Reducción de la factorialidad. Componentes pricipales.

Ricardo Alberich, Juan Gabriel Gomila y Arnau Mir

Introducción a la estadística descriptiva multidimensional

Vectores aleatorios

- Vector aleatorio. Una \red{variable aleatoria p-dimensional, o vector aleatorio de dimensión p, es un vector (fila) compuesto por p variables aleatorias:

$$\underline{X} = (X_1, X_2, \ldots, X_p)$$

- Como en el caso de las variables aleatorias unidimensionales, es importante distinguir entre los vectores aleatorios (los modelos teóricos), y las realizaciones o las muestras de los mismos, que corresponden a una o varias mediciones concretas de las variables que forman dichos vectores.

Vectores aleatorios

*Por ejemplo, si llamamos X_1 a la variable aleatoria que da la edad de un individuo (en años), X_2 a la que da su altura (redondeada a cm) y X_3 a la que da su peso (redondeada a kg con una cifra decimal), entonces

$$\underline{X} = (X_1, X_2, X_3)$$

- es un vector aleatorio de dimensión 3. Cada vez que medimos la edad, la altura y el peso de una persona, y organizamos estos datos en este orden como un vector numérico, obtenemos una realización de $\underline{X}$.

Vectores aleatorios

- Sea ahora $\underline{X} = (X_1, X_2, \ldots, X_p)$ un vector aleatorio y, para cada $i = 1, \ldots, p$, sean μ_i y σ_i la media y la desviación típica, respectivamente, de su componente X_i.
- El valor esperado, o vector de medias, de $\underline{X}$ es el vector formado por los valores esperados, o medias, de sus componentes:

$$E(\underline{X}) = (E(X_1), \ldots, E(X_p)) = (\mu_1, \ldots, \mu_p)$$

También lo denotaremos simplemente $\boldsymbol{\mu}$.

Vectores aleatorios

- El vector de varianzas de $\underline{X}$ es el vector formado por las varianzas de sus componentes:

$$Var(\underline{X}) = (Var(X_1), \ldots, Var(X_p)) = (\sigma_1^2, \ldots, \sigma_p^2)$$

- El vector de desviaciones típicas de $\underline{X}$ es el vector formado por las desviaciones típicas de sus componentes:

$$\sigma(\underline{X}) = (\sigma(X_1), \ldots, \sigma(X_p)) = (\sigma_1, \ldots, \sigma_p)$$

Vectores aleatorios; tipificación

Tipificación.

- Cuando sumamos una constante b a una variable aleatoria X, decimos que efectuamos un cambio de origen, puesto que desplazamos todos los valores de X la cantidad b.
- Cuando multiplicamos una variable aleatoria X por una constante $a \neq 0$, decimos que efectuamos un cambio de escala, puesto que la agrandamos (si $|a| > 1$) o la encogemos (si $0 < |a| < 1$) en el factor constante a. Si además $a < 0$, cambiamos el signo de la variable.

Vectores aleatorios; tipificación

- Sea X una variable aleatoria de media μ y desviación típica σ. Recordemos que si $a, b \in \mathbb{R}$, entonces $a \cdot X + b$ es una variable aleatoria de media, varianza y desviación típica, respectivamente,

- $$E(a \cdot X + b) = a \cdot \mu + b$$

- $$Var(aX + b) = a^2 \cdot \sigma^2$$

- $$\sigma(aX + b) = |a| \cdot \sigma.$$

Tipificación de variables aleatorias

Llamaremos la variable tipificada de X a la variable aleatoria

$$Z = \frac{X - \mu}{\sigma}.$$

Por ejemplo, cuando construimos una variable aleatoria normal estándar Z a partir de una variable normal X, lo que hacemos es tipificar esta última.

Las fórmulas anteriores implican que si Z es una variable tipificada, entonces $E(Z) = 0$ y $\sigma(Z) = 1$.

Tipificación de variables aleatorias

- Si $\underline{X} = (X_1, \ldots, X_p)$ es un vector aleatorio, su vector tipificado $\underline{Z}$ se obtiene substituyendo cada X_i por su variable tipificada Z_i.
- Matricialmente, $\underline{Z}$ se puede obtener a partir de $\underline{X}$ de la manera siguiente. Para cada $i = 1, \ldots, p$, sean μ_i y σ_i la media y la desviación típica de X_i, respectivamente. Sean

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p) \qquad A = \begin{pmatrix} \sigma_1^{-1} & 0 & \ldots & 0 \\ 0 & \sigma_2^{-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_p^{-1} \end{pmatrix}$$

el vector de medias de $\underline{X}$ y ....

Tipificación de variables aleatorias

la matriz diagonal que tiene en la diagonal principal los inversos de las desviaciones típicas de las componentes de $\underline{X}$. Entonces,

$$\underline{Z} = (\underline{X} - \boldsymbol{\mu}) \cdot A = \left(A(\underline{X} - \boldsymbol{\mu})^t\right)^t.$$

(Como ejercicio de manejo de matrices, comprobadlo para $p = 3$.)
Esta expresión es un caso particular de transformación lineal multivariante de $\underline{X}$. Una transformación lineal general emplearía una matriz real cualquiera de p columnas con entradas números reales en el lugar de A, y un vector fila real cualquiera de dimensión p en el lugar de $\boldsymbol{\mu}$.

Covarianzas

Covarianzas. Dadas dos variables aleatorias X_1 y X_2 de medias μ_1 y μ_2, respectivamente, se define su covarianza como

$$Cov(X_1, X_2) = E((X_1 - \mu_1)(X_2 - \mu_2)).$$

Es fácil comprobar que la covarianza también se puede calcular mediante la identidad

$$Cov(X_1, X_2) = E(X_1 \cdot X_2) - \mu_1 \cdot \mu_2.$$

Covarianzas

En efecto

$$\begin{aligned} Cov(X_1, X_2) &= E((X_1 - \mu_1)(X_2 - \mu_2)) = E(X_1X_2 - \mu_1X_2 - \mu_2X_1 + \mu_1 \\ &= E(X_1X_2) - \mu_1E(X_2) - \mu_2E(X_1) + \mu_1\mu_2 \\ &= E(X_1X_2) - \mu_1\mu_2 - \mu_2\mu_1 + \mu_1\mu_2 = E(X_1X_2) - \mu_1\mu_2 \end{aligned}$$

Covarianzas; propiedades

- La covarianza de X_1 y X_2 puede tomar cualquier valor real, y mide si las dos variables aleatorias se comportan igual, en el sentido siguiente: si valores grandes de una variable corresponden a valores grandes de la otra, su covarianza es positiva
- En el caso opuesto, si valores grandes de una variable corresponden a valores pequeños de la otra, su covarianza es negativa.
- Si las dos variables aleatorias son independientes, entonces su covarianza es 0, puesto que en este caso

Covarianzas; propiedades

La covarianza es simétrica, $Cov(X_1, X_2) = Cov(X_2, X_1)$. La covarianza de una variable aleatoria consigo misma es su varianza:

$$Cov(X, X) = E((X - \mu)^2) = Var(X).$$

Para simplificar la notación, se suele utilizar σ para indicar las covarianzas. Dadas dos variables aleatorias X_i y X_j que formen parte de un vector aleatorio, escribiremos

$$\sigma_{ij} = Cov(X_i, X_j) \text{ y } \sigma_{ii} = Cov(X_i, X_i) = \sigma_i^2.$$

Covarianzas; propiedades

Igual que en el caso unidimensional, un vector aleatorio $\underline{X} = (X_1, \ldots, X_p)$ posee una medida de su dispersión respecto de su valor esperado $\boldsymbol{\mu}$.

