# GEOMETRÍA MUESTRAL

## Observación Multivariada

Una observación multivariada es una colección de mediciones sobre ’p’ variables medida sobre el mismo objeto o ensayo:

$$X_{nxp} = \begin{array}{cc} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} & \begin{array}{l} \text{1ra observación} \\ \\ \\ \text{n-ésima observación} \end{array} \\ \begin{array}{cccc} y_1 & y_2 & \quad & y_p \end{array} & \end{array}$$

## NOTACIÓN

$X_{nxp}$ : Matriz de datos para una muestra de tamaño 'n' para 'p' variables

$X_{ij}$ : Un dato para la variable 'j' del individuo o unidad muestral 'i'.

$X'_i$ : i-ésima fila.

$X_{(j)}$ : j-ésima columna, correspondiente a la j-ésima variable.

# NOTACIÓN (...continuación)

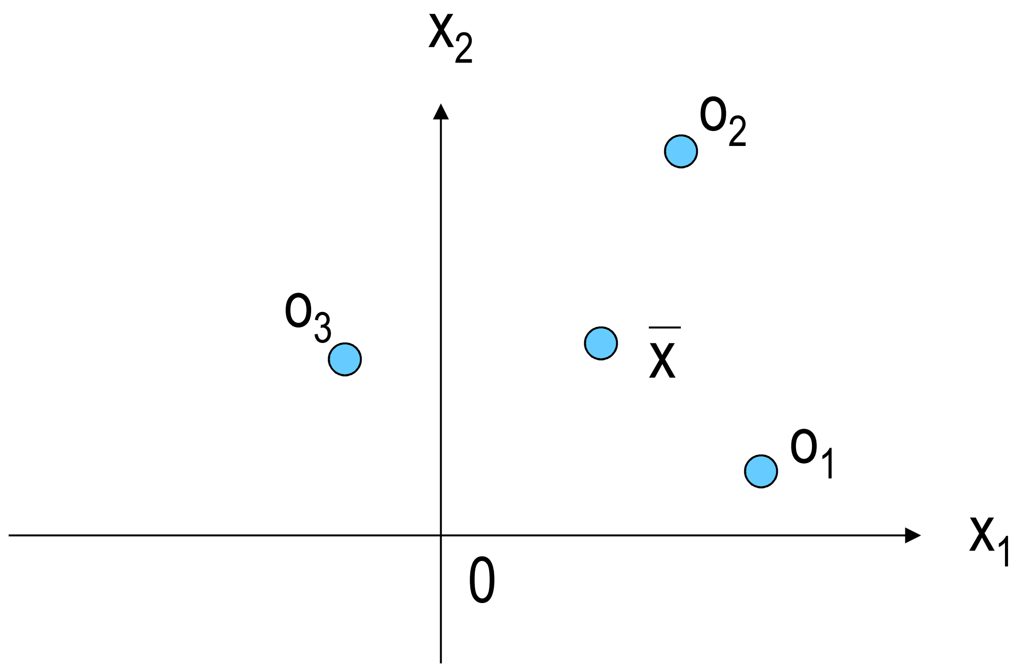
$$X_{nxp} = \begin{pmatrix} X'_1 \\ X'_2 \\ \cdot \\ \cdot \\ \cdot \\ X'_n \end{pmatrix} = \begin{pmatrix} X_{(1)}, & X_{(2)}, \dots, & X_{(p)} \end{pmatrix}$$

donde,

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \cdot \\ \cdot \\ \cdot \\ X_{ip} \end{pmatrix} \quad (i=1, \dots, n) \qquad X_{(j)} = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \cdot \\ \cdot \\ \cdot \\ X_{nj} \end{pmatrix} \quad (j=1, \dots, p)$$

