UNIVERSIDAD MAYOR

Medidas descriptivas de variables aleatorias multivariadas

Hossein T. Dinani

31 de Agosto 2023

Medidas descriptivas

- Para un set de datos describimos datos a través de :
  - Media
  - Desviación estándar : Estima dispersión
    - Varianza es la desviación estándar al cuadrado
  - Covarianza: estima la relación lineal entre dos variables aleatorias
  - Correlación: estima la relación estandarizada entre dos variables
  - Distancia
  - Coeficiente de asimetría y kurtosis

Media

- La media para datos multivariados es un vector que contiene las medias de cada una de las variables:

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

donde $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\,, \quad j = 1, \ldots, p.$

- También se puede escribirlo como $\bar{\mathbf{x}} = \frac{1}{n} X^T \mathbf{1}_{n\times 1}$, donde $\mathbf{1}_{n\times 1}$ es el vector identidad.

Varianza

- La varianza es la media de las distancias de los valores a su media al cuadrado:

$$\mathbf{s} = \begin{bmatrix} s_1^2 \\ \vdots \\ s_p^2 \end{bmatrix}$$

donde $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2, \quad j = 1, \ldots, p.$

- Nota que hemos ocupado la varianza insesgada que es un mejor estimador: Estamos usando la media muestral en lugar de la media poblacional.

Covarianza

- La covarianza determina el grado de variación conjunta de dos variables aleatorias respecto a sus medias:

$$s_{12} = cov(x_1, x_2) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

- La covarianza nos permite determinar si hay dependencia lineal entre las dos variables.

- En el caso multivariado, tenemos una covarianza para cada par de variables y se define una matriz de covarianza $S_{p\times p}$:

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{pmatrix}$$

donde los elementos en la diagonal son las varianzas: $s_{11} = s_1^2$.

- Matriz de covarianza es simétrica: $cov(x_i, x_j) = cov(x_j, x_i)$.

Matriz de covarianza

- El valor de la covarianza depende en las unidades de las variables.
- Por lo tanto con la matriz de covarianza solo se puede decir si hay una relación lineal entre las variables y su tendencia: No se puede comentar sobre la magnitud de la relación (fuerte o débil).
- El signo de la covarianza determina la tendencia en la relación lineal entre las dos variables:
  - Para $s_{ij} > 0$ la relación entre x_i y x_j es directa.
  - Si $s_{ij} < 0$ la relación es inversa.
- Si $s_{ij} = 0$ no hay una relación lineal entre las dos variables.

Matriz de covarianza

- Podemos escribir la matriz de covarianza usando matriz de datos centrados:

$$\tilde{X} = X - \mathbf{1}_{n\times 1}\bar{\mathbf{x}}^T$$

- Por lo tanto podemos escribir:

$$S = \frac{1}{n-1}\tilde{X}^T\tilde{X}$$

Propiedades de la matriz de covarianza

- Propiedades de la matriz de covarianza:
  - Es semi-definida positiva: sus valores propios $\lambda_j \geq 0, \quad j = 1, \ldots, p$
  - $|S| = \prod_{j=1}^{p} \lambda_j \geq 0$.
  - Si $|S| = 0$ tenemos algunas variables que son combinación lineal de otras variables.
  - El rango de la matriz $rg(S)$ está dado por numero de los valores propios que no sean cero
  - $Tr(S) = \sum_{j=1}^{p} s_j^2 = \sum_{j=1}^{p} \lambda_j$

Matriz de correlaciones

- La correlación es la versión normalizada de la covarianza:

$$r_{12} = \frac{cov(x_1, x_2)}{s_1 s_2}$$

- La matriz de correlación:

$$R = \begin{pmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{pmatrix}$$

donde $r_{ij} \in [-1,1]$.

- La correlación mide la magnitud de la relación lineal entre dos variables aleatorias sin influencia de las unidades de las variables.

Matriz de correlación

- Podemos estandarizar las variables (sin unidades): $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$.

