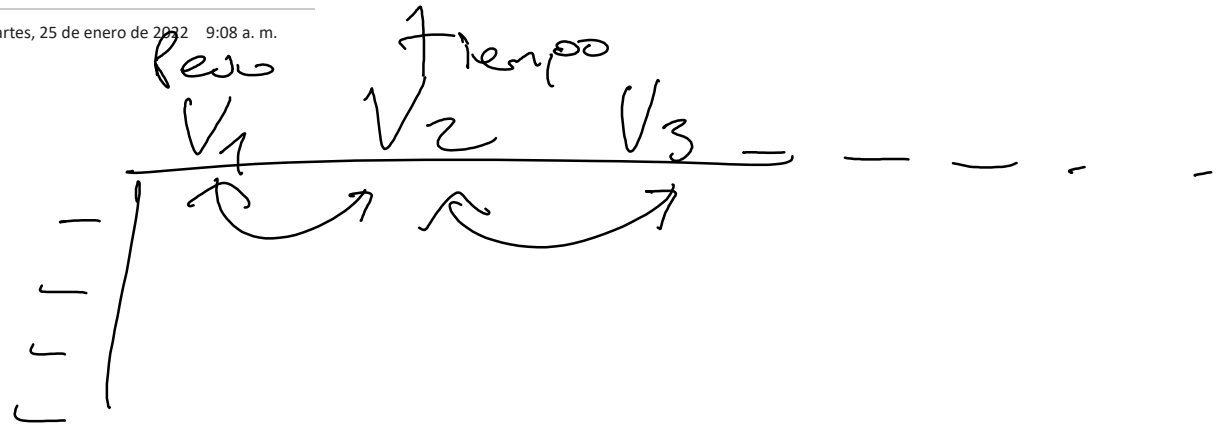
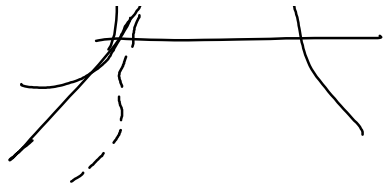
# Introducción

martes, 25 de enero de 2022 9:08 a. m.



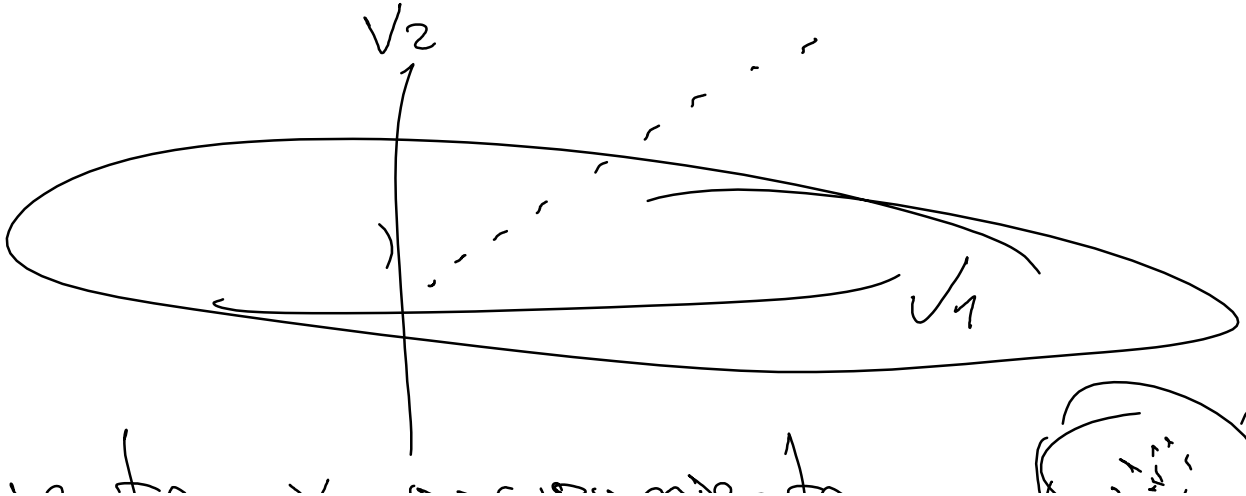

- La mayoría de problemas involucran múltiple variables
- Vamos a extender los métodos vistos previamente (PyE1, PyE2) y veremos otros nuevos involucrando algebra matricial, cálculo de varias variables.



