALES

Considere la partición $\underline{X} = \begin{bmatrix} \underline{X}_1 \\ -- \\ \underline{X}_2 \end{bmatrix}$ en que $\underline{X}_1$ tiene k elementos y $\underline{X}_2$ tiene p-k elementos.

La *distribución marginal* del vector aleatorio $\underline{X}_1$ se define como $\Pr(\underline{X}_1 \leq \underline{x}_1) = F(x_1, x_2, .., x_k, \infty, \infty, .., \infty)$.

Si $\underline{X}$ es continuo y su densidad es $f(\underline{x}) = f(\underline{x}_1, \underline{x}_2)$ entonces la densidad marginal de $\underline{X}_1$ está dada por

$$f_1(x_1) = \int_{-\infty}^{\infty} f(\underline{t}_1, \underline{t}_2) d\underline{t}_2$$

La *densidad condicional* de $\underline{X}_2$ dado $\underline{X}_1 = \underline{x}_1$ se define como

$f(\underline{X}_2/\underline{X}_1 = \underline{x}_1) = \frac{f(\underline{x}_1, \underline{X}_2)}{f_1(\underline{x}_1)}$ en los puntos en que $f_1(x_1) \neq 0$, y se define como 0 en los demás puntos.

Dos vectores aleatorios distintos pueden tener densidades marginales iguales .

El caso discreto es análogo. En adelante sólo se darán las definiciones y propiedades para el caso contínuo, entendiéndose que hay una situación análoga para el caso discreto.

INDEPENDENCIA

Dos vectores aleatorios $\underline{X}_1$ y $\underline{X}_2$ son *independientes* si y sólo si la densidad conjunta de $\underline{X}$=$(\underline{X}_1, \underline{X}_2)$ cumple la propiedad

$$f(\underline{x}_1, \underline{x}_2) = f_1(x_1)\ f_2(x_2)$$