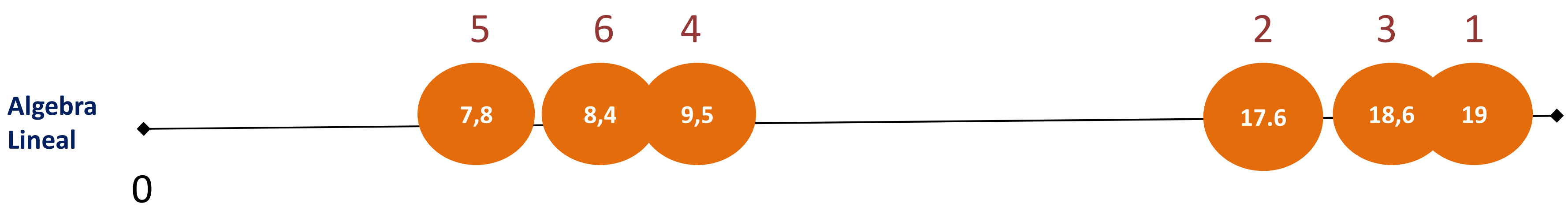
ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

Definición

Es una técnica matemática que permite reducir la dimensionalidad centrándose en las varianzas de un conjunto de datos multivariados obtenidos de una población cuya distribución de probabilidades no necesita ser conocida.

CALIFICACION FINAL DE ESTUDIANTES SEGÚN CURSOS

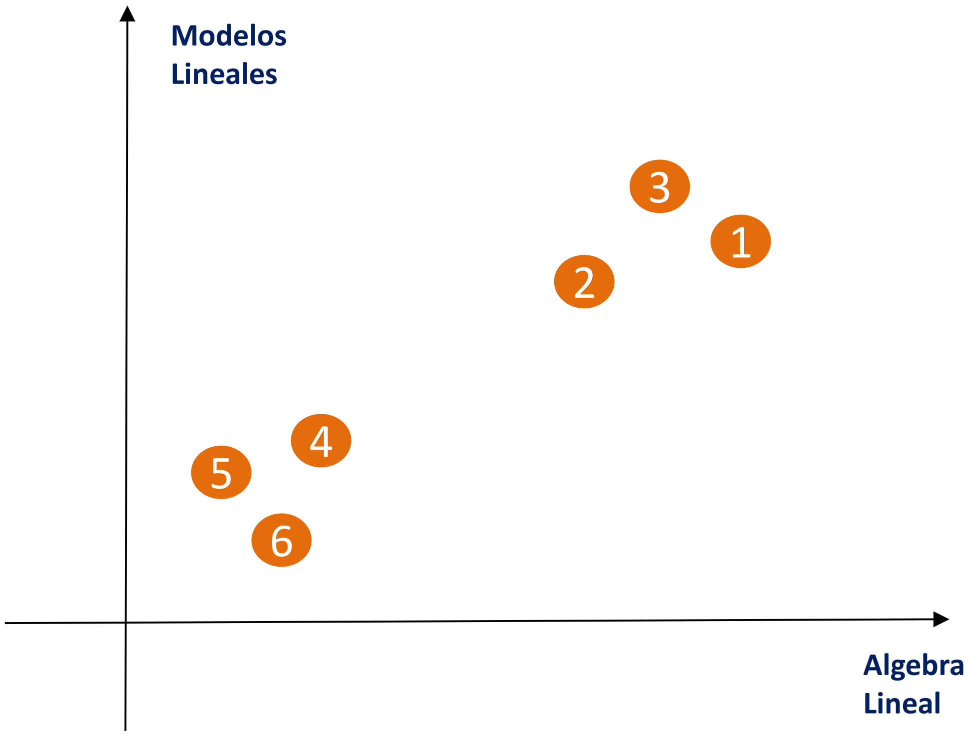
GRAFICO EN UNA DIMENSIÓN



	Estudiante 1	Estudiante 2	Estudiante 3	Estudiante 4	Estudiante 5	Estudiante 6
Algebra Lineal	19,0	17,6	18,6	9,5	7,8	8,4