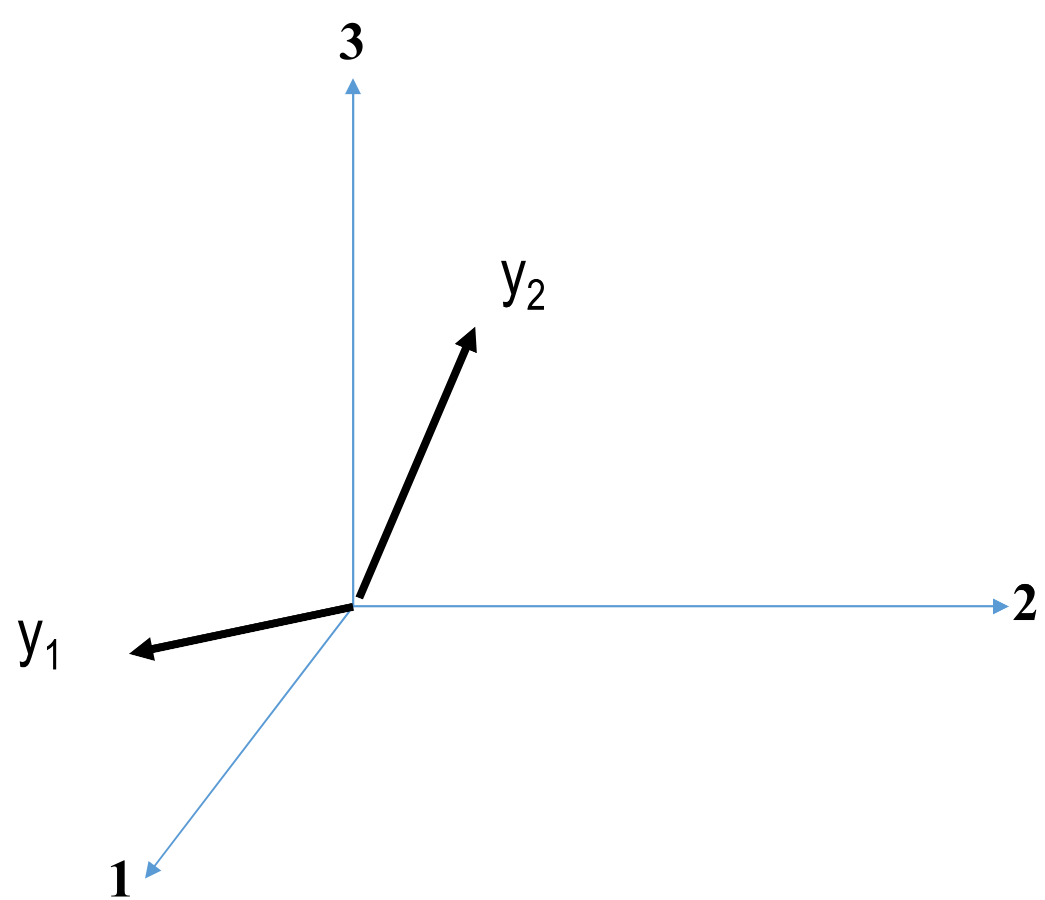
# Caso n=3 y p=2

## Ploteo de puntos (observaciones) para una matriz X



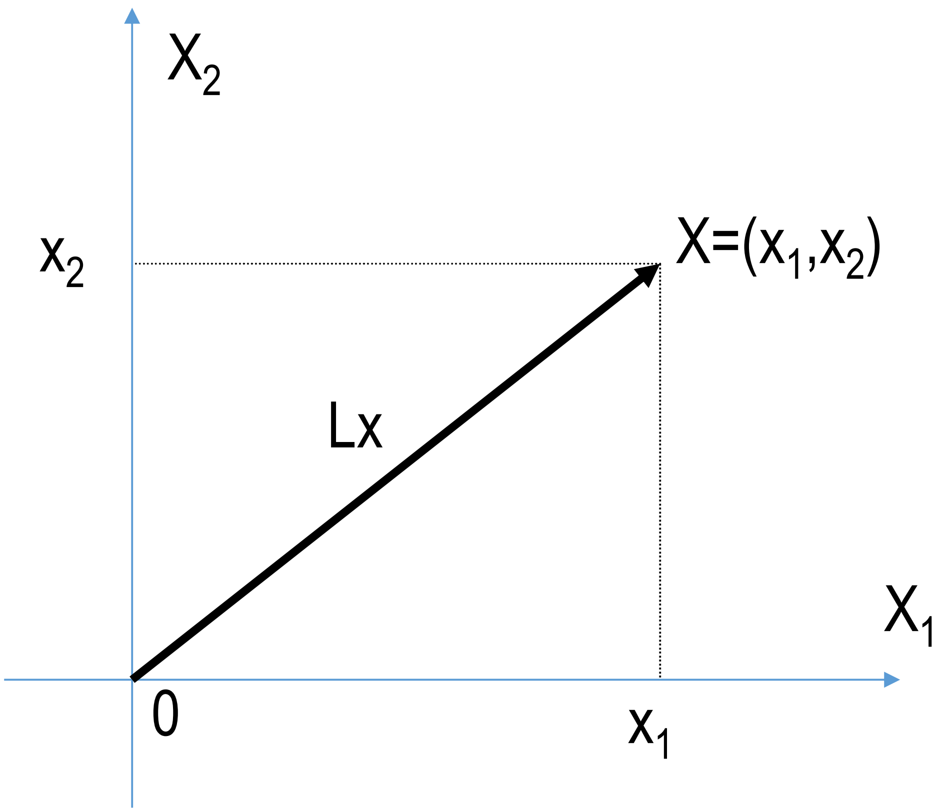
## Caso n=3 y p=2

# Gráfico de variables para la matriz X



# VECTORES ALEATORIOS

# Longitud o Norma de un Vector



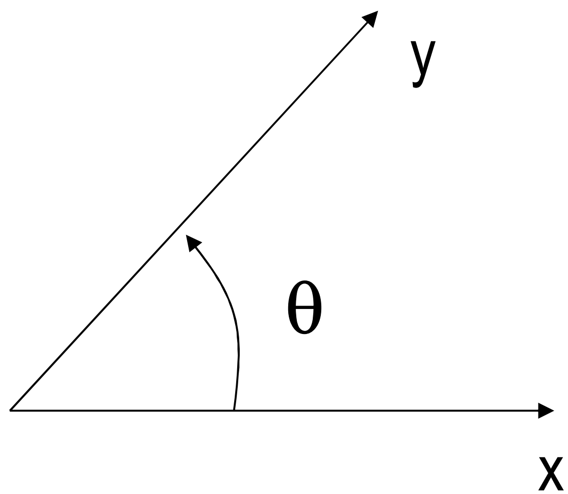

$$L_x = \sqrt{X_1^2 + X_2^2}$$

$$L_x = \sqrt{X^I X}$$

Caso General:

$$X = (x_1, x_2, \ldots, x_p)$$

# Ángulo entre Dos Vectores




$$\text{Cos}\ \theta = \frac{x_1y_1 + x_2y_2 + \ldots + x_py_p}{LxLy}$$

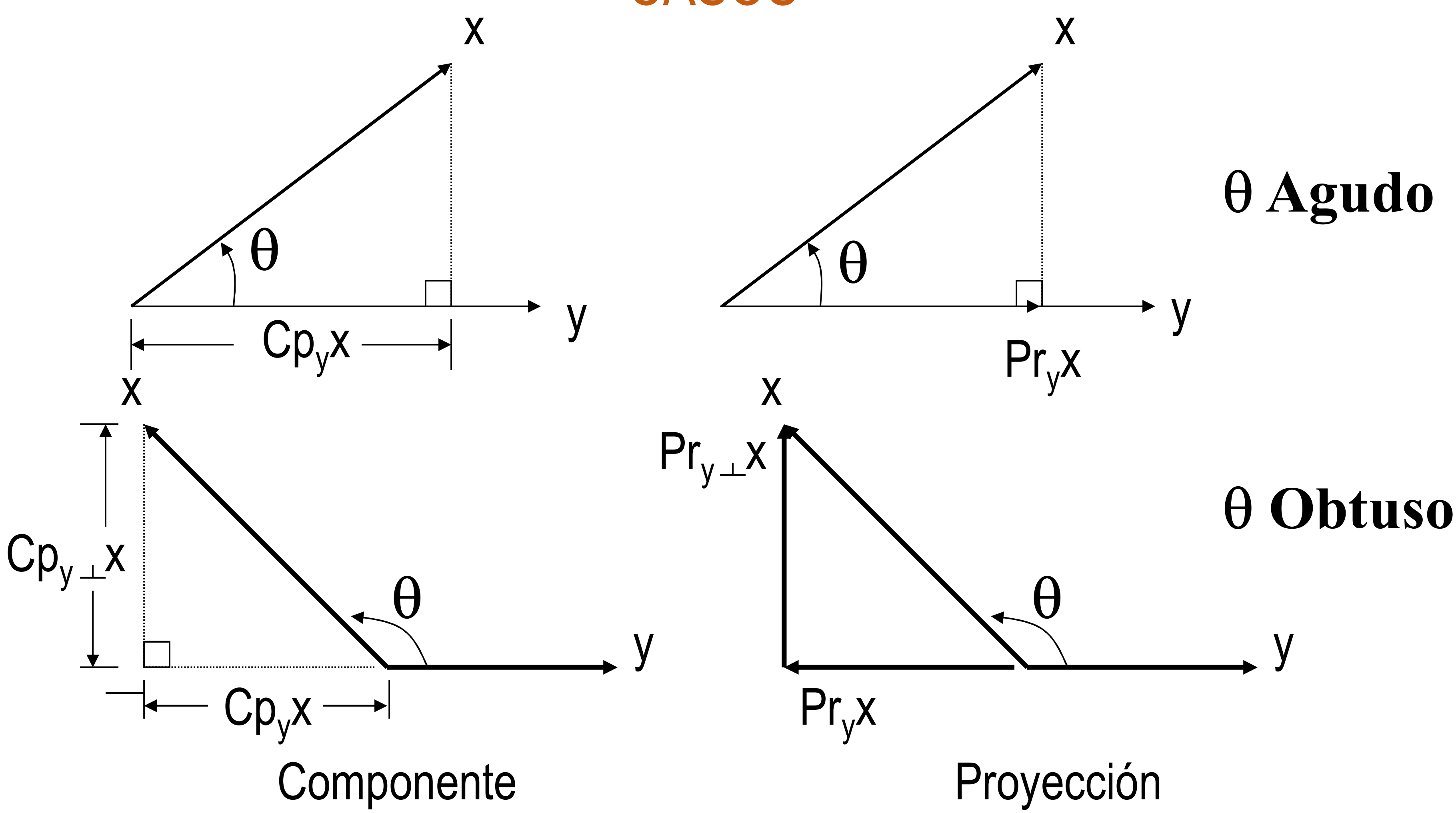
$$\text{Cos}\ \theta = \frac{X^{l}Y}{\sqrt{X^{l}X}\ \sqrt{Y^{l}Y}}$$

# Componente y Proyección de X sobre Y

La Proyección Ortogonal de X sobre Y ($Pr_y x$) es el vector cuya dirección está dada por la dirección y sentido del vector Y, y la longitud por la Componente de X sobre Y ($Cp_y x$)

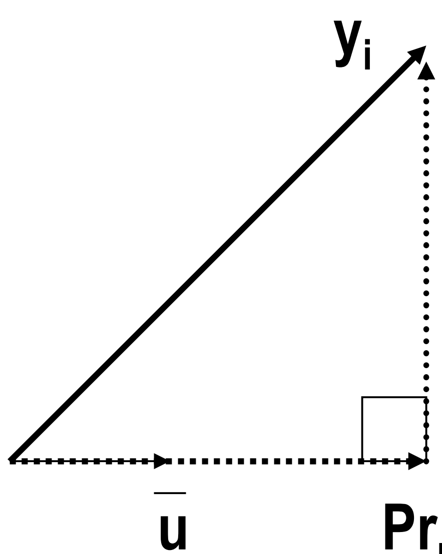
$$Pr_y x = \frac{x^l y}{L^2 y} \cdot y = \frac{x^l y}{Ly} \cdot \frac{y}{Ly} = Cp_y x \cdot \frac{y}{Ly}$$

# CASOS


x
θ
y
$Cp_y x$
x
θ
y
$Pr_y x$
θ Agudo
x
$Cp_{y\perp} x$
θ
y
$Cp_y x$
x
$Pr_{y\perp} x$
θ
y
$Pr_y x$
θ Obtuso
Componente
Proyección