Es su llamada matriz de covarianzas y se define como

$$
\begin{aligned}
Cov(\underline{X}) &= E((\underline{X} - \boldsymbol{\mu})^t \cdot (\underline{X} - \boldsymbol{\mu}) \\
&= E\left(\left(X_1 - \mu_1 X_2 - \mu_2 \vdots X_p - \mu_p\right) \cdot (X \right. \\
&= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}
\end{aligned}
$$

Covarianzas; propiedades

Es decir, la matriz de covarianzas de $\underline{X}$ tiene como entrada (i, j) la covarianza σ_{ij} de X_i y X_j.

Se puede comprobar fácilmente que esta matriz se puede calcular como

$$Cov(\underline{X}) = E(\underline{X}^t \cdot \underline{X}) - \boldsymbol{\mu}^t \cdot \boldsymbol{\mu}$$

La matriz de covarianzas de $\underline{X}$ también se suele representar por $\mathring{}$.

Covarianzas; Matriz semi-definida positiva

Las matrices de covarianzas satisfacen la propiedad fundamental siguiente.

Definición

Diremos que una matriz cuadrada de números reales

$$M = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

es semi definida positiva si para todo $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$.

Covarianzas; Matriz semi-definida positiva

$$\begin{pmatrix} y_1 & y_2 & \dots y_n \end{pmatrix} \cdot \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = y^t \cdot M \cdot y \geq 0.$$

Si la desigualdad es estricta se dice que la matriz es definida positiva,

Covarianzas; propiedades

Teorema

- Las matrices de covarianzas son simétricas (y por lo tanto diagonalizables) y semi-definidas positivas.
- Las matrices reales semi-definidas positivas diagonalizan y todos sus valores propios $\lambda \geq 0$.

Correlaciones

Correlaciones. Como las covarianzas son difíciles de comparar, para medir si dos variables aleatorias se comportan igual o no, se usa el llamado coeficiente de correlación lineal de Pearson, que es una medida adimensional de la relación entre dos variables.

Definimos la correlación de las variables X_i y X_j como

$$Cor(X_i, X_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

A menudo denotaremos $Cor(X_i, X_j)$ por medio de ρ_{ij}.

Correlaciones; propiedades

Las correlaciones tienen las propiedades siguientes:

- $-1 \leq \rho_{ij} \leq 1$.
- $\rho_{ij} = \rho_{ji}$ y $\rho_{ii} = 1$.
- Si $\sigma_i = \sigma_j = 1$, entonces $\rho_{ij} = \sigma_{ij}$.
- Salvo en el signo, ρ_{ij} es invariante por cambios de origen y escala: es decir, si $a_i, a_j, b_i, b_j \in \mathbb{R}$ y $a_i, a_j \neq 0$,

$$Cor(a_i X_i + b_i, a_j X_j + b_j) = \pm Cor(X_i, X_j)$$

donde el signo que aparece es el del cociente a_i/a_j.

Correlaciones; propiedades

- Si $\rho_{ij} = \pm 1$, las variables tienen una relación lineal perfecta, es decir, existen $\alpha \neq 0$ y β tales que $X_i = \alpha X_j + \beta$. La pendiente α tiene el mismo signo que la correlación.
- Si $\rho_{ij} = 0$, decimos que las variables X_i y X_j son incorreladas. Notemos que la correlación es 0 si, y sólo si, la covarianza es 0. Por lo tanto, dos variables aleatorias independientes son incorreladas. El recíproco en general es falso.

Matriz de correlaciones

La matriz de correlaciones de un vector aleatorio $\underline{X} = (X_1, \ldots, X_p)$ como

$$Cor(\underline{X}) = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1p} \\ \rho_{21} & 1 & \ldots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \ldots & 1 \end{pmatrix}$$

Matriz de correlaciones; propiedades

Esta matriz tiene la propiedad fundamental siguiente.

Teorema

{La matriz de correlaciones de un vector aleatorio $\underline{X}$ es igual a la matriz de covarianzas de su vector tipificado $\underline{Z}$:}

$$Cor(\underline{X}) = Cov(\underline{Z}).$$

En efecto, si $Z_i = \frac{X_i - \mu_i}{\sigma_i}$ y $Z_j = \frac{X_j - \mu_j}{\sigma_j}$ son las variables tipificadas de X_i y X_j, respectivamente, entonces

$$Cov(Z_i, Z_j) = Cor(Z_i, Z_j) = Cor\left(\frac{1}{\sigma_i}X_i - \frac{\mu_i}{\sigma_i}, \frac{1}{\sigma_j}X_j - \frac{\mu_j}{\sigma_j}, \right) = Cor(X_i, X_j$$

donde la primera igualdad es consecuencia de la tercera propiedad de las correlaciones en la lista anterior, y la última igualdad es consecuencia de la cuarta propiedad en la misma.

Descripción de datos multivariantes

- Los vectores aleatorios son el modelo teórico que usaremos cuando manejemos simultáneamente diversas variables aleatorias sobre los mismos individuos. Las realizaciones de un vector aleatorio serán las observaciones de las variables que lo componen sobre individuos concretos de la población.
- Datos multivariantes. Supongamos que hemos obtenido los valores de p variables aleatorias $X_1, \ldots, X_p$ sobre un conjunto de n individuos u objetos. Es decir, tenemos n observaciones de p variables. En cada observación, los valores que toman estas variables forman un vector que será una realización del vector aleatorio $\underline{X} = (X_1, X_2, \ldots, X_p)$.

Descripción de datos multivariantes

Estas observaciones se pueden organizar de manera matricial, de manera que cada fila sea una realización de $\underline{X}$:

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

Descripción de datos multivariantes; notación

Utilizaremos las notaciones siguientes:

*Denotaremos por

$$\mathbf{x}_{i\bullet} = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

la i-ésima realización de $\underline{X}$, es decir, el vector fila compuesto por las observaciones de las p variables sobre el i-ésimo individuo.