CALIFICACION FINAL DE ESTUDIANTES SEGÚN CURSOS

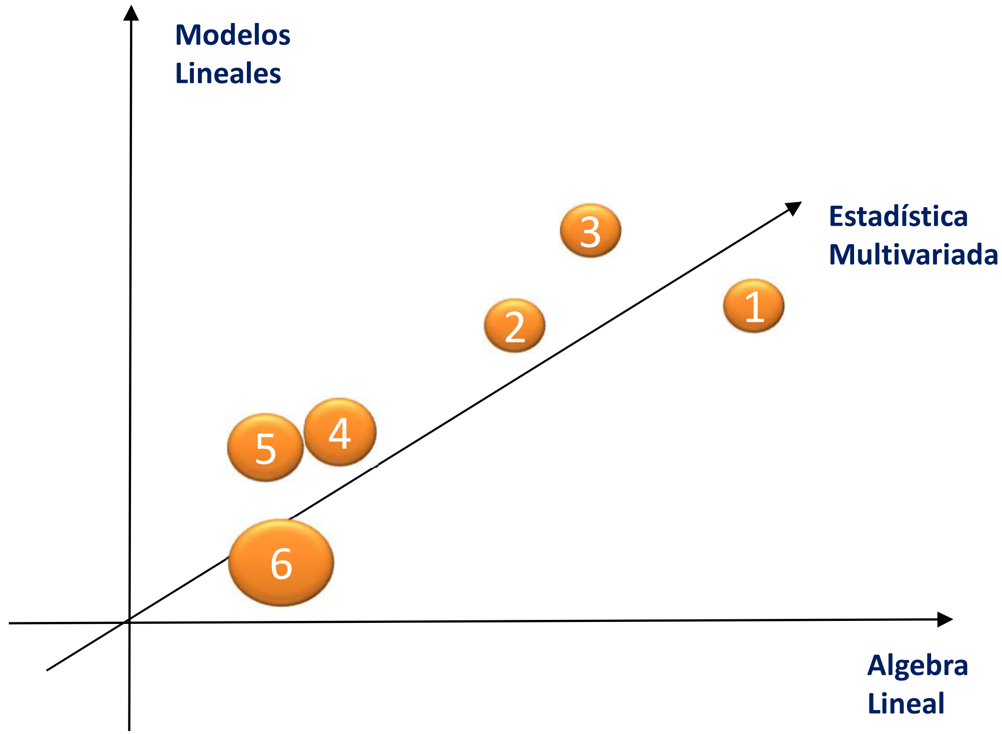
GRAFICO EN
DOS DIMENSIONES



	Estudiante 1	Estudiante 2	Estudiante 3	Estudiante 4	Estudiante 5	Estudiante 6
Algebra Lineal	19,0	17,6	18,6	9,5	7,8	8,4
Modelos Lineales	19,2	15,8	19,6	8,7	7,2	5,0

CALIFICACION FINAL DE ESTUDIANTES SEGÚN CURSOS

BOSQUEJO EN TRES DIMENSIONES



	Estudiante 1	Estudiante 2	Estudiante 3	Estudiante 4	Estudiante 5	Estudiante 6
Algebra Lineal	19,0	17,6	18,6	9,5	7,8	8,4
Modelos Lineales	19,2	15,8	19,6	8,7	7,2	5,0
Estadística Multivariada	17,2	11,0	15,1	5,0	4.4	3,0

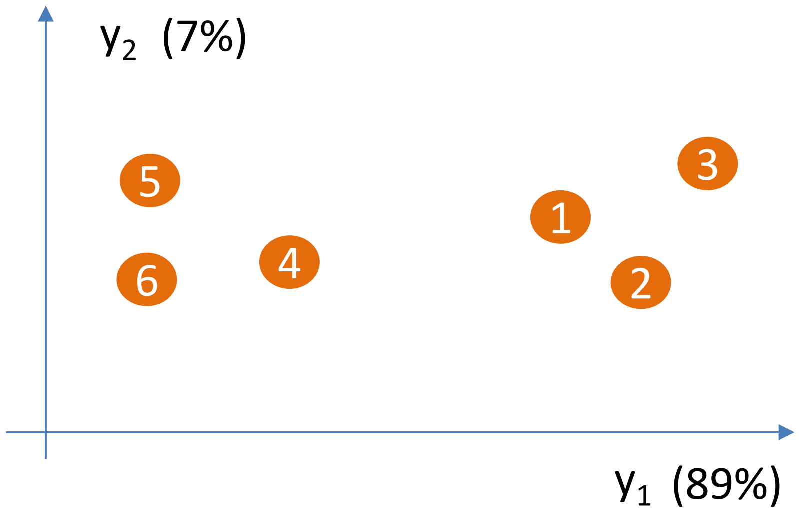
¿EN CUATRO DIMENSIONES?

