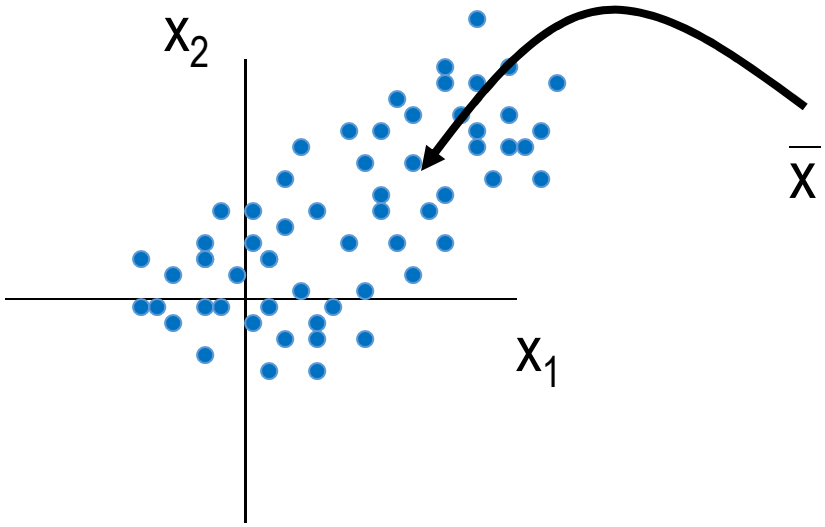
ANALISIS EXPLORATORIO DE DATOS MULTIVARIADOS

Métodos Cuantitativos

El Vector de Medias

$$\underset{(px1)}{\bar{\mathbf{x}}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$



$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i} \mathbf{x}_i = \frac{1}{n}\mathbf{X}^t\mathbf{1}$$

$$\bar{x}_i = \frac{1}{n}\sum_{r} x_{ri} = \frac{x_{.i}}{n}$$

Matriz Varianza-Covarianza

$$s_{ij} = \frac{1}{n}\sum_{i} x_{ir}\, x_{rj} \; - \; \bar{x}_i \; \bar{x}_j$$

$$S = \frac{1}{n}\sum_{r} (x_r - \bar{X})(x_r - \bar{X})^t$$

$$\boxed{S = \frac{1}{n} X^t H X}$$

"H" se conoce como matriz central

$$H = I - \frac{1}{n} 11^t$$

Resultado

Mostrar que S es semidefinida positiva y H es simétrica e idempotente

La matriz $\underset{(nxn)}{M} = \sum_{i} z_r \, z_r^t = Z^t Z$ se conoce como matriz suma

de cuadrados y productos cruzados y se define como:

$$Z^t Z = \begin{bmatrix} z_1 & z_2 & \dots & z_n \end{bmatrix} \begin{pmatrix} z^t_1 \\ z^t_2 \\ . \\ . \\ z^t_n \end{pmatrix} = z_1 z^t_1 + z_2 z^t_2 + \dots + z_n z^t_n$$

Otras Medidas de Dispersión

Varianza Generalizada

$|\mathbf{S}|$

Estimación Máximo Verosímil

Determinante de la Matriz Varianza-Covarianza Muestral

Variación Total

$\mathbf{Tr\ (S)}$

Traza de la Matriz Varianza-Covarianza

Teorema

La matriz de correlación R de un vector aleatorio "y" con matriz de covarianzas **R** se calcula a partir de **S** mediante la relación:

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

donde D es una matriz diagonal con el i-ésimo elemento de la diagonal igual a σ_{ii}

Teorema

La matriz de correlación R es definida positiva

Distancia

Sean ***P*** y ***Q*** dos puntos que representan medidas ***x*** e ***y*** respecto a dos objetos. Una función real valorada ***d(P,Q)*** es una función distancia si tiene las siguientes propiedades:

I) ***Simetría*** $d(P,Q) = d(Q,P)$

II) ***No negatividad*** $d(P,Q) > 0$, si $P \neq Q$

III) ***Identidad*** $P = Q \Rightarrow d(P,Q) = 0$

Métrica

Una distancia es una MÉTRICA si cumple:

IV) ***Definición*** $d(P,Q) = 0 \Rightarrow P = Q$

V) ***Desigualdad Triangular*** $d(P,R) + d(R,Q) \geq d(P,Q)$

Ultra Métrica

Una distancia es una ULTRAMÉTRICA si cumple:

VI) $d(P,Q) \leq \text{máx}\{ d(P,X) , d(X,Q) \}$

Distancias para Datos Cuantitativos

a) Distancia Euclidea
b) Distancia Estadística
c) Distancia de Mahalanobis

a) Distancia Euclidea

La distancia más corta entre dos puntos $P=(x_1, x_2, ..., x_n)$ y $Q=(y_1, y_2, ..., y_n)$ está definido por:

$$d^2(P,Q) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2$$

Dada X (nxp) una matriz de datos con filas $x'_1, x'_2, \ldots, x'_n$, entonces, la distancia Euclidea entre los puntos (objetos) x'_i y x'_j es d_{ij}, donde:

$$d^2ij = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = || x_i - x_j ||^2$$