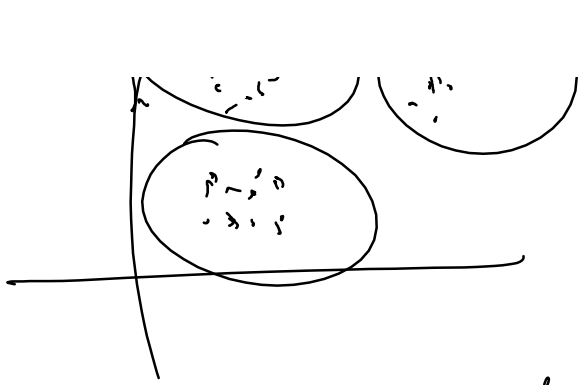
- Vamos a utilizar principalmente la distr. Normal multivariada.
- Vamos a trabajar con R y R Markdown

Objetivos

* Reducción de datos o simplificación estructural

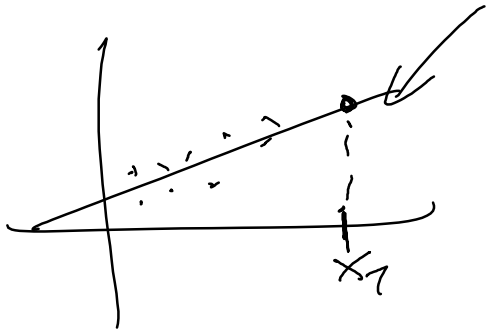


* Agrupamiento y agrupamiento



* Investigación acerca de la dependencia entre variables

* Predicción




$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

* Construcción de pruebas de hipótesis

Usaremos la notación matricial

| | Var 1 | Var 2 | Var 3 --- Var p |
|---|---|---|---|
| Item 1 | $X_{11}$ | $X_{12}$ | $X_{1p}$ |

Item 2 $x_{21}$

Item 3

⋮

Item n $x_{np}$

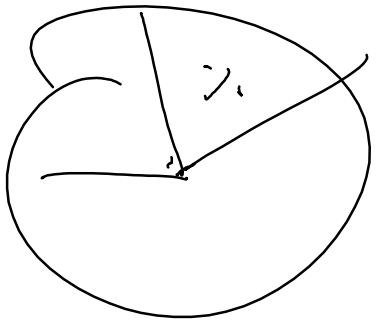
$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ & & x_{np} \end{bmatrix}$$

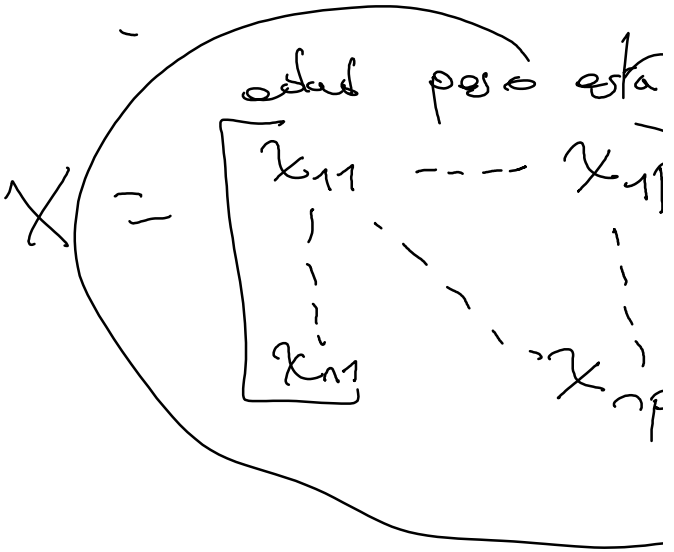
Generalmente las conjuntos de datos son muy grandes y es complejo extraer info.