- La covarianza entre las variables estandarizadas z_i y z_j está dado por

$$r_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} z_{ij} z_{ik}$$

- Se puede escribir la matriz de correlación en términos de la matriz de covarianza:

$$R = D^{-1/2} S D^{-1/2}$$

Donde D es una matriz diagonal que contiene las varianzas: $D = \begin{pmatrix} s_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_p^2 \end{pmatrix}$

Matriz de correlación

- Propiedades de la matriz de correlación
  - Es semi-definida positiva: sus autovalores $\lambda'_j \geq 0, \quad j = 1, \ldots, p$
  - $|R| = \prod_{j=1}^{p} \lambda'_j \geq 0$
  - Si $|R| = 0$ tenemos algunas variables que son combinación lineal de otras variables
  - El rango de la matriz $rg(R)$ esta dado por numero de los valores propios que no sean cero
  - $Tr(R) = \sum_{j=1}^{p} \lambda'_j =?$

Transformación lineal

- Se utiliza una transformación lineal sobre la matriz de datos para cambio de unidades o cambiar numero de las variables.

- Consideramos $C_{p\times r}$ y definimos $Y = XC$, el vector de media $\bar{\mathbf{y}}_{r\times 1}$ y la matriz de covarianza están dados por

$$\bar{\mathbf{y}} = C^T\bar{\mathbf{x}}, \qquad S_y = C^T S_x C$$

➢ Demuestra.

Transformación lineal: Ejemplo

- Consideramos datos de Iris. La media y la matriz de covarianza están dados por

$$\bar{\mathbf{x}}^T = (5.84, 3.05, 3.75, 1.19), \qquad S_x = \begin{pmatrix} 0.68 & -0.04 & 1.27 & 0.51 \\ -0.04 & 0.18 & -0.32 & -0.12 \\ 1.27 & -0.32 & 3.11 & 1.29 \\ 0.51 & -0.12 & 1.29 & 0.58 \end{pmatrix}$$

- Para crear una nueva variable *y* que es la suma de largos y anchos de sépalo y pétalo, definimos

$$\mathbf{c} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

➢ La media y varianza de $Y = X\mathbf{c}$ son?

Transformación lineal: Ejercicio

- Escribe el set de datos de Iris basado en dos siguientes variables

  Variable 1: La suma de la longitud del pétalo y del sépalo de cada flor
  Variable 2: La suma del ancho del pétalo y del sépalo de cada flor

Estandarización

- Matriz de datos centrados: $\tilde{X} = X - \mathbf{1}_{n\times 1}\bar{\mathbf{x}}^T$

- Definimos estandarización univariada: $Y = \tilde{X}D_x^{-1/2}$

  donde D_x es la matriz diagonal que contiene las varianzas.

- El vector de media y la matriz de covarianza de la variable $\mathbf{y}$:

$$\bar{\mathbf{y}} = \mathbf{0}_{p\times 1}, \qquad S_y = D_x^{-1/2} S_x D_x^{-1/2} = R_x$$