# Media Muestral

*desviación*




$y_i - \bar{x}_i \mathbf{1} = e_i$

$Pr_u y_i = \bar{x}_i \,.\, \mathbf{1}$

*componente de medias*

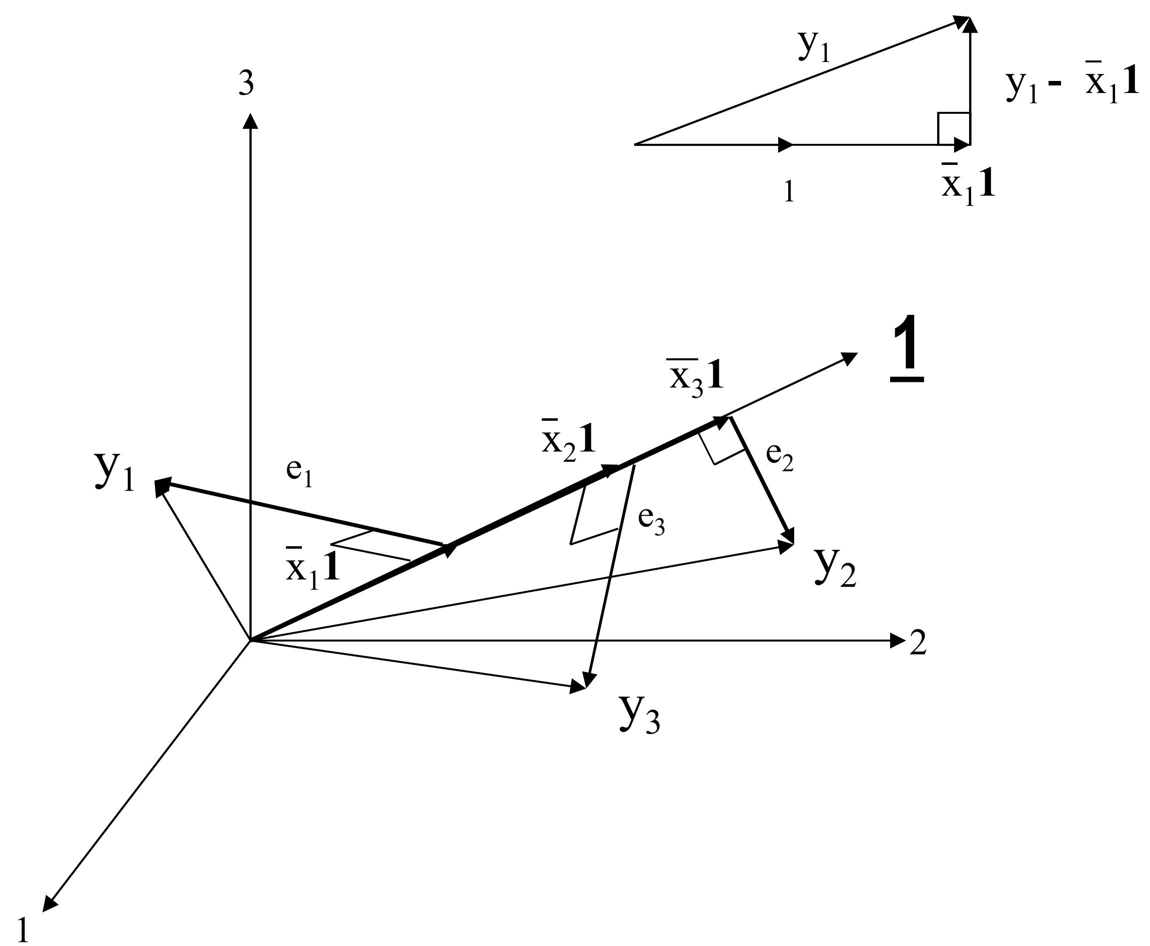
Sea

$$y_i = (x_{i1}, x_{i2}, ..., x_{1n}) \text{ y } \mathbf{1} = (1, 1, ..., 1)_{nx1}$$

$$Pr_u y_i = \frac{y_i' \bar{u}}{|\bar{u}|^2} \,.\, \bar{u} \quad \text{donde} \quad \bar{u} = \frac{1}{\sqrt{n}} \,.\, \mathbf{1}$$

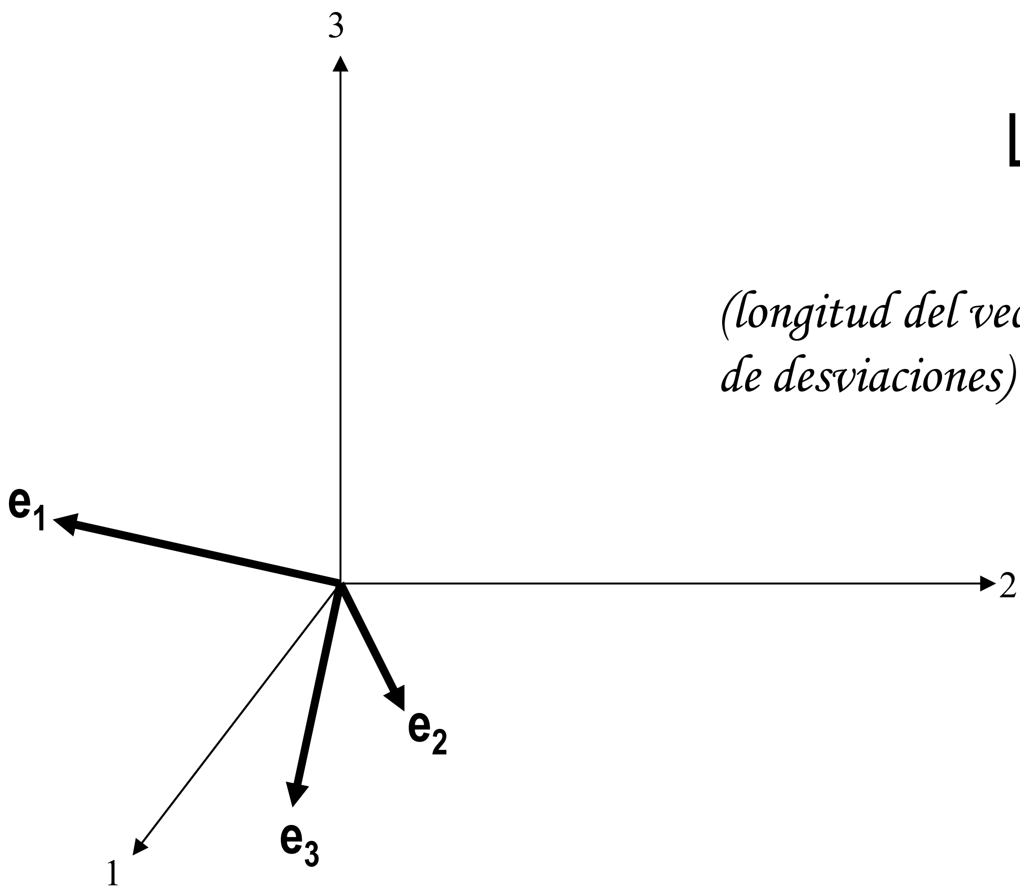
$$Pr_u y_i = y_i' \bar{u} \,.\, \bar{u} = \bar{x}_i \,.\, \mathbf{1}$$

# Caso para p=3 y n=3



# Caso para p=3 y n=3

## Graficando las Desviaciones




$$L^2_{e_i} = e_i' e_i = \sum_j (x_{ij} - \bar{x}_i)^2$$

*(longitud del vector de desviaciones)*$^2$ = *(suma de cuadrados de las desviaciones)*