| | | |
|---|---|---|
| 100 | 23 | |
| 4 | 25 | |
| Rio | 22 | |

Podemos obtener info con medidas que resumen los datos

media, mediana y varianza ------






Media muestral

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad k = 1, \dots p$$

Varianza muestral

$$S_k^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2 \qquad k=1,\dots,p.$$

$$\underline{S_k^2 = S_{kk}} \qquad \underline{\underline{S}} = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{12} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1p} & S_{2p} & \cdots & S_{pp} \end{pmatrix}$$

$S_1^2$

desv. estándar $\sqrt{S_{kk}} = S_k$

Supongamos que tenemos $n$ observaciones sobre $p$ variables.

$$\begin{matrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{matrix}$$

$$\begin{matrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{matrix}$$

$$S = \begin{bmatrix} & \\ & \\ & \end{bmatrix}$$

$$S_{kk} = \frac{1}{n}\sum (x_{jk} - \bar{x}_k)^2$$

$$\underline{S_{12}} = \frac{1}{n}\sum (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2)$$

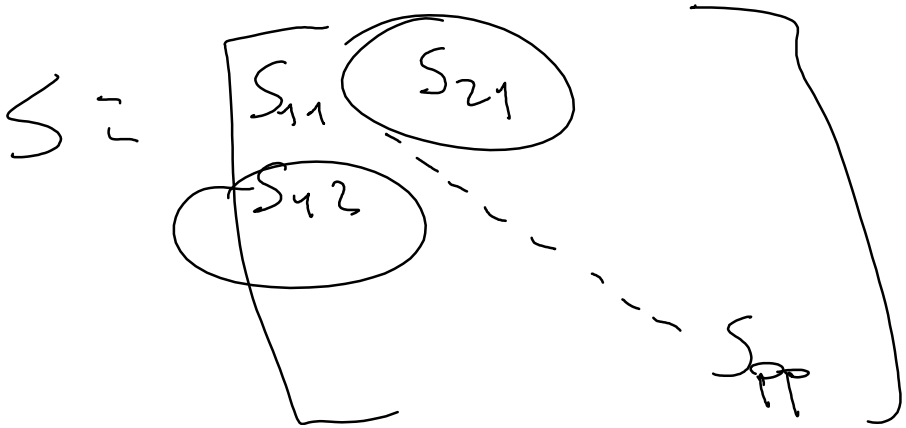
Mide la asociación lineal entre las dos variables.

* Es el promedio del producto entre las desviaciones de sus respectivos medios.

* Si se [illegible] valores grandes y pequeños valores también se presenta de forma conjunta $S_{12}$ será positiva

* Si se presenta valores grandes en una var y pequeños en otra $S_{12}$ será negativa

* Si no hay asociación entre las dos varia $S_{12}$ será aprox 0.




$S_{12} = S_{21}$

$$S_{iK} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jK} - \bar{x}_K)$$

* Cuando $i = K$ es la varianza muestral
* La matriz es simétrica.

Normalizamos para no depender de las unidades.

Coeficiente de correlación muestral

$$r_{iK} = \frac{S_{iK}}{\sqrt{S_{ii}} \cdot \sqrt{S_{KK}}} = \frac{\sum (x_{ji} - \bar{x}_i)(x_{jK} - \bar{x}_K)}{\sqrt{\sum (x_{ji} - \bar{x}_i)^2} \sqrt{\sum (x_{jK} - \bar{x}_K)^2}}$$

* Es simétrica $r_{iK} = r_{Ki}$ para todo $i \neq K$

- Es una medida de asociación lineal entre dos variables, que no depende de las unidades.