Propiedades Adicionales

a) **Es semi definida positiva**

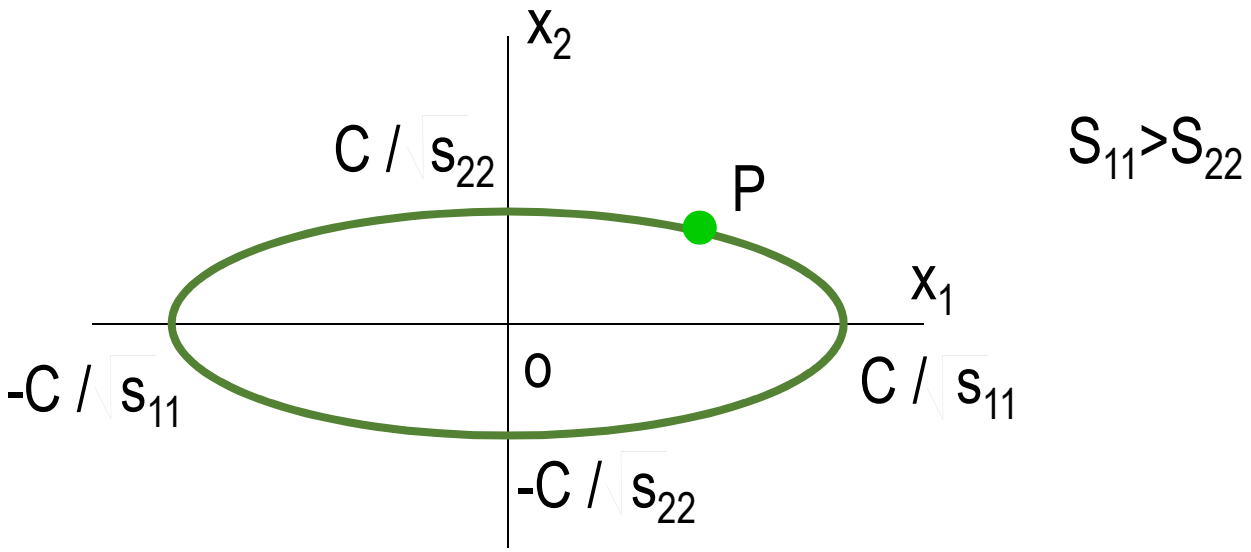
b) **Es invariante frente a transformaciones ortogonales en las x**

c) **Cumple la ley de cosenos**

b) Distancia Estadística

Es un concepto de distancia que además de incluir la variabilidad también incorpora la presencia de correlación

La forma de equilibrar las ponderaciones de acuerdo a la variabilidad consiste en dividir cada coordenada por la desviación estándar, así se obtiene las coordenadas estandarizadas.



Si hacemos:

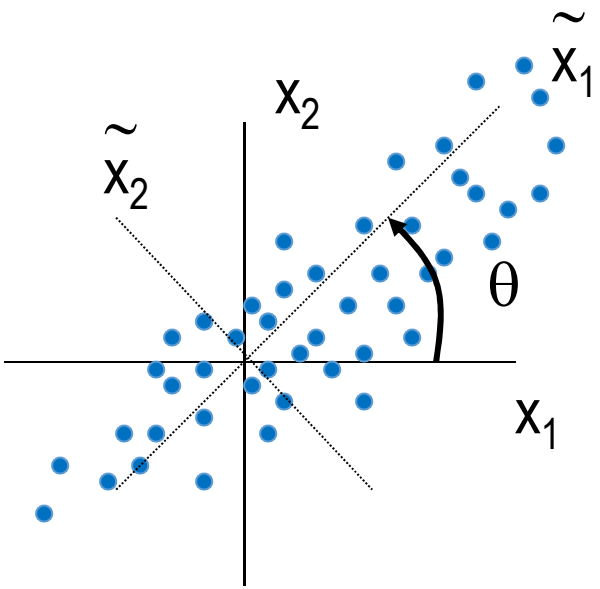
$$d(O,P) = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$$

La distancias estadística de P a Q se define como:

$$d^2(P,Q) = \frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \ldots + \frac{(x_n - y_n)^2}{s_{nn}}$$

Si $s_{11}=s_{22}=\ldots=s_{nn}$ se puede utilizar la fórmula de la distancia Euclidea

Si x_1 no varía independientemente de x_2



La variabilidad de x_1 es diferente a la de x_2 y, además, ambas están correlacionados

$$d(O,P) = \sqrt{\frac{\tilde{x}_1^2}{s_{11}} + \frac{\tilde{x}_2^2}{s_{22}}}$$

donde;

$$\tilde{X}_1 = x_1 \cos\theta + x_2 \operatorname{sen}\theta$$
$$\tilde{X}_2 = -x_1 \operatorname{sen}\theta + x_2 \cos\theta$$