- Estandarización multivariada: $Z = \tilde{X}S_x^{-1/2}$

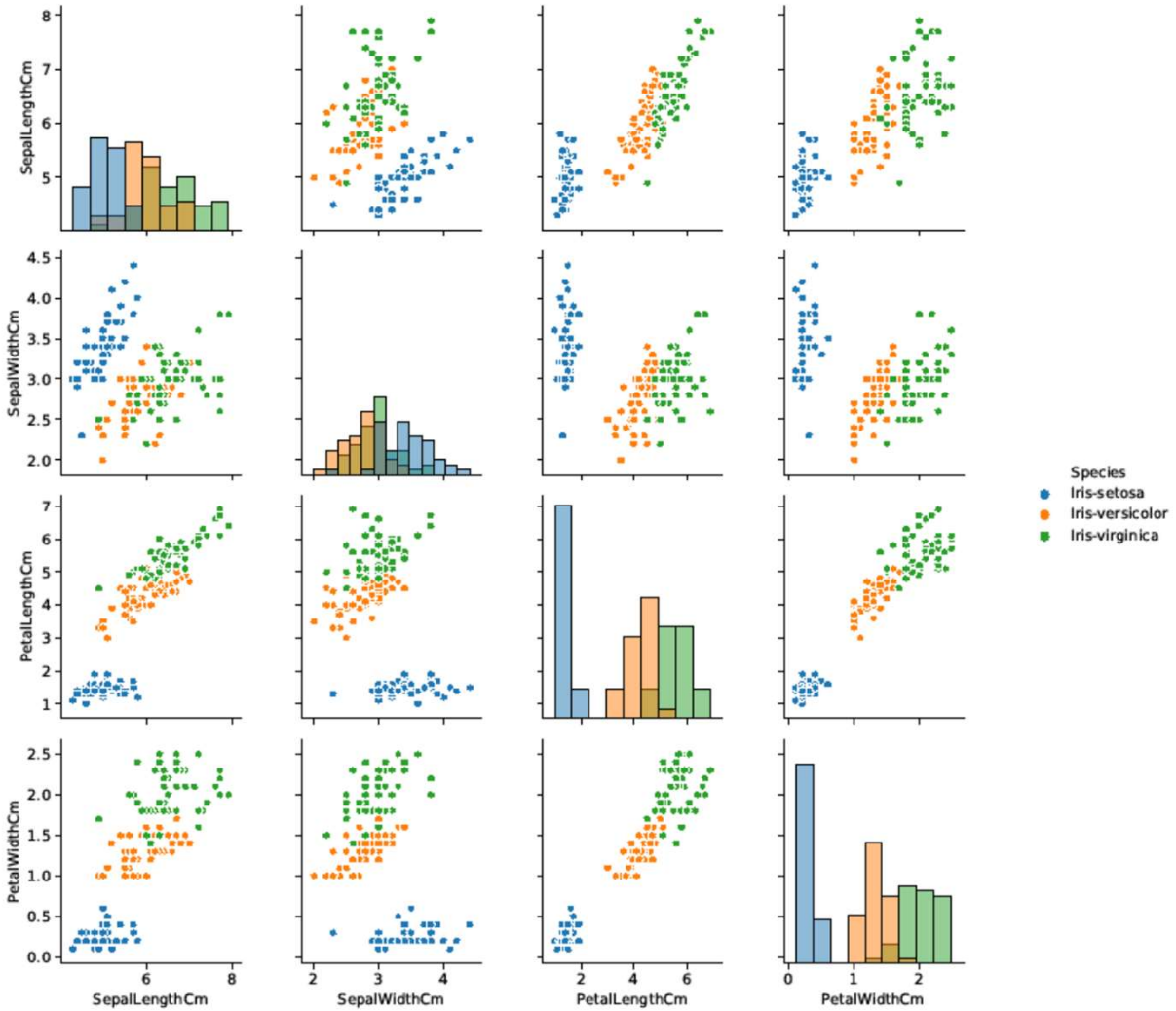
- El vector de media y la matriz de covarianza de la variable $\mathbf{z}$:

$$\bar{\mathbf{z}} = \mathbf{0}_{p\times 1}, \qquad S_z = \left(S_x^{-1/2}\right)^T S_x S_x^{-1/2} = I_{p\times p}$$