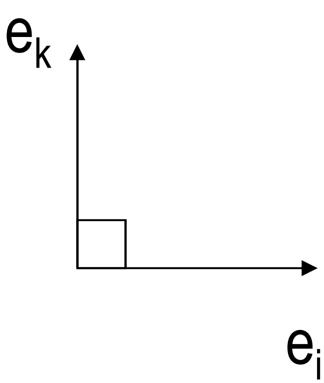
***Longitud más grande del vector de desviaciones representa mayor variabilidad que vectores más pequeños***

# Coeficiente de Correlación

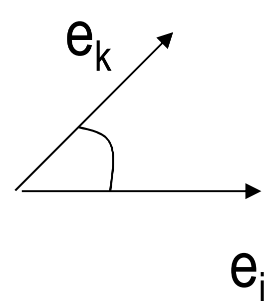
El coseno del ángulo formado por dos vectores de desviaciones es el coeficiente de correlación

$$e_i^{\prime} e_k = L_{ei}\, L_{ek} \cos(\theta_{ik})$$

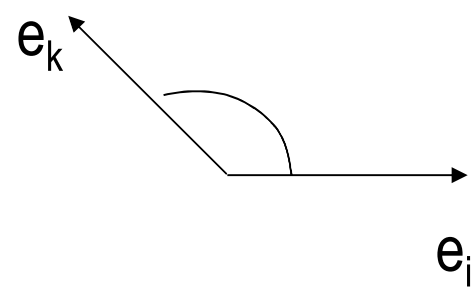
$$\cos(\theta_{ik}) = \frac{e_i^{\prime} e_k}{L_{ei}\, L_{ek}} \qquad \cos(\theta_{ik}) = \frac{S_{ik}}{\sqrt{S_{ii}}\,\sqrt{S_{kk}}}$$




$\cos(\theta_{ik}) = 0$

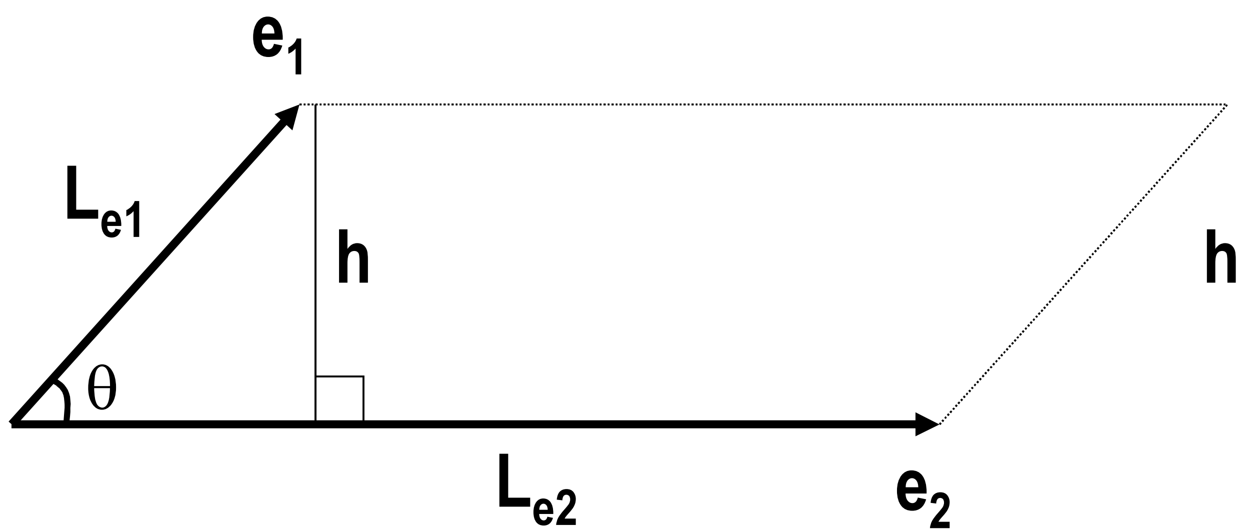



$0 < \cos(\theta_{ik}) < 1$




$\cos(\theta_{ik}) < 0$

# Varianza Generalizada




$$h = Pr_{e_2^{\perp}}\ e_1$$

$$h = L_{e1}\,\mathrm{sen}(\theta)$$

$$\text{Area} = L_{e1}L_{e2}\sqrt{1-\cos^2(\theta)}$$

$$L_{e1} = \sqrt{\sum_{j=1}^{n} (x_{1j} - \bar{x}_1)^2} = \sqrt{(n-1)s_{11}}$$

$$L_{e2} = \sqrt{\sum_{j=1}^{n} (x_{2j} - \bar{x}_2)^2} = \sqrt{(n-1)s_{22}}$$

$$\cos(\theta) = r_{12}$$

Luego,

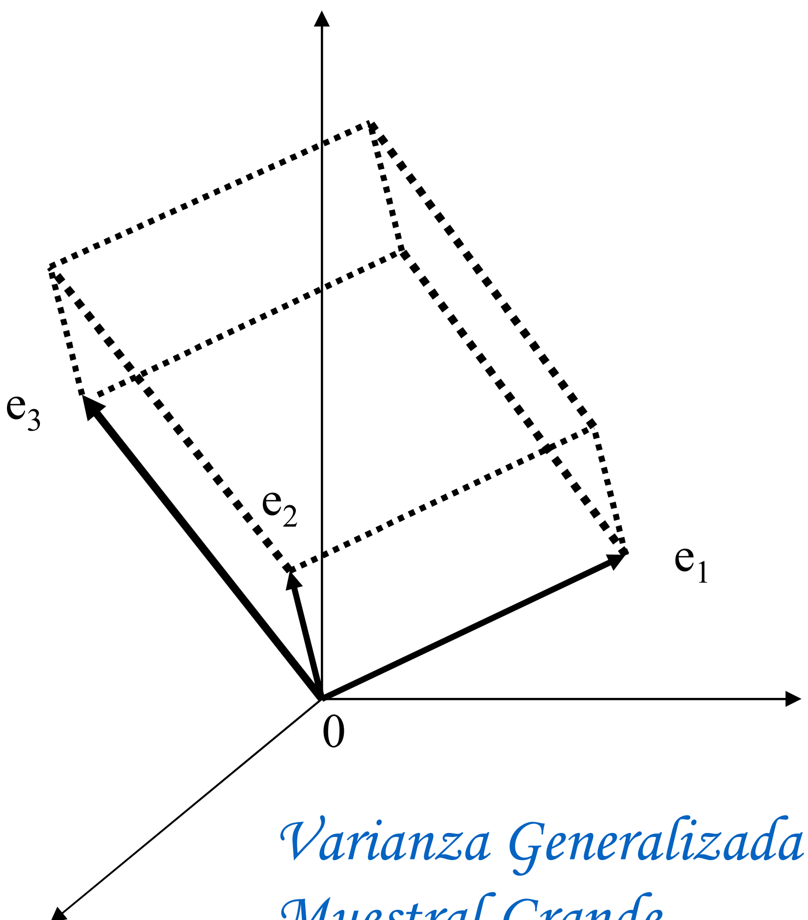
$$Area = (n-1)\sqrt{s_{11}}\sqrt{s_{22}}\sqrt{1 - r^2_{12}}$$