*Denotaremos por

$$\mathbf{x}_{\bullet j} = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

el vector columna compuesto por las n observaciones de la j-ésima variable.

De esta manera, podremos expresar la matriz de datos $\mathbf{X}$ por filas o por columnas:

Descripción de datos multivariantes; fórmulas matriciales

Con estas notaciones podemos generalizar al caso multidimensional algunas definiciones ya conocidas de los estadísticos más usuales de una muestra. Dada una matriz $\mathbf{X}$ de n observaciones de un vector aleatorio $\underline{X} = (X_1, X_2, \ldots, X_p)$:

*El vector de medias de $\mathbf{X}$ es el vector fila cuyas entradas son las medias aritméticas de las realizaciones de cada variable X_i:

$$\overline{\mathbf{X}} = (\overline{\mathbf{x}}_{\bullet 1}, \overline{\mathbf{x}}_{\bullet 2}, \ldots, \overline{\mathbf{x}}_{\bullet p})$$

donde, para cada $j = 1, \ldots, p$,

$$\overline{\mathbf{x}}_{\bullet j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}.$$

Observemos que

$$\overline{\mathbf{X}} = (\overline{\mathbf{x}}_{\bullet 1}, \overline{\mathbf{x}}_{\bullet 2}, \ldots, \overline{\mathbf{x}}_{\bullet p}) = \frac{1}{n} \Big(\sum_{i=1}^{n} x_{i1}, \sum_{i=1}^{n} x_{i2}, \ldots, \sum_{i=1}^{n} x_{ip} \Big)$$

Descripción de datos multivariantes; fórmulas matriciales

Es decir, el vector de medias de $\mathbf{X}$ es la media aritmética de sus filas.

*El vector de varianzas de $\mathbf{X}$ es el vector fila cuyas entradas son las varianzas de las realizaciones de cada variable X_i:

$$s^2_{\mathbf{X}} = (s_1^2, s_2^2, \ldots, s_p^2)$$

donde

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{\mathbf{x}}_{\bullet j})^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 - \overline{\mathbf{x}}_{\bullet j}^2.$$

*El vector de varianzas muestrales de $\mathbf{X}$ es el vector fila cuyas entradas son las varianzas muestrales de las realizaciones de cada variable X_i:

$$\widetilde{s}^2_{\mathbf{X}} = (\widetilde{s}_1^2, \widetilde{s}_2^2, \ldots, \widetilde{s}_p^2)$$

donde

$$\widetilde{s}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{\mathbf{x}}_{\bullet j})^2 = \frac{n}{n-1}s_j^2$$

Descripción de datos multivariantes; fórmulas matriciales

*Los vectores de desviaciones típicas $s_{\mathbf{X}}$ y desviaciones típicas muestrales $\widetilde{s}_{\mathbf{X}}$ se definen como las raíces cuadradas positivas de los vectores de varianzas y varianzas muestrales, respectivamente.

Como en el caso unidimensional, $\overline{\mathbf{X}}$ es un estimador de $E(\underline{X}) = \boldsymbol{\mu}$. Tanto $s^2_{\mathbf{X}}$ como $\widetilde{s}^2_{\mathbf{X}}$ son estimadores del vector de varianzas de $\underline{X}$: el primero es el máximo verosímil y el segundo es insesgado. Y tanto $s_{\mathbf{X}}$ como $\widetilde{s}_{\mathbf{X}}$ son estimadores del vector de desviaciones típicas de $\underline{X}$: de nuevo, el primero es el máximo verosímil y el segundo es insesgado.

Observación. Supondremos en lo que sigue que nuestras matrices de datos no tienen ninguna columna constante, es decir, ninguna columna de varianza 0.

Centralización de una matriz de datos

Centralización de una matriz de datos.

Para centrar una matriz de datos $\mathbf{X}$, se resta a cada columna su media aritmética:

$$\widetilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \overline{\mathbf{x}}_{\bullet 1} & x_{12} - \overline{\mathbf{x}}_{\bullet 2} & \cdots & x_{1p} - \overline{\mathbf{x}}_{\bullet p} \\ x_{21} - \overline{\mathbf{x}}_{\bullet 1} & x_{22} - \overline{\mathbf{x}}_{\bullet 2} & \cdots & x_{2p} - \overline{\mathbf{x}}_{\bullet p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{\mathbf{x}}_{\bullet 1} & x_{n2} - \overline{\mathbf{x}}_{\bullet 2} & \cdots & x_{np} - \overline{\mathbf{x}}_{\bullet p} \end{pmatrix}$$

Llamamos al resultado la matriz de datos centrados de $\mathbf{X}$.

Centralización de una matriz de datos; cálculo matricial

Vamos a ver que esta igualdad admite un cálculo matricial sencillo. Sea $\mathbf{1}_n$ un vector fila formado por n 1's, de manera que $\mathbf{1}_n^t$ es un vector columna de n filas, cada una de las cuales es un 1. Observemos que

$$\mathbf{1}_n^t \cdot \overline{\mathbf{X}} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \cdot (\overline{\mathbf{x}}_{\bullet 1}, \overline{\mathbf{x}}_{\bullet 2}, \ldots, \overline{\mathbf{x}}_{\bullet p}) = \begin{pmatrix} \overline{\mathbf{x}}_{\bullet 1} & \overline{\mathbf{x}}_{\bullet 2} & \ldots & \overline{\mathbf{x}}_{\bullet p} \\ \overline{\mathbf{x}}_{\bullet 1} & \overline{\mathbf{x}}_{\bullet 2} & \ldots & \overline{\mathbf{x}}_{\bullet p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\mathbf{x}}_{\bullet 1} & \overline{\mathbf{x}}_{\bullet 2} & \ldots & \overline{\mathbf{x}}_{\bullet p} \end{pmatrix}$$

y por lo tanto

Centralización de una matriz de datos; cálculo matricial

$$\widetilde{\mathbf{X}} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} - \begin{pmatrix} \overline{\mathbf{x}}_{\bullet 1} & \overline{\mathbf{x}}_{\bullet 2} & \dots & \overline{\mathbf{x}}_{\bullet p} \\ \overline{\mathbf{x}}_{\bullet 1} & \overline{\mathbf{x}}_{\bullet 2} & \dots & \overline{\mathbf{x}}_{\bullet p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\mathbf{x}}_{\bullet 1} & \overline{\mathbf{x}}_{\bullet 2} & \dots & \overline{\mathbf{x}}_{\bullet p} \end{pmatrix} = \mathbf{X} - \mathbf{1}_n^t \cdot \overline{\mathbf{X}}$$

Centralización de una matriz de datos; cálculo matricial

Vamos a calcular ahora $\overline{\mathbf{X}}$ de manera matricial a partir de $\mathbf{X}$.