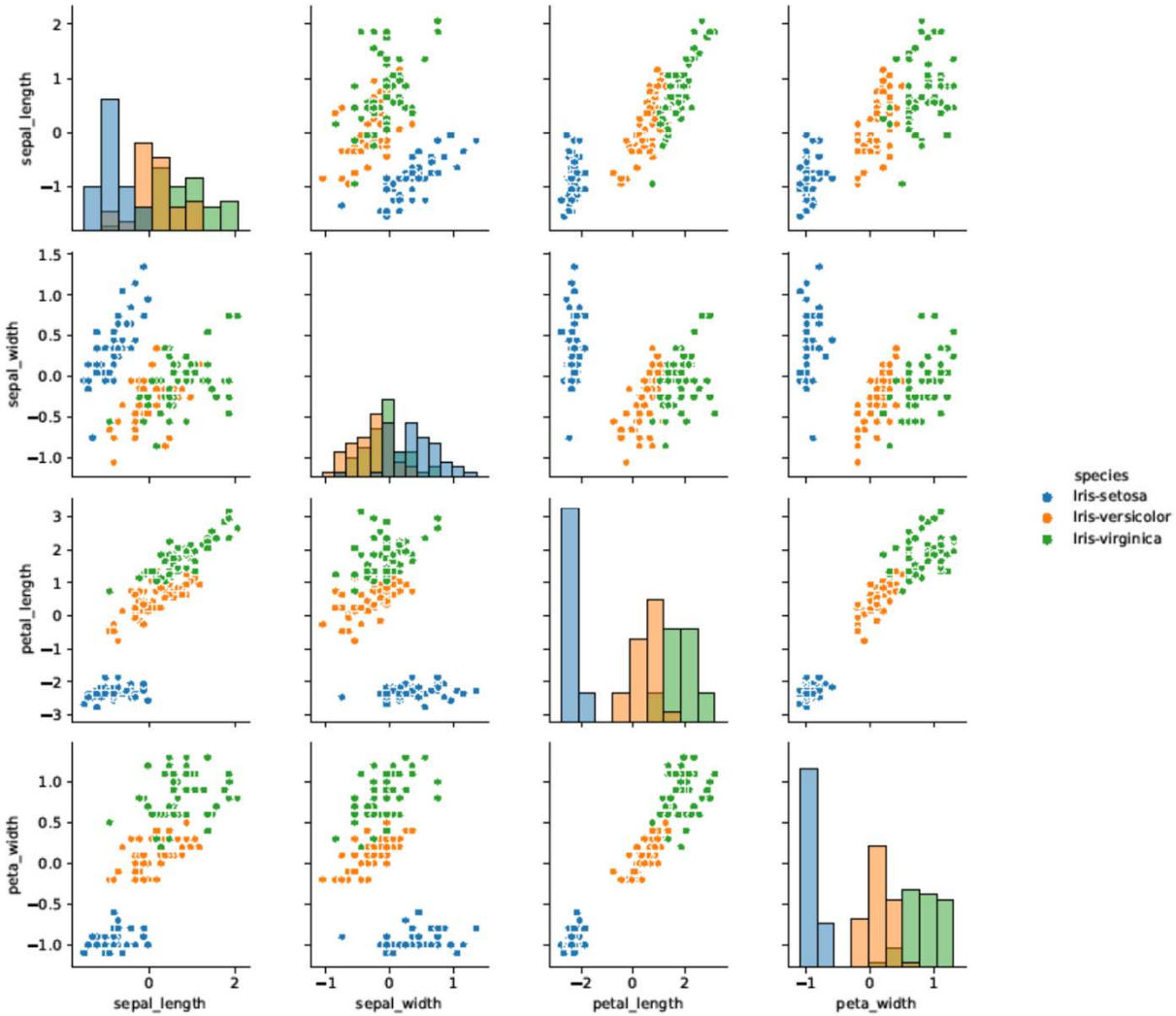
  donde S_x es la matriz de covarianza de variable $\mathbf{x}$.

- Las variables $\mathbf{z}$ son incorreladas.