$$\boxed{|S| = (Area)^2 / (n-1)^2}$$

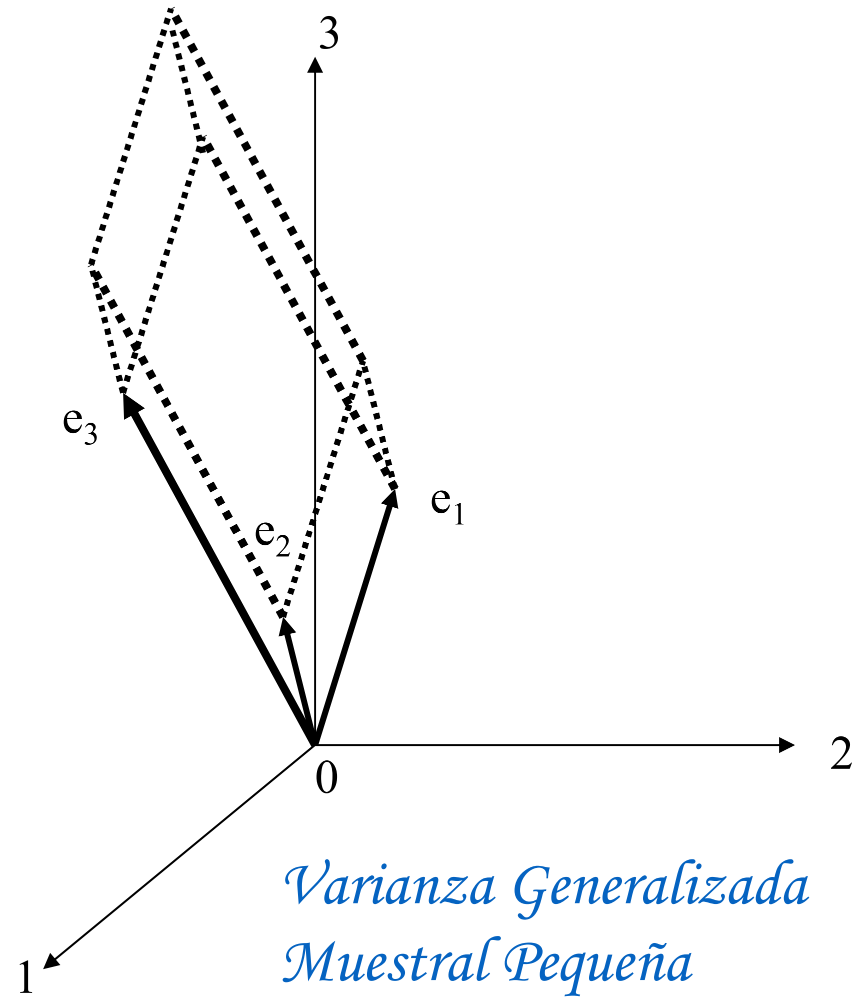
Generalizando para un espacio ***n***-dimensional y ***p*** vectores de desviaciones,