$r_{11} = 1$

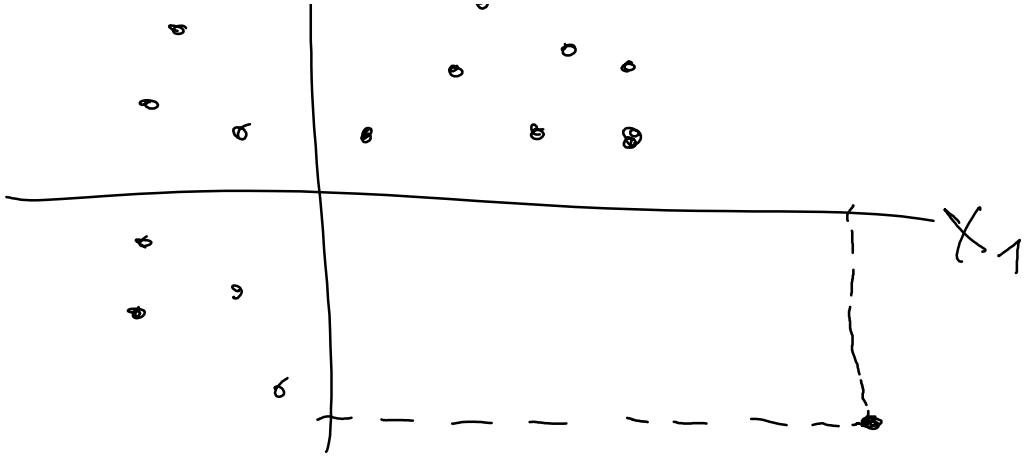
- $r$ está entre $-1$ y $1$

$r = 0$ No hay asociación.

$r = -1$ correlación negativa → una aumenta y la otra disminuye (en promedio)

$r = 1$ correlación positiva → las dos aumentan o disminuyen (en promedio)

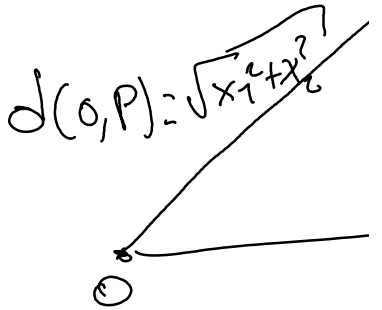
$X_2$




# Distancia Euclidiana

distancia entre un punto $P = (x_1, x_2 \dots, x_p)$ al origen $O = (0, 0, \dots, 0)$

$$d(O,P) = \sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$$



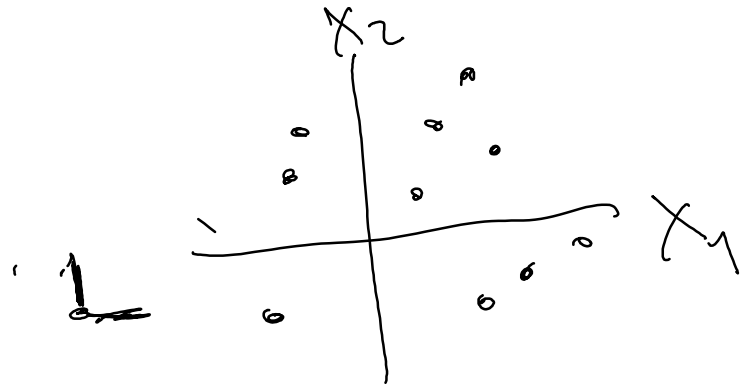

Todos los puntos que estén a una distancia cuadrada $C^2$ respecto al origen se encuentran

Sobre una hiperesfera

Distancia entre un punto P y un punto Q

$P = (x_1, x_2 \cdots, x_p)$ $Q = (y_1, y_2 \cdots, y_p)$

$$d(P,Q) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 \cdots + (x_p-y_p}$$




Para la mayoría de aplicaciones estadísticas la distancia Euclidiana NO es apropiada