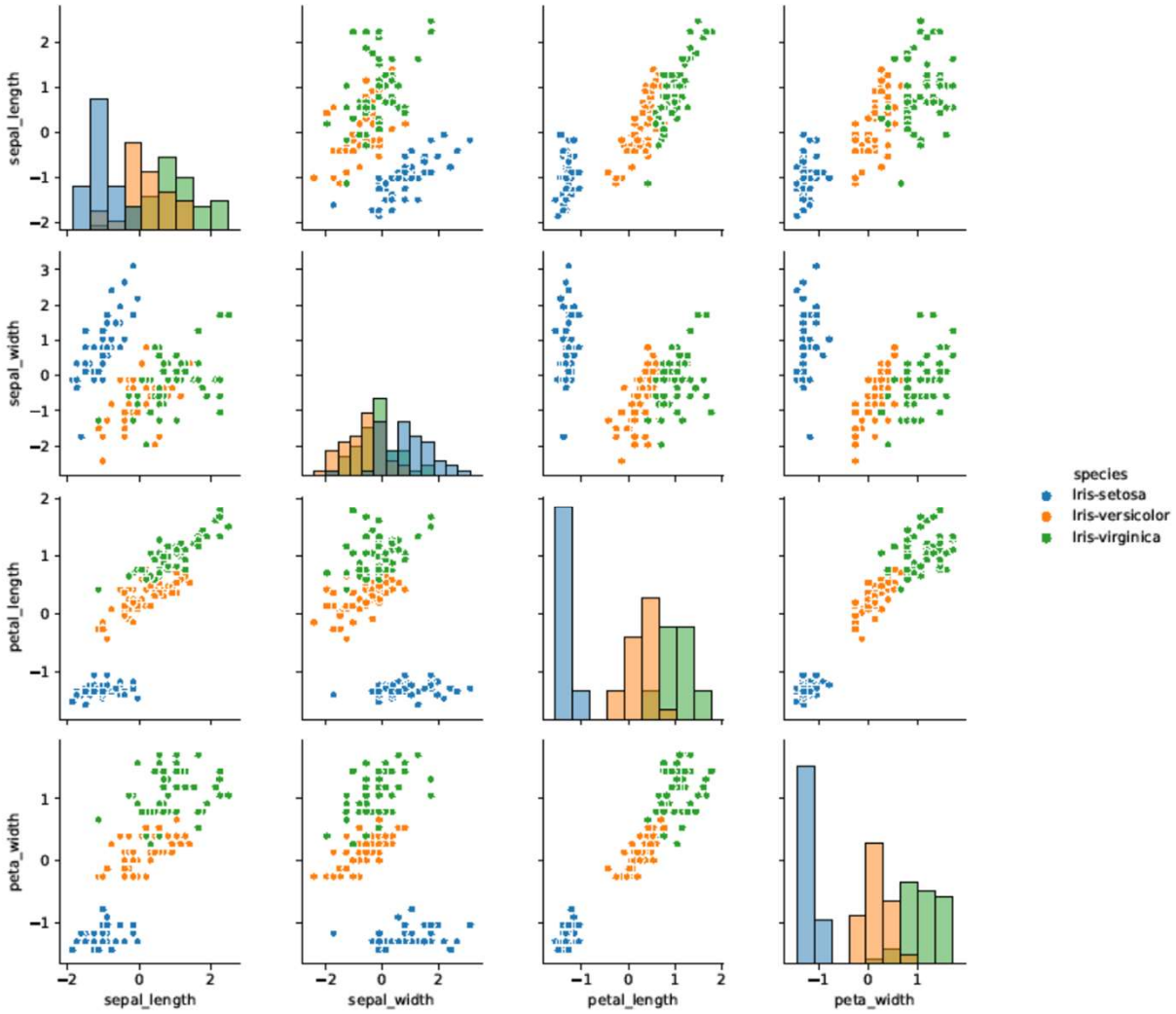
Iris


SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Species
Iris-setosa
Iris-versicolor
Iris-virginica