Observemos que cuando multiplicamos $\mathbf{1}_n$ por una matriz de n filas, obtenemos un vector fila formado por las sumas de sus columnas:

$$(1, 1, \dots, 1) \cdot \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \Big(x_{11} + x_{21} + \dots + x_{n1}, \dots, x_{1p} + x_{2}$$

Por lo tanto

Centralización de una matriz de datos; cálculo matricial

$$
\begin{aligned}
\text{`}\mathit{dataframe}\text{`}\mathit{rac1n}\mathbf{1}_n \cdot \mathbf{X} &= \frac{1}{n}(1,\ldots,1)\cdot\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \\
&= \frac{1}{n}\left(x_{11}+\cdots+x_{n1},\ldots,x_{1p}+\cdots+x_{np}\right) = (\bar{\mathbf{x}}
\end{aligned}
$$

Centralización de una matriz de datos; cálculo matricial

Combinando las igualdades

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n^t \cdot \overline{\mathbf{X}} \quad \text{y} \quad \overline{\mathbf{X}} = \text{`}dataframe\text{`}rac1n\mathbf{1}_n \cdot \mathbf{X}$$

concluimos que

$$\widetilde{\mathbf{X}} = \mathbf{X} - \text{`}dataframe\text{`}rac1n\mathbf{1}_n^t \cdot \mathbf{1}_n \cdot \mathbf{X} = (I_n - \text{`}dataframe\text{`}rac1n\mathbf{1}_n^t \cdot \mathbf{1}_n)\mathbf{X}$$

Si definimos

$$\mathbf{H}_n = I_n - \frac{1}{n}\mathbf{1}_n^t \cdot \mathbf{1}_n.$$

Centralización de una matriz de datos; cálculo matricialel propiedades

Teorema $\widetilde{\mathbf{X}} = \mathbf{H}_n \cdot \mathbf{X}$.

A esta matriz $\mathbf{H}_n$ se la llama la matriz centralizadora de orden n. Notemos que

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{pmatrix} \quad 1 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Centralización de una matriz de datos; cálculo matricial propiedades

Se comprueban fácilmente las dos propiedades siguientes:

- $\mathbf{H}_n \cdot \mathbf{H}_n = \mathbf{H}_n$; es decir, $\mathbf{H}_n$ es una matriz idempotente.
- $\mathbf{H}_n$ es simétrica, tiene rango $n-1$ y $\mathbf{H}_n \cdot \mathbf{1}_n^t = 0$.

Ejemplo centrado matricial

Ejemplo

Consideremos la matriz de datos

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

que contiene $n = 4$ observaciones de $p = 3$ variables. Para centralizarla a mano, basta restar a cada columna su media aritmética. Como

$$\mathbf{x}_{\bullet 1} = \frac{1+1+2+3}{4} = 1.75,\ \mathbf{x}_{\bullet 2} = \frac{-1+0+3+0}{4} = 0.5,\ \mathbf{x}_{\bullet 3} = \frac{3+3+}{4}$$

Ejemplo centrado matricial

su matriz centralizada es

$$\widetilde{\mathbf{X}} = \begin{pmatrix} 1-1.75 & -1-0.5 & 3-1.75 \\ 1-1.75 & 0-0.5 & 3-1.75 \\ 2-1.75 & 3-0.5 & 0-1.75 \\ 3-1.75 & 0-0.5 & 1-1.75 \end{pmatrix} = \begin{pmatrix} -0.75 & -1.5 & 1.25 \\ -0.75 & -0.5 & 1.25 \\ 0.25 & 2.5 & -1.75 \\ 1.25 & -0.5 & -0.75 \end{pmatrix}$$

Ejemplo centrado matricial

Para calcularla matricialmente, observemos que

$$\mathbf{H}_4 = \begin{pmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{pmatrix} = \begin{pmatrix} 0.75 & -0.25 & -0.25 & - \\ -0.25 & 0.75 & -0.25 & - \\ -0.25 & -0.25 & 0.75 & - \\ -0.25 & -0.25 & -0.25 & \end{pmatrix}$$

Ejemplo centrado matricial

y por lo tanto

$$\widetilde{\mathbf{X}} = \mathbf{H}_4 \cdot \mathbf{X} = \begin{pmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -0. \\ -0. \\ 0.2 \\ 1.2 \end{pmatrix}$$

Ejemplo centrado matricial

Si queremos realizar estos cálculos con R , entramos la matriz de datos, definimos la matriz centralizadora y operamos:

```
X=cbind(c(1,1,2,3),c(-1,0,3,0),c(3,3,0,1))
# vector de medias
colMeans(X)
```

```
## [1] 1.75 0.50 1.75
```

```
#Construimos la matriz centralizadora H4
H4=diag(4)-1/4
H4
```

```
##       [,1]  [,2]  [,3]  [,4]
## [1,]  0.75 -0.25 -0.25 -0.25
## [2,] -0.25  0.75 -0.25 -0.25
## [3,] -0.25 -0.25  0.75 -0.25
## [4,] -0.25 -0.25 -0.25  0.75
```

Tipificación de datos

Dado un vector de datos, llamaremos vector de datos tipificados al vector que se obtiene restando a cada valor la media aritmética del vector y dividiendo el resultado por su desviación típica. De esta manera, se obtiene un vector de datos de media aritmética 0 y varianza$\sim$1.

Dada una matriz de datos $\mathbf{X}$, llamaremos su matriz tipificada a la matriz $\mathbf{Z}$ que se obtiene tipificando cada columna. Es decir, la tipificación de una matriz de datos $\mathbf{X}$ consiste en primero calcular su matriz centrada $\widetilde{\mathbf{X}}$ y a continuación dividir cada columna de esta última por la desviación típica s_i de la correspondiente columna de $\mathbf{X}$:

$$\mathbf{Z} = \begin{pmatrix} \frac{x_{11}-\overline{\mathbf{x}}_{\bullet 1}}{s_1} & \frac{x_{12}-\overline{\mathbf{x}}_{\bullet 2}}{s_2} & \cdots & \frac{x_{1p}-\overline{\mathbf{x}}_{\bullet p}}{s_p} \\ \frac{x_{21}-\overline{\mathbf{x}}_{\bullet 1}}{s_1} & \frac{x_{22}-\overline{\mathbf{x}}_{\bullet 2}}{s_2} & \cdots & \frac{x_{2p}-\overline{\mathbf{x}}_{\bullet p}}{s_p} \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