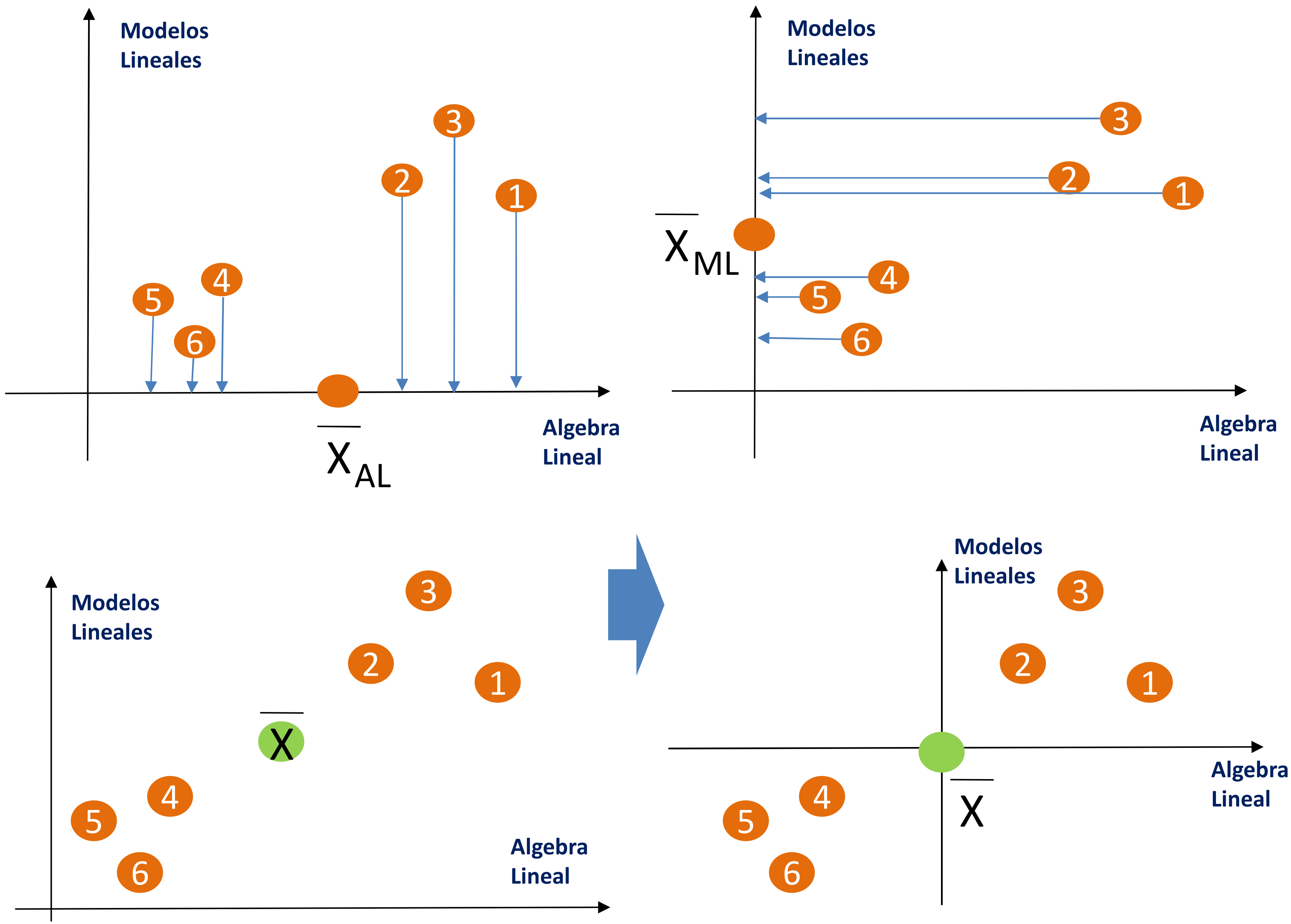
	Estudiante 1	Estudiante 2	Estudiante 3	Estudiante 4	Estudiante 5	Estudiante 6
Algebra Lineal	19,0	17,6	18,6	9,5	7,8	8,4
Modelos Lineales	19,2	15,8	19,6	8,7	7,2	5,0
Estadística Multivariada	17,2	11,0	15,1	5,0	4.4	3,0
Técnicas de redacción	12,1	15,0	13,3	6,0	8.3	15,0

¿Como se puede representar en más de tres dimensiones gráficamente?

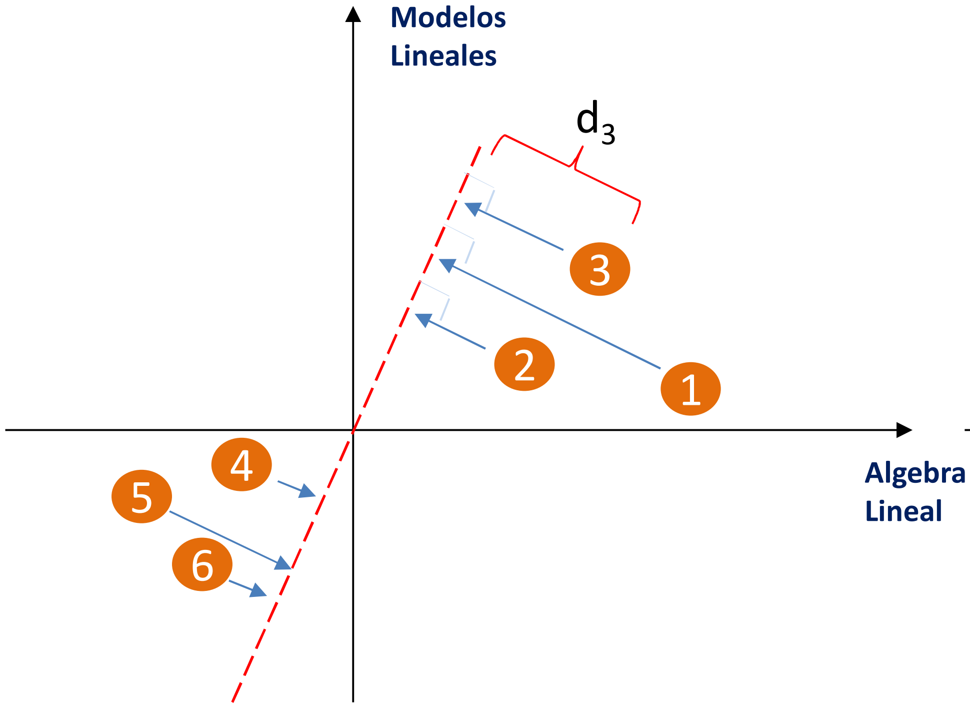
Calificación Final de Estudiantes según Cursos