Porque cada componente contribuye igualmente al cálculo de la distancia

metros
23
82

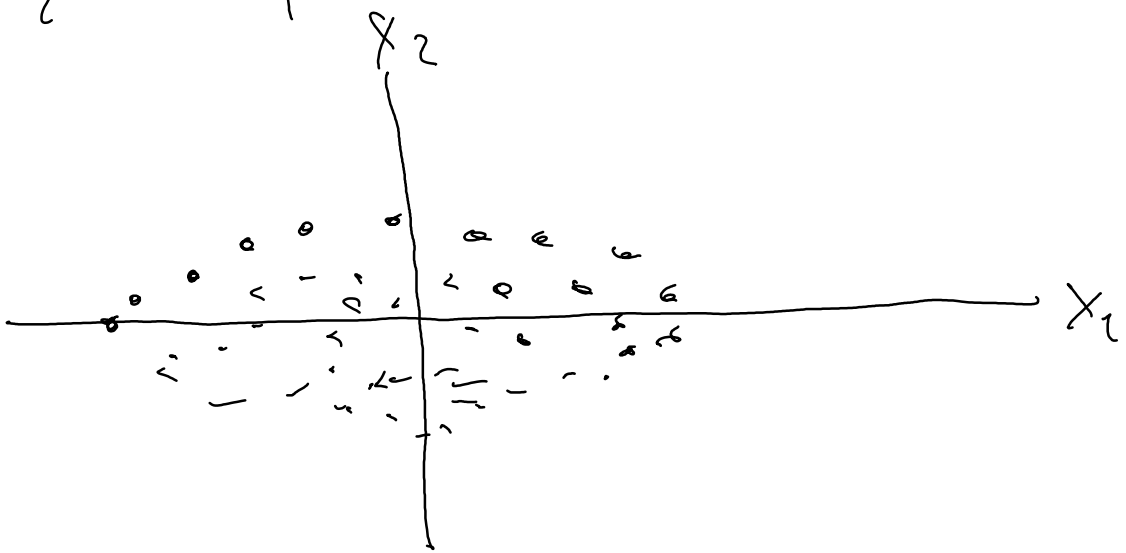
cm
1.12

Cuando las coordenadas representan medidas que pue
presentar variaciones absolutas de dif. magnitudes
es conveniente ponderar con mayor peso a las
variables que varían menos y con menor peso a
las que varían más

El objetivo es desarrollar una distancia estadística

* El objetivo [illegible]

que tenga en cuenta tanto las diferencias en variacion y la presencia de correlación.




Suposiciones

- Hay n pares de medidas en las variables $X_1$ y $X_2$
- Tienen media cero y varianzas independientes

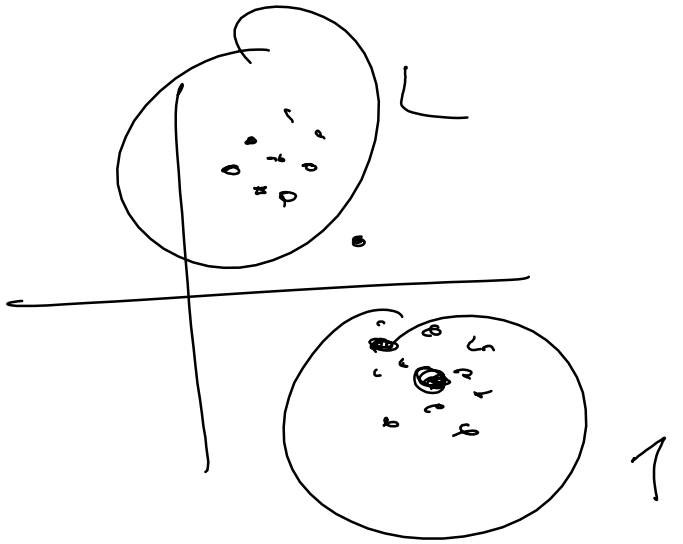
los puntos en $X_1$ tienen más variabilidad que los de $X_2$

Por ésto, $X_2$ tendrá más ponderación.

$$X_1^* = X_1 / \sqrt{S_{11}} \qquad X_2^* = X_2 / \sqrt{S_{22}}$$

$P = (X_1, X_2)$

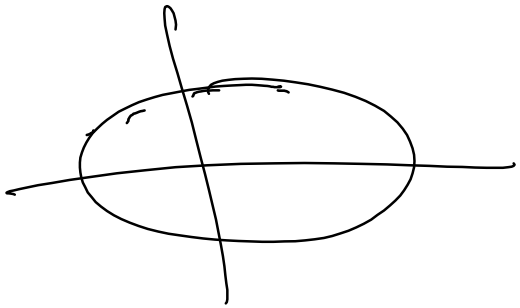
$$d(O, P) = \sqrt{(X_1^*)^2 + (X_2^*)^2} = \sqrt{\frac{X_1^2}{S_{11}} + \frac{X_2^2}{S_{22}}}$$



Si la variabilidad en ambos [illegible]
igual, es mejor utilizar la distancia Euclidiana

Todos los puntos con coordenadas $(x_1, x_2)$ con distancia cuadrada $C^2$ satisfacen: $\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = C^2$