$$|S| = (n-1)^{-p}\,(\text{Volumen})^2$$

# Caso: p=3

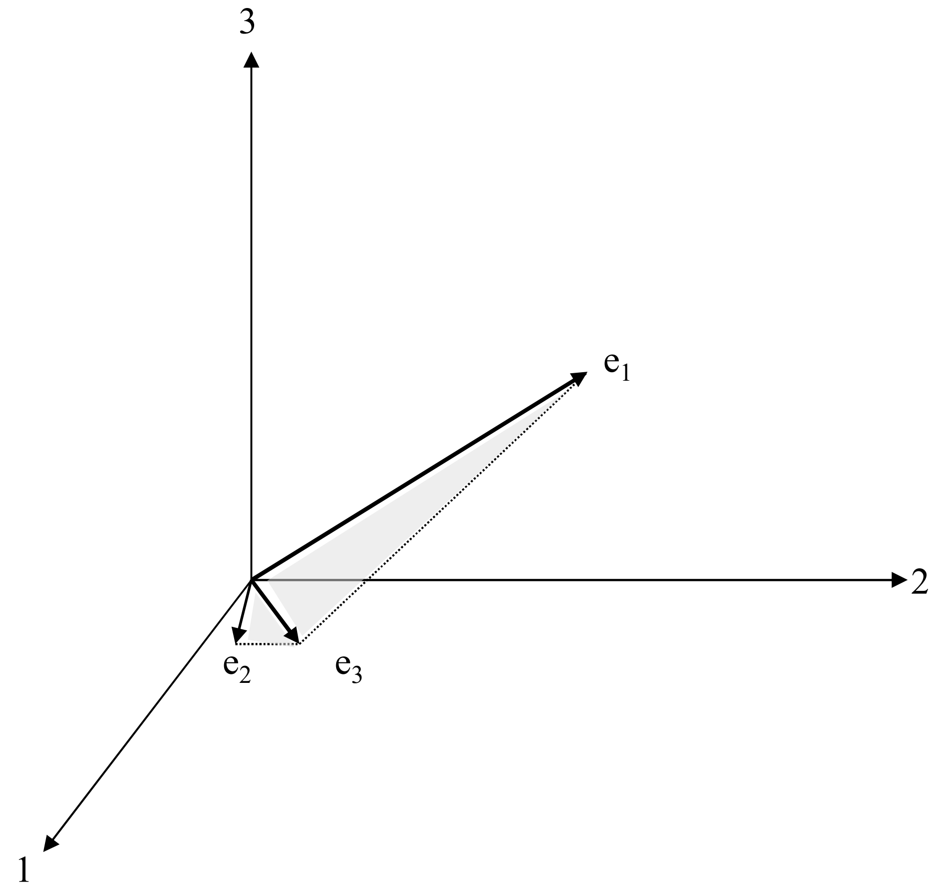



*Varianza Generalizada Muestral Grande*

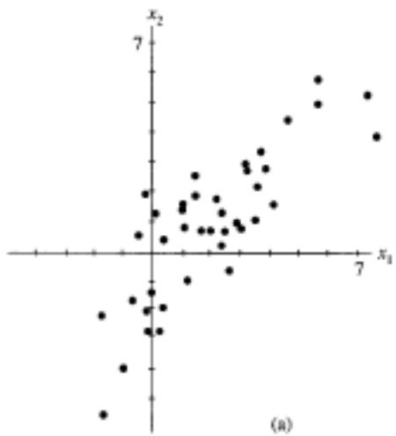


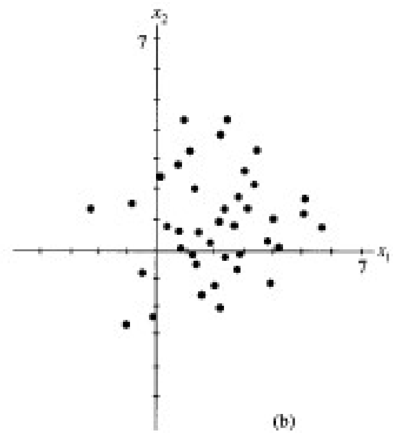

*Varianza Generalizada Muestral Pequeña*

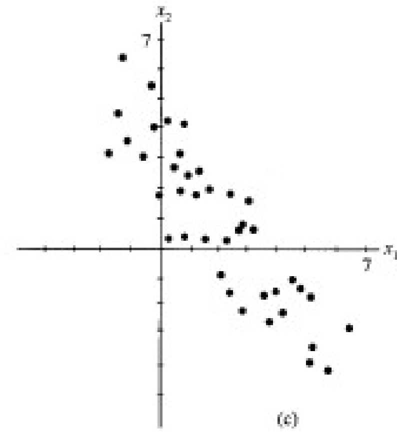
# Caso en que el volumen tri-dimensional es cero |S|=0



# Interpretación de la varianza generalizada







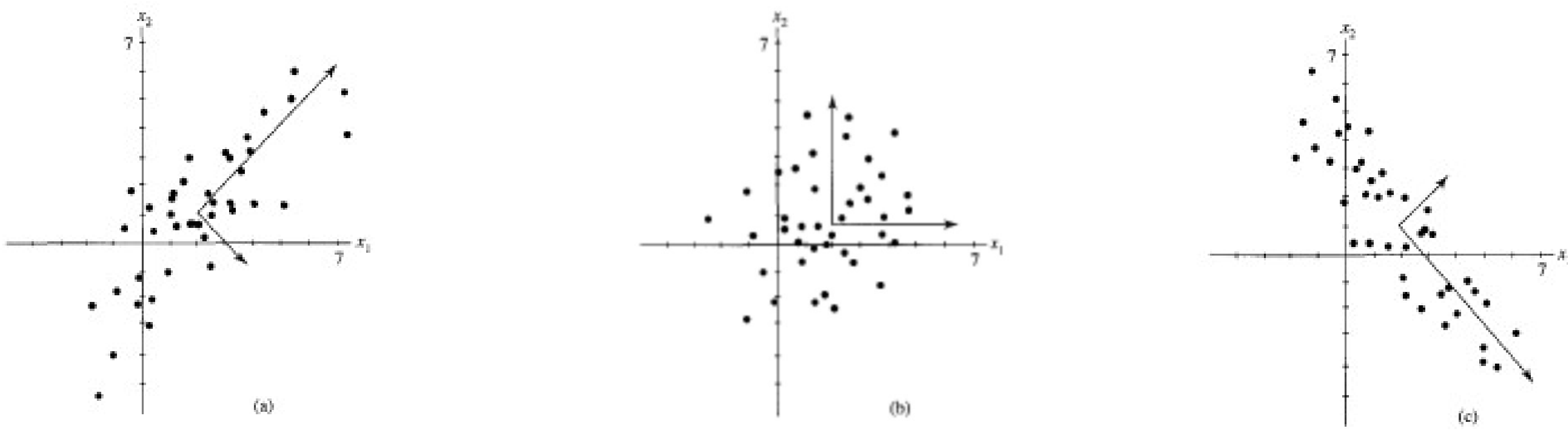
$$S = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \qquad S = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \qquad S = \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}$$

$r = 0.8$ $\qquad$ $r = 0$ $\qquad$ $r = -0.8$

$\bar{x}' = [2, 1]$ $\qquad$ $\bar{x}' = [2, 1]$ $\qquad$ $\bar{x}' = [2, 1]$

$|S| = 9$ $\qquad$ $|S| = 9$ $\qquad$ $|S| = 9$

## Diferentes estructuras de correlación no son detectadas por $|S|$




$\lambda_1 = 9 \quad e_1 = \left[1/\sqrt{2}\,,\ 1/\sqrt{2}\right]$

$\lambda_2 = 1 \quad e_2 = \left[1/\sqrt{2}\,,\ -1/\sqrt{2}\right]$

$\overline{x}' = \left[2\,,\ 1\right]$

$\lambda_1 = 3 \quad e_1 = \left[1\,,\ 0\right]$

$\lambda_2 = 3 \quad e_2 = \left[0\,,\ 1\right]$

$\overline{x}' = \left[2\,,\ 1\right]$

$\lambda_1 = 9 \quad e_1 = \left[1/\sqrt{2}\,,\ -1/\sqrt{2}\right]$

$\lambda_2 = 1 \quad e_2 = \left[1/\sqrt{2}\,,\ 1/\sqrt{2}\right]$

$\overline{x}' = \left[2\,,\ 1\right]$

## Ejemplo 3.3 (*) *Descomposición de un vector en su media y desviaciones centrales*

$$X = \begin{pmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{pmatrix}$$

En este caso,

$$\bar{x}_1 = (4 - 1 + 3)/3 = 2 \text{ and } \bar{x}_2 = (1 + 3 + 5)/3 = 3, \text{ so}$$

$$\bar{x}_1\mathbf{1} = 2\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} \qquad \bar{x}_2\mathbf{1} = 3\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