	Estudiante 1	Estudiante 2	Estudiante 3	Estudiante 4	Estudiante 5	Estudiante 6
Algebra Lineal	19,0	17,6	18,6	9,5	7,8	8,4
Modelos Lineales	19,2	15,8	19,6	8,7	7,2	5,0
Estadística Multivariada	17,2	11,0	15,1	5,0	4.4	3,0
Técnicas de redacción	12,1	15,0	13,3	6,0	8.3	15,0

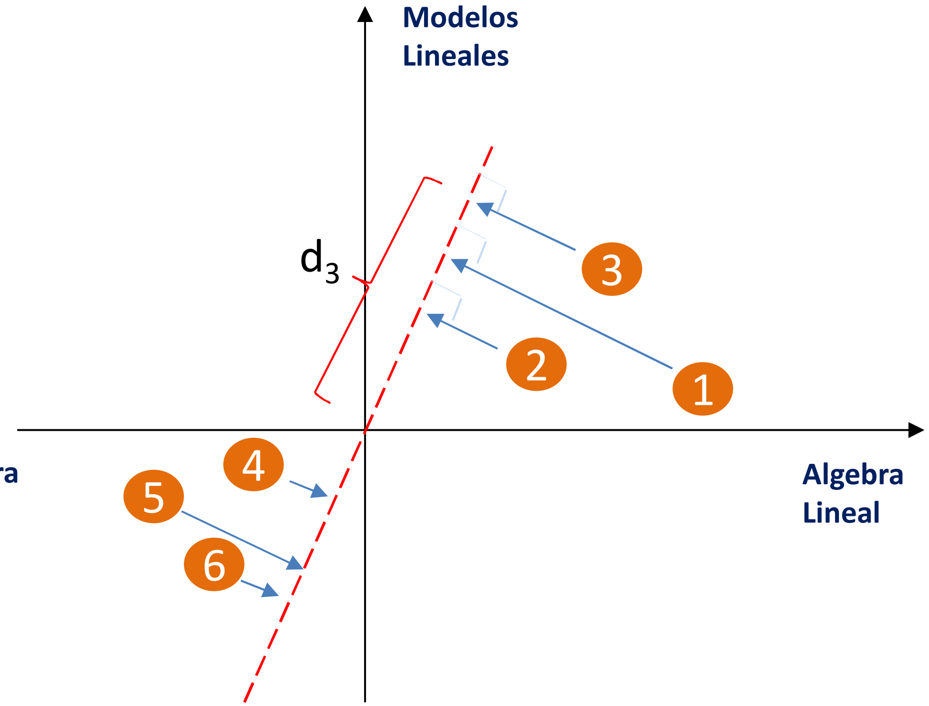



Modelos Lineales
Algebra Lineal
$\overline{X}_{AL}$
Modelos Lineales
Algebra Lineal
$\overline{X}_{ML}$
Modelos Lineales
Algebra Lineal
$\overline{X}$
Modelos Lineales
Algebra Lineal
$\overline{X}$

¿Como trabaja el análisis de componentes principales?

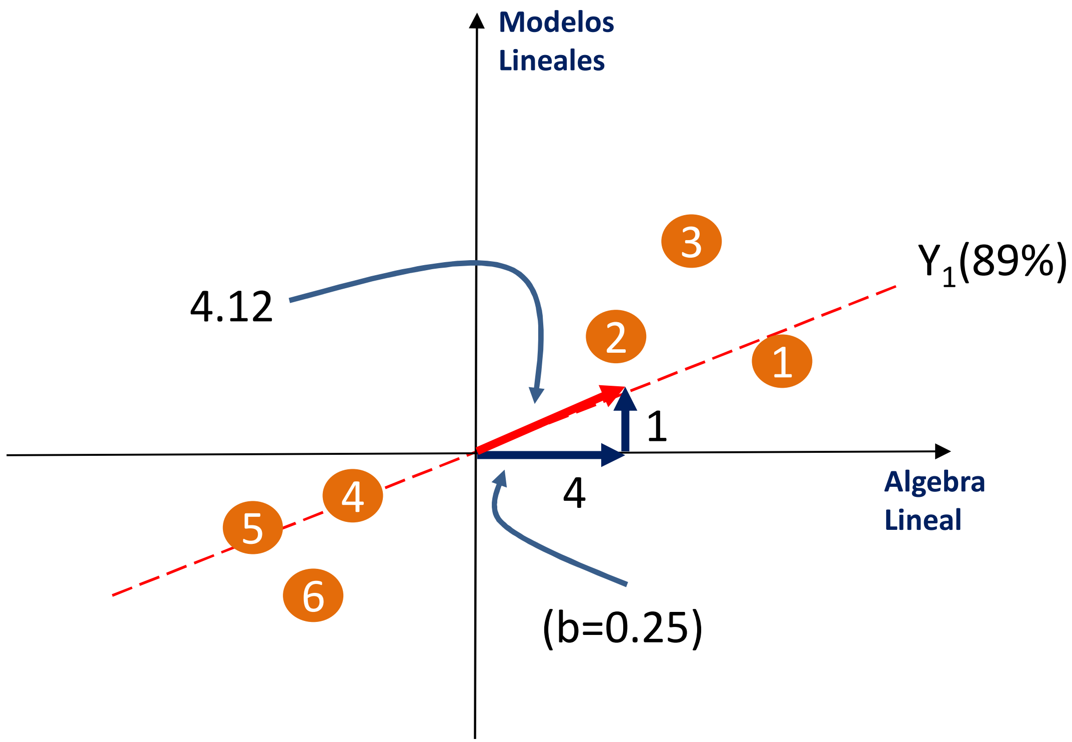


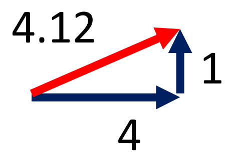
Min $d_1^2+d_2^2+d_3^2+d_4^2+d_5^2+d_6^2$=SC distancias