Tipificación de datos

Si llamamos $\mathbf{D}$ a la matriz diagonal $p \times p$ que tiene en su diagonal principal las desviaciones típicas de las columnas correspondientes de $\mathbf{X}$, entonces $\mathbf{D}^{-1}$ es la matriz diagonal $p \times p$ que tiene en su diagonal principal los inversos de estas desviaciones típicas.

$$\mathbf{D} = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & s_p \end{pmatrix} \qquad \mathbf{D}^{-1} = \begin{pmatrix} \frac{1}{s_1} & 0 & \dots & 0 \\ 0 & \frac{1}{s_2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{s_p} \end{pmatrix}$$

es inmediato comprobar que la matriz tipificada $\mathbf{Z}$ de $\mathbf{X}$ se obtiene como

$$\mathbf{Z} = \widetilde{\mathbf{X}} \cdot \mathbf{D}^{-1} = \mathbf{H}_n \cdot \mathbf{X} \cdot \mathbf{D}^{-1}$$

Ejemplo tipificación de datos

Ejemplo Vamos a tipificar la matriz de datos

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

del Ejemplo R . Calculamos las varianzas de las columnas: son

$$s_1^2 = \frac{11}{16}, \quad s_2^2 = \frac{9}{4}, \quad s_3^2 = \frac{27}{16}$$

Entonces

$$\mathbf{D}^{-1} = \begin{pmatrix} \frac{1}{\sqrt{11/16}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{9/4}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{27/16}} \end{pmatrix}$$

y por lo tanto

Ejemplo tipificación de datos

$$
\begin{aligned}
\mathbf{Z} &= \mathbf{H}_4 \cdot \mathbf{X} \cdot \mathbf{D}^{-1} \\
&= \begin{pmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{11/16}} & & \\ & 0 & \\ & & \\ & & 0 \end{pmatrix} \\
&= \begin{pmatrix} -3/\sqrt{11} & -1 & 5/(3\sqrt{3}) \\ -3/\sqrt{11} & -1/3 & 5/(3\sqrt{3}) \\ 1/\sqrt{11} & 5/3 & -7/(3\sqrt{3}) \\ 5/\sqrt{11} & -1/3 & -3/(3\sqrt{3}) \end{pmatrix}
\end{aligned}
$$

Ejemplo tipificación de datos

Para tipificar con R , hay que recordar que para aplicar funciones a las columnas de una matriz, hay que usar la construcción `apply(matriz,MARGIN=2,fun)`.

```
X=cbind(c(1,1,2,3),c(-1,0,3,0),c(3,3,0,1))
#Vector de medias
m=apply(X,MARGIN=2,mean)  #o m=colMeans(X)
m
```

```
## [1] 1.75 0.50 1.75
```

```
#Vector de desviaciones típicas
desv.tip.muest=apply(X,MARGIN=2,sd)
n=dim(X)[1]
desv.tip=desv.tip.muest*sqrt((n-1)/n)
desv.tip
```

```
## [1] 0.8291562 1.5000000 1.2990381
```

Covarianzas muestrales

Dada una matriz de datos $\mathbf{X} = (\mathbf{x}_{\bullet 1}, \mathbf{x}_{\bullet 2}, \ldots, \mathbf{x}_{\bullet p})$, se define la covarianza de las columnas $\mathbf{x}_{\bullet i}$ y $\mathbf{x}_{\bullet j}$ como

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left((x_{ki} - \overline{\mathbf{x}}_{\bullet i})(x_{kj} - \overline{\mathbf{x}}_{\bullet j})\right) = \frac{1}{n}\Big(\sum_{k=1}^{n} x_{ki}x_{kj}\Big) - \overline{\mathbf{x}}_{\bullet i}\overline{\mathbf{x}}_{\bullet j}$$

y se define la covarianza muestral de $\mathbf{x}_i$ y $\mathbf{x}_j$ como

$$\widetilde{s}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}\left((x_{ki} - \overline{\mathbf{x}}_{\bullet i})(x_{kj} - \overline{\mathbf{x}}_{\bullet j})\right) = \frac{n}{n-1}s_{ij}$$

Matriz de covarianz<as muestrales

El estadístico s_{ij} es el estimador máximo verosímil de la covarianza σ_{ij} de las variables aleatorias X_i y X_j, mientras que $\tilde{s}_{ij}$ es un estimador insesgado de dicha covarianza.

Es inmediato comprobar que

$$s_{ij} = s_{ji},\ \widetilde{s}_{ij} = \widetilde{s}_{ji},\ s_{ii} = s_i^2,\ \widetilde{s}_{ii} = \widetilde{s}_i^2$$

Ejemplo cálculo de la matriz de covarianz<as muestrales

La covarianza de las dos primeras columnas de la matriz de datos

$$\mathbf{x} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

se obtendría de la manera siguiente

$$s_{12} = \frac{1}{4}(1\cdot(-1)+1\cdot 0+2\cdot 3+3\cdot 0)-1.75\cdot 0.5 = 1.25-0.875 = 0.375$$

Su covarianza muestral se obtendría multiplicando por $4/3$ este valor

$$\widetilde{s}_{12} = \frac{4}{3}s_{12} = 0.5.$$

Ejemplo cálculo de la matriz de covarianz<as muestrales

La covarianza muestral de dos vectores de datos de la misma longitud se puede calcular con R mediante la función `cov`:

```
x1=c(1,1,2,3)
x2=c(-1,0,3,0)
cov(x1,x2)    #covarianza muestral
```

```
## [1] 0.5
```

```
(3/4)*cov(x1,x2)  #covarianza a secas
```

```
## [1] 0.375
```

```
sum(x1*x2)/4-mean(x1)*mean(x2)
```

```
## [1] 0.375
```

Queremos recalcar que, con en el caso de la varianza con `var`, R calcula con `cov` la versión muestral de la covarianza. Para pasar de

Matriz de varinzas covarianzas; propiedades

La matriz de covarianzas de $\mathbf{X}$ representa la variabilidad conjunta de los datos de dicha matriz. Es una matriz simétrica y semidefinida positiva (tiene todos sus valores propios ≥ 0). Además, se tiene el resultado siguiente, que permite su cálculo matricial.