Se encuentran en una elipse centrada en el origen



Distancia estadística entre dos puntos.

$$d(O, P) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}}$$

$$d(P,Q) = \sqrt{\frac{(x_1-y_1)^2}{s_{11}} + \frac{(x_2-y_2)^2}{s_{22}} + \cdots}$$

Si $P = (x_1, x_2 \cdots x_p)$ y $Q = (y_1, y_2 \cdots y_p)$

$$d(P,Q) = \sqrt{\frac{(x_1-y_1)^2}{s_{11}} + \frac{(x_2-y_2)^2}{s_{22}} + \cdots + \frac{(x_p-y_p)^2}{s_{pp}}}$$

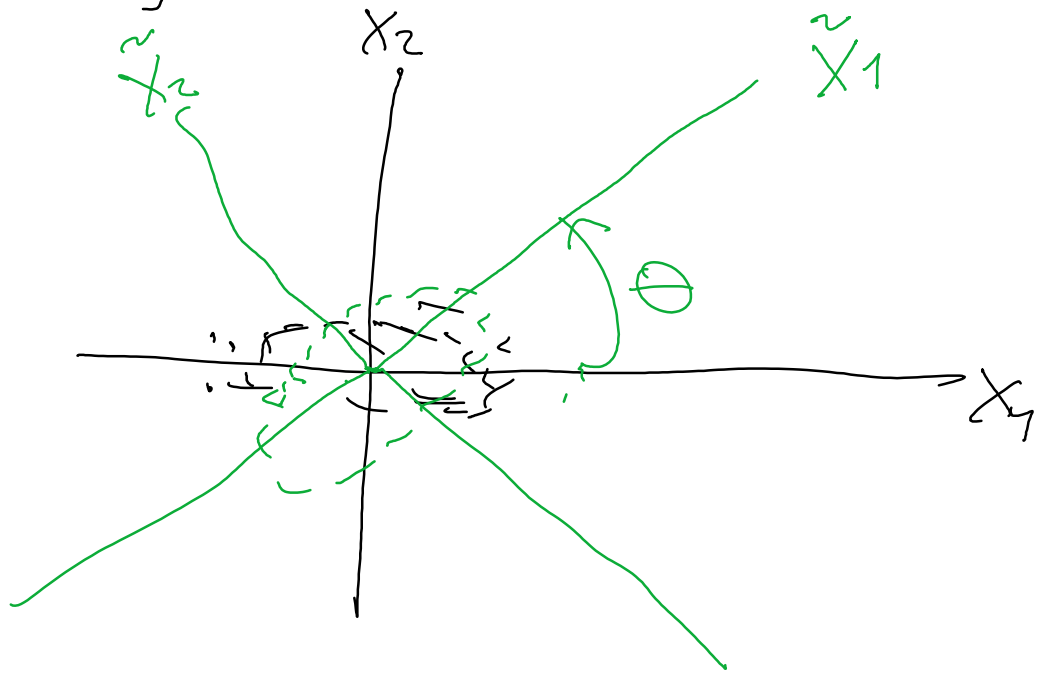
$y_1 = y_2 = \cdots = 0$

$$d(0,P)$$

¿Siempre es adecuada esa distancia estadística?

Esta distancia estadística asume independencia

en las variables




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

$$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \operatorname{Sen}(\theta)$$
$$\tilde{x}_2 = -x_1 \operatorname{Sen}(\theta) + x_2 \cos(\theta)$$

$$\sqrt{\qquad\qquad x_2^2}$$

$\rightarrow$ $Q = \sum a_{ij} x_i x_j$

$a_{ij}$ son funciones de $\theta$.

$P = (x_1, x_2)$ $\quad Q = (y_1, y_2)$

$$d(P,Q) = \sqrt{a_{11}(x_1-y_1)^2 + 2a_{12}(x_1-y_1)(x_2-y_2) + a_{22}(x_2-y_2)^2}$$

Con p variable

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

$$+2a_{12}(x_1-y_1)(x_2-y_2)+2a_{13}(x_1-y_1)(x_3-y_3)$$

$$\cdots+2a_{p-1,p}(x_{p-1}-y_{p-1})(x_p-y_p)$$