Max $d_1^2+d_2^2+d_3^2+d_4^2+d_5^2+d_6^2$=SC distancias

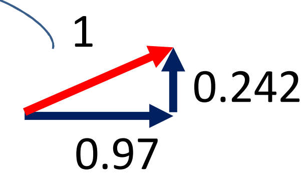
Combinación lineal = **4** de AL por **1** de ML





Relación pitagórica

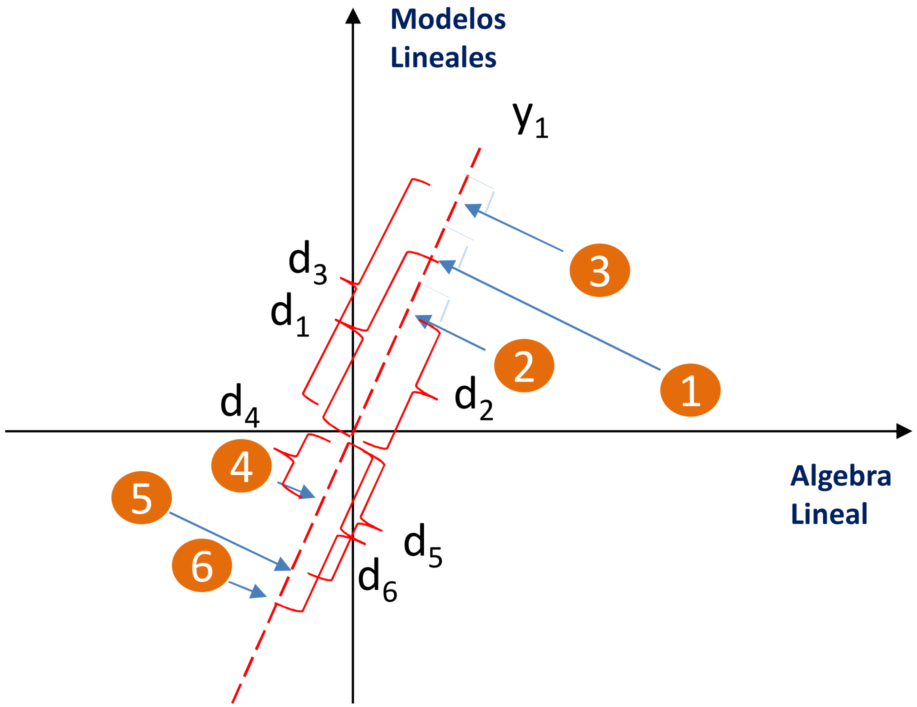
$$\frac{4.12}{4.12} = \sqrt{\left(\frac{1}{4.12}\right)^2 + \left(\frac{4}{4.12}\right)^2}$$



$$\begin{pmatrix} 0.242 \\ 0.970 \end{pmatrix}$$

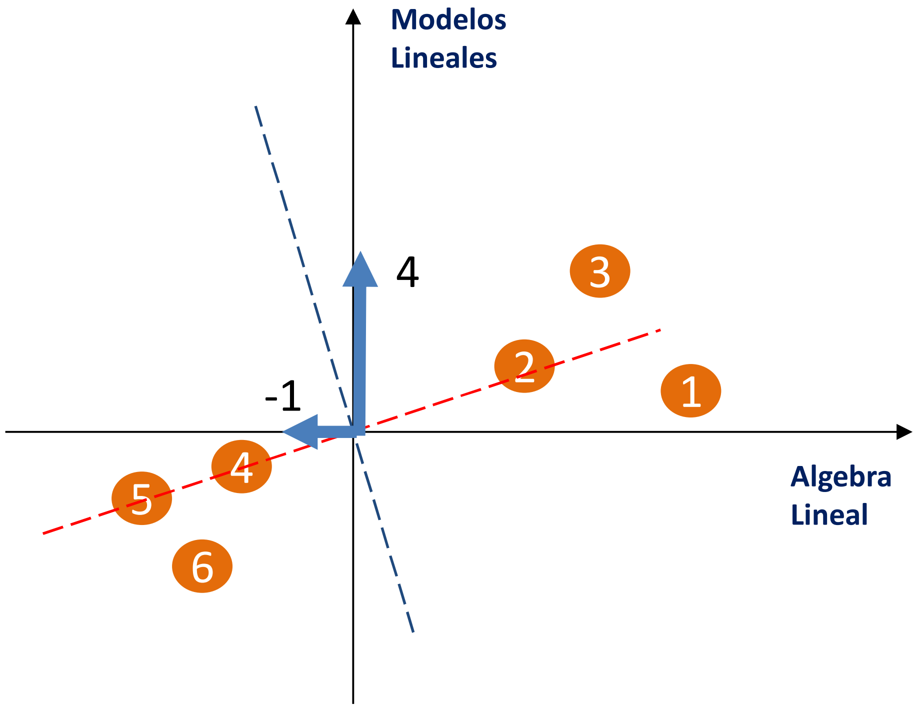
Vector característico, vector singular o "eigenvector"