Teorema $\{$*Si* $\mathbf{S}$ *es la matriz de covarianzas de* $\mathbf{X}$*, entonces*$\}$

$$\mathbf{S} = \frac{1}{n}\widetilde{\mathbf{X}}^t \cdot \widetilde{\mathbf{X}} = \frac{1}{n}\mathbf{X}^t \cdot \mathbf{H}_n \cdot \mathbf{X}.$$

Matriz de varianzas covarianzas; ejemplo

Ejemplo Continuemos con el Ejemplo R , donde, recordemos,

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

Vamos a calcular a mano su matriz de covarianzas y comprobar que coincide con la dada por la fórmula anterior. Para realizar los cálculos a mano, es costumbre organizar los datos y los cálculos intermedios necesarios en una tabla como la siguiente:

i	x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3
1	1	-1	3	1	1	9	-1	3	-3
2	1	0	3	1	0	9	0	3	0
3	2	3	0	4	9	0	6	0	0
4	3	0	1	9	0	1	0	3	0
Suma	7	2	7	15	10	19	5	9	-3

Matriz de varianzas covarianzas; ejemplo

Así tenemos que

$$\overline{x}_1 = \frac{7}{4}, \quad \overline{x}_2 = \frac{2}{4}, \quad \overline{x}_3 = \frac{7}{4}$$

$$s_1^2 = \frac{1}{4}\sum_{i=1}^{4} x_{i1}^2 - \overline{\mathbf{x}}_1^2 = \frac{15}{4} - \left(\frac{7}{4}\right)^2 = \frac{11}{16}$$

$$s_2^2 = \frac{1}{4}\sum_{i=1}^{4} x_{i2}^2 - \overline{\mathbf{x}}_2^2 = \frac{10}{4} - \left(\frac{2}{4}\right)^2 = \frac{9}{4}$$

$$s_3^2 = \frac{1}{4}\sum_{i=1}^{4} x_{i3}^2 - \overline{\mathbf{x}}_3^2 = \frac{19}{4} - \left(\frac{7}{4}\right)^2 = \frac{27}{16}$$

$$s_{12} = \frac{1}{4}\sum_{i=1}^{n} x_{i1}x_{i2} - \overline{\mathbf{x}}_1\overline{\mathbf{x}}_2 = \frac{5}{4} - \frac{7}{4}\cdot\frac{2}{4} = \frac{3}{8}$$

$$s_{13} = \frac{1}{4}\sum_{i=1}^{n} x_{i1}x_{i3} - \overline{\mathbf{x}}_1\overline{\mathbf{x}}_3 = \frac{9}{4} - \frac{7}{4}\cdot\frac{7}{4} = -\frac{13}{16}$$

$$s_{23} = \frac{1}{4}\sum_{i=1}^{n} x_{i2}x_{i3} - \overline{\mathbf{x}}_2\overline{\mathbf{x}}_3 = \frac{-3}{4} - \frac{2}{4}\cdot\frac{7}{4} = -\frac{13}{16}$$

Matriz de varianzas covarianzas; ejemplo

. . . por lo tanto, la matriz de covarianzas es

$$\mathbf{S} = \begin{pmatrix} 11/16 & 3/8 & -13/16 \\ 3/8 & 9/4 & -13/8 \\ -13/16 & -13/8 & 27/16 \end{pmatrix}$$

Matriz de varianzas covarianzas; ejemplo

En forma matricial, sería

$$
\begin{aligned}
\mathbf{S} &= \text{`}dataframe\text{`}rac14\mathbf{X}^t \cdot \mathbf{H}_4 \cdot \mathbf{X} \\
&= \text{`}dataframe\text{`}rac14 \begin{pmatrix} 1 & 1 & 2 & 3 \\ -1 & 0 & 3 & 0 \\ 3 & 3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 3/4 & -1/4 & -1/4 & - \\ -1/4 & 3/4 & -1/4 & - \\ -1/4 & -1/4 & 3/4 & - \\ -1/4 & -1/4 & -1/4 & \end{pmatrix} \\
&= \ldots = \begin{pmatrix} 11/16 & 3/8 & -13/16 \\ 3/8 & 9/4 & -13/8 \\ -13/16 & -13/8 & 27/16 \end{pmatrix}
\end{aligned}
$$

Matriz de varianzas covarianzas; ejemplo

Con R , efectuaríamos esta operación de la manera siguiente:

```
X=cbind(c(1,1,2,3),c(-1,0,3,0),c(3,3,0,1))
n=dim(X)[1]
H4=diag(4)-1/4
S=(1/n)*t(X)%*%H4%*%X
S
```

```
##          [,1]   [,2]    [,3]
## [1,]  0.6875  0.375 -0.8125
## [2,]  0.3750  2.250 -1.6250
## [3,] -0.8125 -1.625  1.6875
```

La instrucción cov aplicada a una matriz de datos $\mathbf{X}$ calcula la matriz de covarianzas muestrales de $\mathbf{X}$

$$\widetilde{\mathbf{S}} = (\widetilde{s}_{ij})_{i,j=1,\ldots,p} = \begin{pmatrix} \widetilde{s}_{11} & \widetilde{s}_{12} & \ldots & \widetilde{s}_{1p} \\ \widetilde{s}_{21} & \widetilde{s}_{22} & \ldots & \widetilde{s}_{2p} \\ \cdot & \cdot & & \cdot \end{pmatrix}$$

Matriz de varianzas covarianzas; ejemplo

Para obtener la matriz de covarianzas, es suficiente multiplicar dicha matriz por $(n-1)/n$. Por ejemplo:

```
X=cbind(c(1,1,2,3),c(-1,0,3,0),c(3,3,0,1))
cov(X)  #matriz de covarianzas muestrales(3/4)*cov(X)  #ma
```

```
##              [,1]      [,2]       [,3]
## [1,]  0.9166667  0.500000 -1.083333
## [2,]  0.5000000  3.000000 -2.166667
## [3,] -1.0833333 -2.166667  2.250000
```

Variables redundantes

Deficnición.

Decimos que en una matriz de datos hay variables redundantes cuando una o más columnas aportan la misma información que otra. La redundancia de variables se puede manifestar por ejemplo si una columna $\mathbf{x}_{\bullet i}$ depende linealmente de otras columnas $\mathbf{x}_{\bullet i_1}, \ldots, \mathbf{x}_{\bullet i_k}$, es decir, si existen $a_1, \ldots, a_k, b \in \mathbb{R}$ tales que

$$\mathbf{x}_{\bullet i} = a_1 \mathbf{x}_{\bullet i_1} + \cdots + a_k \mathbf{x}_{\bullet i_k} + b \cdot \mathbf{1}_n^t.$$

En este caso diremos que se da \red{{redundancia por dependencia lineal}, y es la única que consideraremos en esta lección, por lo que cuando hablemos de variables redundantes, nos referiremos realmente a variables redundantes por dependencia lineal.

Diremos que una matriz de datos \red{{tiene k variables redundantes} $\mathbf{x}_{\bullet l_1}, \ldots, \mathbf{x}_{\bullet l_k}$ cuando estas k variables dependen linealmente del resto de variables, $\{\mathbf{x}_{\bullet 1}, \ldots, \mathbf{x}_{\bullet p}\} - \{\mathbf{x}_{\bullet l_1}, \ldots, \mathbf{x}_{\bullet l_k}\}$, pero no entre ellas.