Consecuentemente,

$$\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\mathbf{1} = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

(*) Johnson y Wichern (2002) Applied Multivariate Statistical Analysis y,

$$\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2\mathbf{1} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

Se comprueba que $\bar{x}_1 1$ y $d_1$ son perpendiculares

$$(\bar{x}_1\mathbf{1})'(\mathbf{y}_1 - \bar{x}_1\mathbf{1}) = [2 \quad 2 \quad 2] \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = 4 - 6 + 2 = 0$$

El mismo resultado se obtiene para $\bar{x}_2 1$ y $d_2$ .

La descomposición de $y_1$ e $y_2$ es:

$$\mathbf{y}_1 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

$$\mathbf{y}_2 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

Para cualquiera dos vectores de desviaciones $d_1$ y $d_2$ :

$$\mathbf{d}_i'\mathbf{d}_k = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

Si $\theta_{ik}$ es el ángulo formado por los vectores $\bar{d}_i$ y $\bar{d}_k$ :

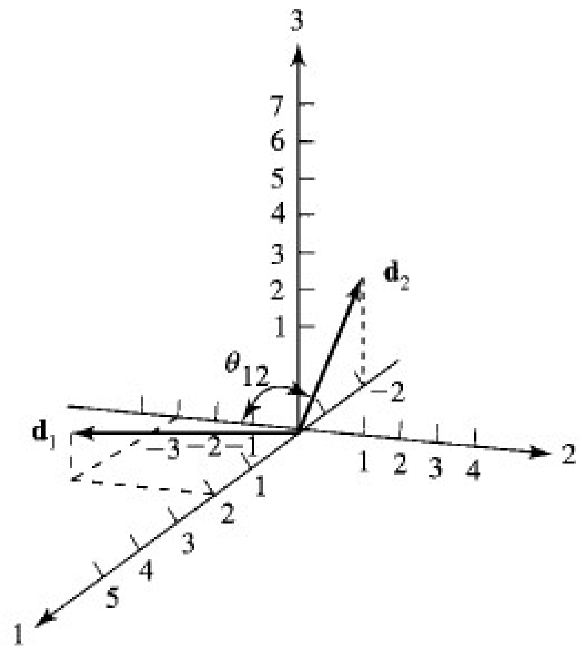
$$\mathbf{d}_i'\mathbf{d}_k = L_{\mathbf{d}_i} L_{\mathbf{d}_k} \cos(\theta_{ik})$$

O, equivalentemente con las observaciones muestrales:

$$\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) = \sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2} \sqrt{\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2} \cos(\theta_{ik})$$

El coseno del ángulo es el coeficiente de correlación muestral entre $d_i$ y $d_k$:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos(\theta_{ik})$$




Calculando la matriz varianza covarianza **Sn** y de correlaciones **R**

Se sabe que:

$$\mathbf{d}_1 = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{d}_2 = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

Veamos los componentes de la matriz varianza-covarianza:

$$\mathbf{d}_1'\mathbf{d}_1 = [2 \quad -3 \quad 1] \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = 14 = 3s_{11} \longrightarrow s_{11} = 14 / 3$$

$$\mathbf{d}_2'\mathbf{d}_2 = [-2 \quad 0 \quad 2] \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix} = 8 = 3s_{22} \longrightarrow s_{22} = 8 / 3$$

$$\mathbf{d}_1'\mathbf{d}_2 = [2 \quad -3 \quad 1] \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix} = -2 = 3s_{12} \longrightarrow s_{12} = -2 / 3$$

Consecuentemente,

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{-\frac{2}{3}}{\sqrt{\frac{14}{3}}\sqrt{\frac{8}{3}}} = -.189 \longrightarrow \theta_{12} = 100.65^{\circ}$$

La matriz varianza-covarianza y de correlación muestral serían:

$$\mathbf{S}_n = \begin{bmatrix} \frac{14}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{8}{3} \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} 1 & -.189 \\ -.189 & 1 \end{bmatrix}$$

# INTERPRETACIÓN GEOMÉTRICA DE LA MUESTRA

1. La proyección de la columna x(j) de la matriz X sobre el vector equiangular ***1*** es el vector $\bar{x}_j$ 1. Éste tiene norma $\sqrt{n}\,|\bar{x}_i|$. Así, la j-ésima media muestral, $\bar{x}_j$, se relaciona con la longitud de la proyección de x(j) sobre 1 o la componente de x(j) sobre ***1***.

2. La información contenida en $\boldsymbol{S_n}$ se obtuvo de los vectores de desviaciones $\boldsymbol{d_j} = \boldsymbol{y_j} - \bar{x}_{(j)}\,\boldsymbol{1} = \left(x_{1j} - \bar{x}_{(j)},\ x_{2j} - \bar{x}_{(j)},\ \ldots,\ x_{nj} - \bar{x}_{(j)}\right)$. El cuadrado de la longitud $\boldsymbol{d_j}$ es $\boldsymbol{ns_{jj}}$ y el producto interno entre $\boldsymbol{d_i}$ y $\boldsymbol{d_j}$ es $\boldsymbol{ns_{ij}}$.

3. La correlación muestral $\boldsymbol{r_{ik}}$ es el coseno del ángulo formado por $\boldsymbol{d_i}$ y $\boldsymbol{d_k}$.