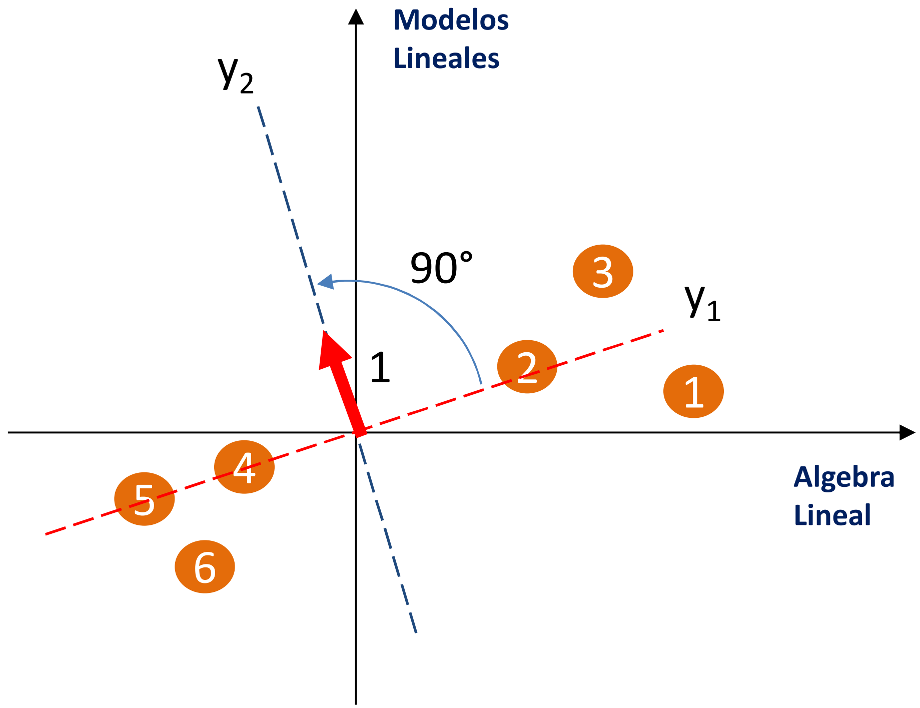
Combinación lineal = **0.97** puntos de AL por **0.242** de ML



$d_1^2+d_2^2+d_3^2+d_4^2+d_5^2+d_6^2$=SC distancias=**(valor característico de y_1)**^{2}

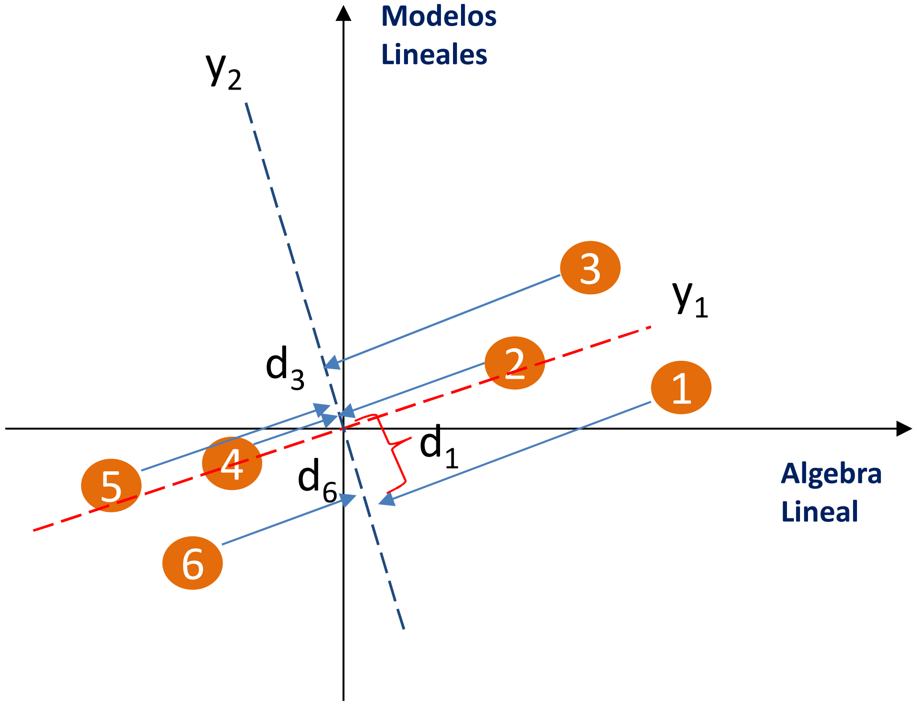
Combinación lineal =
0.97 puntos de ML por -**0.242** de AL





$$\begin{pmatrix} -0.242 \\ 0.970 \end{pmatrix}$$

Vector característico, vector singular o "eigenvector" de y_2



$d_1^2+d_2^2+d_3^2+d_4^2+d_5^2+d_6^2$=SC distancias=(**valor característico de y_2**)^2

Componentes Principales utilizando la descomposición de los valores característicos