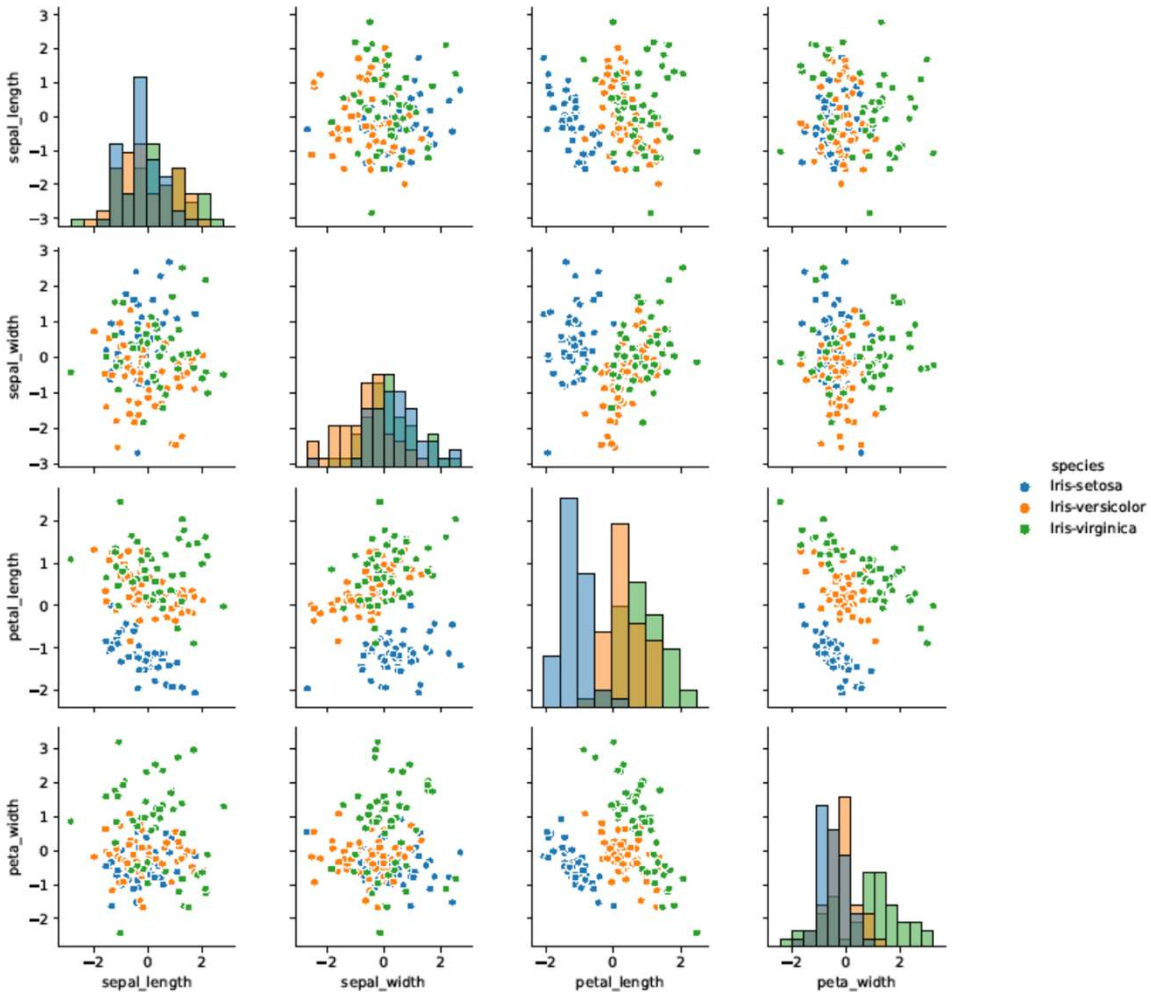
Iris centrado



Iris: Estandarización univariada



Iris: Estandarización multivariada



Distancia

- Un procedimiento alternativo para estudiar la variabilidad de las observaciones es el concepto de distancia entre puntos.