Variables redundantes; ejemplo

Por ejemplo, en la matriz

$$\begin{pmatrix} 1 & -1 & 2 \\ 3 & 0 & 7 \\ -2 & 4 & 1 \end{pmatrix}$$

$\mathbf{x}_{\bullet 3}$ es redundante, puesto que

$$\begin{pmatrix} 2 \\ 7 \\ 1 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 3 \\ -2 \end{pmatrix} + \begin{pmatrix} -1 \\ 0 \\ 4 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Variables redundantes; ejemplo

Despejando en esta igualdad cada una de las otras dos columnas, vemos que también son redundantes, pero esta redundancia se debe a la misma relación. Por tanto, no podemos decir que esta matriz tenga tres variables redundantes, sólo tiene una. En cambio la matriz de datos

$$\begin{pmatrix} 1 & 4 & 3 & -5 & 4 \\ 3 & 9 & 7 & -13 & 8 \\ -2 & -2 & 0 & 4 & 1 \end{pmatrix}$$

tiene dos variables redundantes, dadas por las relaciones lineales

$$\mathbf{x}_{\bullet 4} = \mathbf{x}_{\bullet 1} - 2\mathbf{x}_{\bullet 2} + 2 \cdot \mathbf{1}_3, \quad \mathbf{x}_{\bullet 5} = \mathbf{x}_{\bullet 3} + \mathbf{1}_3$$

Observad que, en estas relaciones, $\mathbf{x}_{\bullet 4}$ no depende de $\mathbf{x}_{\bullet 5}$ ni viceversa.

Propiedades para detectar variables redundantes

La propiedad siguiente de la matriz de covarianzas muestra cómo se pueden detectar cuantas variables redundantes tiene una matriz de datos.

Teorema Sea $\mathbf{S}$ la matriz de covarianzas de una matriz de datos $\mathbf{X}$ de p variables. Entonces, el número de variables redundantes de $\mathbf{X}$ es igual a la multiplicidad de 0 como valor propio de $\mathbf{S}$. En particular:

- Si $|\mathbf{S}| \neq 0$, entonces no existe ninguna variable redundante.
- Si $|\mathbf{S}| = 0$, entonces existe al menos una variable redundante.
- Si el rango de $\mathbf{S}$ es k, entonces existen $p - k$ variables redundantes.

Como la matriz de covarianzas muestrales $\widetilde{\mathbf{S}}$ de $\mathbf{X}$ es un múltiplo escalar de $\mathbf{S}$, este resultado también vale para $\widetilde{\mathbf{S}}$.

Variables redundantes; ejemplo

Recordemos que en un ejemplo anterior

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

Para comprobar si tiene variables redundantes, vamos a calcular el determinante de su matriz de covarianzas muestrales:

```
X=cbind(c(1,1,2,3),c(-1,0,3,0),c(3,3,0,1))
det(cov(X))
```

```
## [1] 0.1481481
```

Como este determinante es diferente de 0, esta matriz de datos no contiene variables redundantes.

Variables redundantes; ejemplo

Consideremos la tabla de datos siguiente:

i	x_1	x_2	x_3
1	1	0	-1
2	1	2	1
3	1	1	0
4	0	3	0

Deseamos saber si contiene variables redundantes. Vamos a calcular a mano su matriz de covarianzas:

Variables redundantes; ejemplo

$$
\begin{aligned}
\mathbf{S} &= \text{`}\mathit{dataframe}\text{`}\mathit{rac}14\mathbf{X}^t \cdot \mathbf{H}_4 \cdot \mathbf{X} \\
&= \text{`}\mathit{dataframe}\text{`}\mathit{rac}14 \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 2 & 1 & 3 \\ -1 & 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 3/4 & -1/4 & -1/4 & - \\ -1/4 & 3/4 & -1/4 & - \\ -1/4 & -1/4 & 3/4 & - \\ -1/4 & -1/4 & -1/4 & \end{pmatrix} \\
&= \begin{pmatrix} 3/16 & -3/8 & 0 \\ -3/8 & 5/4 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}
\end{aligned}
$$

y se comprueba fácilmente que ...

Variables redundantes; ejemplo

$$\begin{vmatrix} 3/16 & -3/8 & 0 \\ -3/8 & 5/4 & 1/2 \\ 0 & 1/2 & 1/2 \end{vmatrix} = 0$$

Por lo tanto, existe al menos una variable redundante.

y se comprueba fácilmente que

Ahora bien, por un lado, $|\mathbf{S}| = 0$ implica que el rango de $\mathbf{S}$ es ≤ 2, y por otro, está claro que el rango de $\mathbf{S}$ es como mínimo 2, puesto que, por ejemplo, las dos primeras columnas son linealmente independientes (en la última fila, la primera columna tiene un 0 y la segunda, una entrada $\neq 0$). Por lo tanto, el rango de $\mathbf{S}$ es 2 y $\mathbf{X}$ contiene una única variable redundante.

Variables redundantes; ejemplo

Para comprobar esto mismo con R, usaremos la matriz de covarianzas muestrales:

```
X=matrix(c(1,0,-1,1,2,1,1,1,0,0,3,0),nrow=4, byrow=TRUE)
S=cov(X)
det(S)
```

```
## [1] 0
```

```
qr(S)$rank
```

```
## [1] 2
```

```
eigen(S)$values
```

```
## [1] 2.109226e+00 4.741076e-01 3.238150e-17
```

Obtenemos que el rango es 2, y también observamos que **S** tiene un solo valor propio igual a 0 (el último, 3.238150e-17). Cada uno de

Varianza total, varianza media y varianza generalizada

Como la matriz de covarianzas como medida de variabilidad es difícil de interpretar, debido a que no es una única cantidad sino toda una matriz, es de desear la existencia de un índice que mida esta variabilidad. Hay diversas propuestas al respecto. Veamos varias:

*La varianza total de $\mathbf{X}$ es la suma de las varianzas de sus columnas.

*La varianza media de $\mathbf{X}$ es la media de las varianzas de sus columnas, es decir, la varianza total partida por el número de columnas.

*La varianza generalizada de $\mathbf{X}$ es el determinante de su matriz de covarianzas. La desviación típica generalizada de $\mathbf{X}$ es la raíz cuadrada positiva de su varianza generalizada. Se puede demostrar que ésta última es el (hiper)volumen del (hiper)poliedro de $\mathbb{R}^p$ definido por las filas de $\mathbf{X}$ consideradas como puntos de $\mathbb{R}^p$.

Correlaciones

Correlaciones.