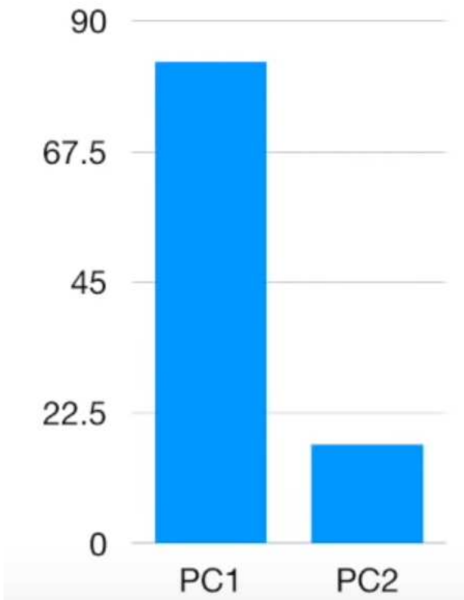
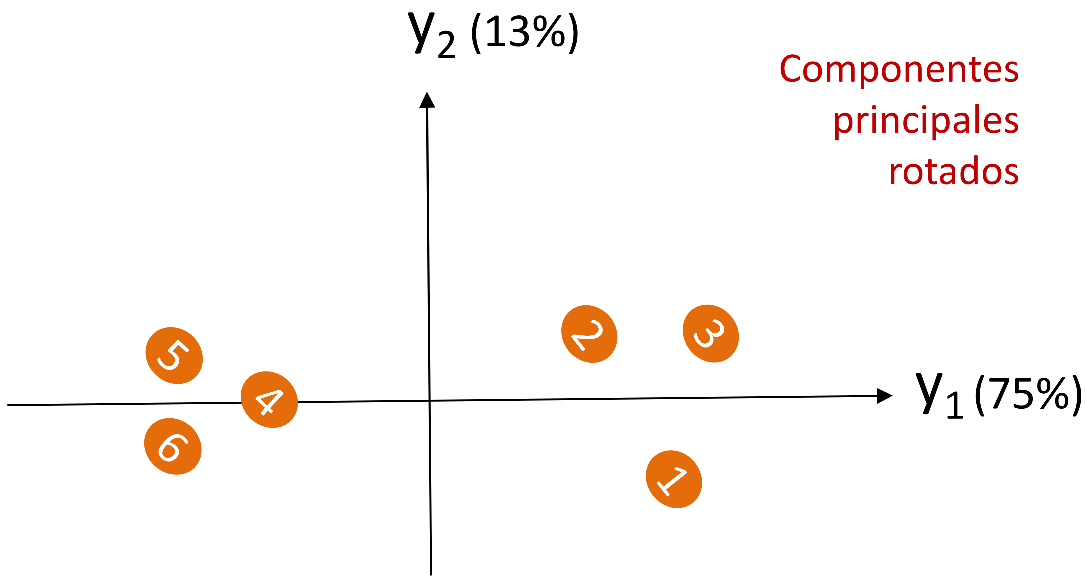


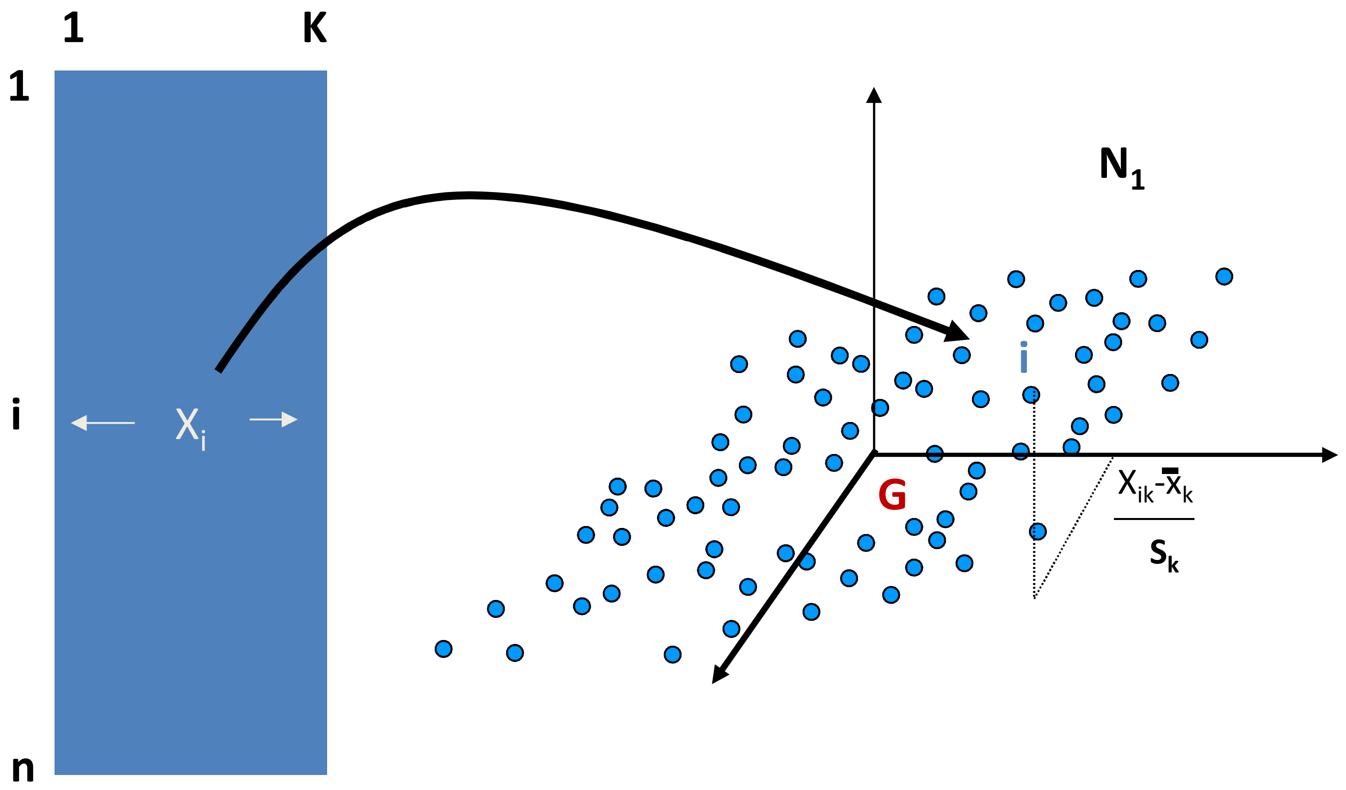
Gráfico de Sedimentación o Descomposición Espectral