- En el caso escalar $\sqrt{(x_i - \bar{x})^2}$

- Para una variable vectorial, cada dato es un punto en $\mathbb{R}^p$. La distancia Euclídea se define

$$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2} = \left[(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) \right]^{1/2}$$

- La distancia Euclídea depende de las unidades de las variables

Distancia

- Ejemplo: Consideramos 3 individuos donde medimos su altura y su peso:

$$d^2_{A,B} = (1.80 - 1.70)^2 + (80 - 72)^2 = 64.01, \qquad d^2_{A,C} = 1.225$$

	Altura (m)	Peso (kg)
A	1.80	80
B	1.70	72
C	1.65	81

- Si cambiamos la altura a cm: $d^2_{A,B} = 164$, $d^2_{A,C} = 226$.

- Una manera para evitar el problema de las unidades es dividir cada variable por un termino que elimina el efecto de la unidad

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)^T M^{-1} (\mathbf{x}_i - \mathbf{x}_j)\right]^{1/2}$$

- Para $M = D$ la matriz diagonal que contiene las varianzas: $d_{ij} = \left(\sum_{k=1}^{p} (\frac{x_{ik} - x_{jk}}{s_k})^2\right)^{1/2}$.

- Si $s_1 = 10$ cm, $s_2 = 10$ kg: $d^2_{A,B} = 1.64$, $d^2_{A,C} = 2.26$.

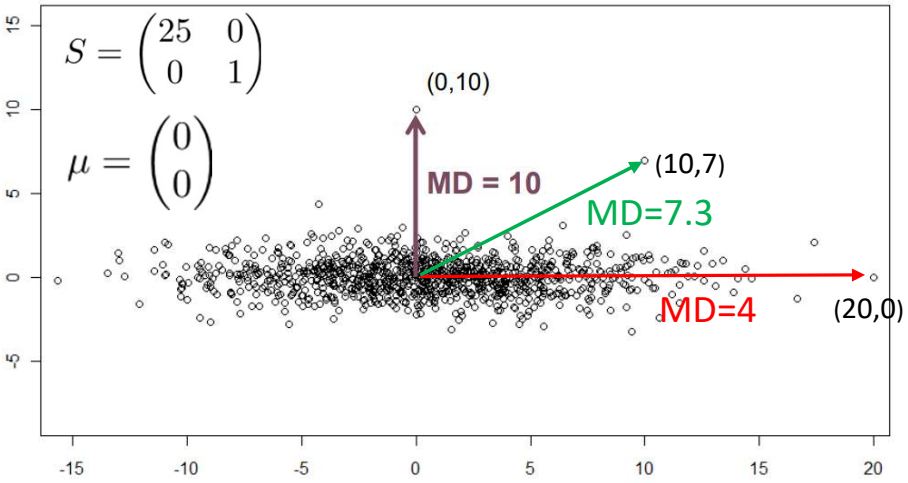
- M debe ser una matriz no singular y definida positiva para que $d_{ij} \geq 0$.

Distancia de Mahalanobis

- La distancia de Mahalanobis (MD), introducida por Mahalanobis en 1936, se define para $M = S$ donde S es la matriz de covarianza:

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)^T S^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right]^{1/2}$$

- Además de las unidades, la MD toma en cuenta la correlación entre las variables.



Distancia de Mahalanobis

Ejemplo: Consideramos el set de datos de peso y altura con la matriz de covarianza

$$S = \begin{pmatrix} s_1^2 & r s_1 s_2 \\ r s_1 s_2 & s_2^2 \end{pmatrix}$$

donde $s_1 = 10$ cm, $s_2 = 10$ kg, $r = 0.7$.

MD entre dos puntos (x_1, y_1) y (x_2, y_2) se puede escribir como

$$d_M = \left[(\mathbf{x}_i - \mathbf{x}_j)^T S^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right]^{1/2}$$

$$d_M^2 = \frac{1}{1 - r^2} \left[\frac{(x_1 - x_2)^2}{s_1^2} + \frac{(y_1 - y_2)^2}{s_2^2} - 2r \frac{(x_1 - x_2)(y_1 - y_2)}{s_1 s_2} \right]$$

$$d_M^2(A, B) = 1.02, d_M^2(A, C) = 4.84$$