Se define la correlación lineal de Pearson de las dos columnas $\mathbf{x}_{\bullet i}$ y $\mathbf{x}_{\bullet j}$ de una matriz de datos $\underline{X}$ como

$$r_{ij} = \frac{s_{ij}}{s_i s_j}$$

Observad que

$$\frac{\widetilde{s}_{ij}}{\widetilde{s}_i \cdot \widetilde{s}_j} = \frac{\frac{n}{n-1} \cdot s_{ij}}{\sqrt{\frac{n}{n-1}} \cdot s_i \cdot \sqrt{\frac{n}{n-1}} \cdot s_j} = \frac{s_{ij}}{s_i \cdot s_j} = r_{ij}$$

y por lo tanto este coeficiente de correlación se puede calcular también a partir de la covarianza y las desviaciones típicas muestrales por medio de esta misma fórmula.

Correlaciones; propiedades

La correlación r_{ij} estima el parámetro poblacional $\rho_{ij} = Cor(X_i, X_j)$, y tiene las propiedades siguientes:

* $-1 \leq r_{ij} \leq 1$.

* $r_{ii} = 1$.

* r_{ij} tiene el mismo signo que s_{ij}.

* $r_{ij} = \pm 1$ si y, sólo si, existe una relación lineal perfecta entre las variables $\mathbf{x}_{\bullet i}$ y $\mathbf{x}_{\bullet j}$. O sea, si, y sólo si, existen valores $a, b \in \mathbb{R}$ tales que $\mathbf{x}_{\bullet j} = a\mathbf{x}_{\bullet i} + b \cdot \mathbf{1}_n$. La pendiente a de esta recta tiene el mismo signo que la correlación entre las variables.

Correlaciones: ejemplo

En ejemplos anteriores hemos calculado la covarianza y las varianzas de las dos primeras columnas de la matriz de datos

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

Hemos obtenido

$$s_{12} = 0.375, \quad s_1 = \sqrt{\frac{11}{16}} = 0.82916, \quad s_2 = \sqrt{\frac{9}{4}} = 1.5$$

y por lo tanto su correlación es

$$r_{12} = \frac{0.375}{0.82916 \cdot 1.5} = 0.3015$$

Correlaciones: ejemplo

Con R, la correlación de Pearson de dos vectores se puede calcular

Matyriz de correlaciones

Definición.

Llamaremos la matriz de correlaciones (de pearson) de la matriz de datos $\mathbf{X}$ a

$$\mathbf{R} = (r_{ij})_{i,j=1,\ldots,p} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}$$

Matriz de correlaciones; propiedades

Propiedades

Esta matriz tiene las propiedades siguientes:

*La matriz $\mathbf{R}$ es semi-definida positiva.

- Si todas las variables son incorreladas entonces $\mathbf{R} = I_p$ y $\det(\mathbf{R}) = 1$.
- $\mathbf{R}$ cumple las mismas propiedades que la matriz de covarianzas por lo que concierne a las variables redundantes. Por ejemplo, si $\det(\mathbf{R}) = 0$, entonces hay al menos una variable redundante.

*$|\mathbf{R}| \leq 1$.

Matriz de correlaciones; cálculo matricial

La matriz de correlaciones de puede calcular de forma matricial de la manera siguiente. Recordemos que

$$\mathbf{D} = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & s_p \end{pmatrix} \qquad \mathbf{D}^{-1} = \begin{pmatrix} \frac{1}{s_1} & 0 & \dots & 0 \\ 0 & \frac{1}{s_2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{s_p} \end{pmatrix}$$

Matriz de correlaciones; cálculo matricial

Propiedad

$\mathbf{R} = \mathbf{D}^{-1} \cdot \mathbf{S} \cdot \mathbf{D}^{-1}$.

De este resultado podemos despejar la matriz de covarianzas, y obtenemos que

$$\mathbf{S} = \mathbf{D} \cdot \mathbf{R} \cdot \mathbf{D}$$

Tenemos también el resultado siguiente.

Teorema $\{$*Si* $\mathbf{Z}$ *es la matriz de datos tipificados de* $\mathbf{X}$*, entonces la matriz de covarianzas de* $\mathbf{Z}$ *es igual a la matriz de correlaciones de* $\mathbf{X}$*.*$\}$

Matriz de correlaciones; ejemplo cálculo matricial

Ejemplo Continuemos con el Ejemplo R , donde, recordemos,

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

En otros ejemplos ya hemos calculado su matriz de inversos de desviaciones típicas y su matriz de covarianzas:

$$\mathbf{D}^{-1} = \begin{pmatrix} \frac{1}{\sqrt{11/16}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{9/4}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{27/16}} \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} 11/16 & 3/8 & -13/16 \\ 3/8 & 9/4 & -13/8 \\ -13/16 & -13/8 & 27/16 \end{pmatrix}$$

y por lo tanto su matriz de correlaciones es

Matriz de correlaciones; ejemplo cálculo matricial

$$\mathbf{R} = \begin{pmatrix} \frac{1}{\sqrt{11/16}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{9/4}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{27/16}} \end{pmatrix} \cdot \begin{pmatrix} \frac{11}{16} & \frac{3}{8} & -\frac{13}{16} \\ \frac{3}{8} & \frac{9}{4} & -\frac{13}{8} \\ -\frac{13}{16} & -\frac{13}{8} & \frac{27}{16} \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{11/16}} & 0 \\ 0 & \frac{1}{\sqrt{9/4}} \\ 0 & 0 \end{pmatrix}$$

Matriz de correlaciones; ejemplo cálculo matricial

La matriz de correlaciones de una matriz de datos se puede calcular con R con la misma instrucción \texttt{cor}.cor‘.

```
X=matrix(c(1,0,-1,1,2,1,1,1,0,0,3,0),nrow=4, byrow=TRUE)
cor(X)
```

```
##              [,1]       [,2]      [,3]
## [1,]  1.0000000 -0.7745967 0.0000000
## [2,] -0.7745967  1.0000000 0.6324555
## [3,]  0.0000000  0.6324555 1.0000000
```

Comprobemos que da lo mismo que el producto de matrices anterior

```
#Matriz diagonal de inversas de desviaciones típicas
desv.tip=apply(X,MARGIN=2,sd)*sqrt(3/4)
Dm=diag(1/desv.tip)
#Matriz de covarianzas
S=(3/4)*cov(X)
```

Ejercicio

Consideremos la siguiente matriz de datos

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & -2 & 0 \end{pmatrix}$$

Calculad de forma matricial las matrices siguientes:

- Su matriz de datos tipificados **Z**
- Su matriz de covarianzas **S**
- Su matriz de correlaciones **R**

Determinad si esta matriz de datos tiene variables redundantes.