La **distancia estadística** más corta entre dos puntos $P=(x_1, x_2, ..., x_n)$ y $Q=(y_1, y_2, ..., y_n)$ está definido por:

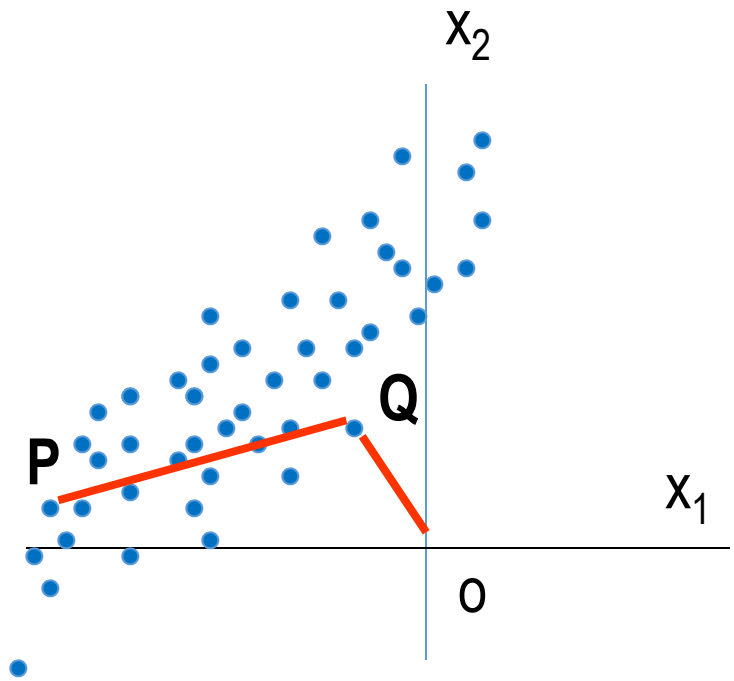
$$d^2(P,Q) = \sqrt{a_{11}(x_1 - y_1)^2 + a_{22}(x_2 - y_2)^2 + ... + a_{pp}(x_n - y_n)^2}$$

$$\sqrt{a_{12}(x_1 - y_1)(x_2 - y_2) + a_{13}(x_1 - y_1)(x_3 - y_3) + ... + a_{p-1,p}(x_{p-1} - y_{p-1})(x_p - y_p)}$$

Los coeficientes de la expresión anterior pueden representarse mediante un arreglo matricial, así:

$$\begin{pmatrix} a_{11} & a_{12} & ... & a_{1p} \\ a_{12} & a_{22} & ... & a_{2p} \\ . & & & \\ . & & & \\ a_{1p} & a_{2p} & ... & a_{pp} \end{pmatrix}$$

Distancia euclidea y distancia estadística



d(P,Q) > d(Q,O) **Distancia euclidea**

d(P,Q) < d(Q,O) **Distancia estadística**

b) Distancia de Mahalanobis

La **distancia al cuadrado de Mahalanobis** entre los puntos x_i y x_j se define como:

$$D^2_{ij} = (x_i - x_j)' S^{-1} (x_i - x_j)$$

Matrices Definidas Positivas

Dado A una matriz simétrica kxk, entonces, A tiene k pares de vectores y valores característicos:

$$\lambda_1, \mathbf{e}_1 \quad \lambda_2, \mathbf{e}_2 \quad \ldots \quad \lambda_k, \mathbf{e}_k$$

tal que,

$$\mathbf{e}_i^t \mathbf{e}_j = \begin{cases} 1 & \text{Si } i=j \\ 0 & \text{si } i \neq j \end{cases}$$

Descomposición Espectral

La descomposición espectral de una matriz simétrica kxk está dado por:

$$A = \lambda_1 e_1 e_1^t + \lambda_2 e_2 e_2^t + \ldots + \lambda_k e_k e_k^t$$

RESULTADOS
(continuación)

Si A es una matriz kxk definida positiva con descomposición espectral:

$$\mathbf{A} = \sum \lambda_1 \, \mathbf{e}_1 \, \mathbf{e}_1^{t} = \mathbf{P}\Lambda\mathbf{P}^{t}$$

donde $\mathbf{PP}^{t} = \mathbf{P}^{t}\mathbf{P} = \mathbf{I}$

Probar que la matriz raíz cuadrada $\mathbf{A}^{1/2} = \mathbf{P}\Lambda^{1/2}\mathbf{P}^{t}$ tiene las siguientes propiedades :

a. $A^{1/2}$ es simétrica
b. $A^{1/2}A^{1/2} = A$
c. $(A^{1/2})^{-1}$ existe
d. $A^{1/2}\, A^{-1/2} = A^{-1/2}A^{1/2} = I$