Componentes principales rotados

Generalización del Análisis Geométrico

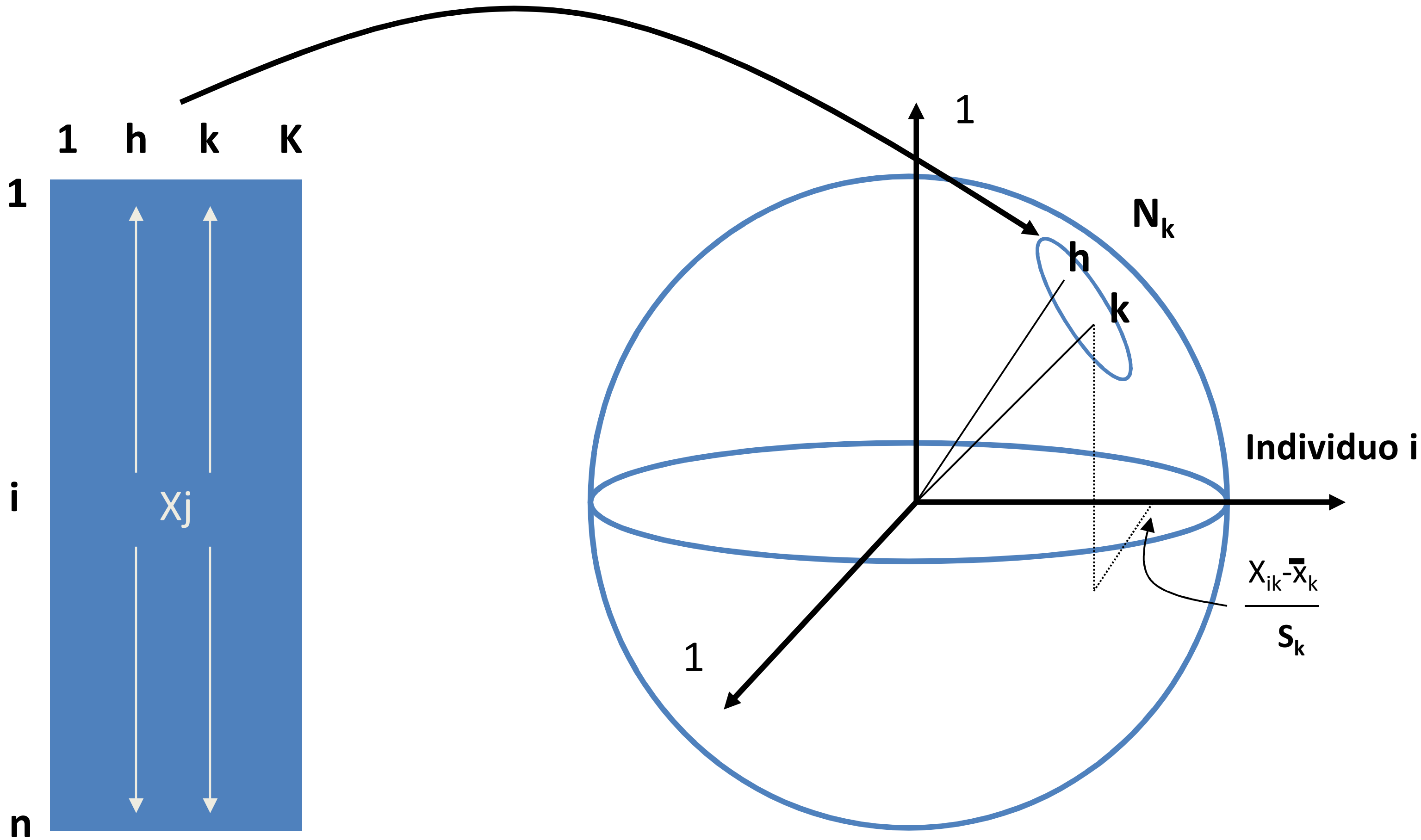
La Nube de Individuos



Individuos como yuxtaposición de "n" filas en R^k

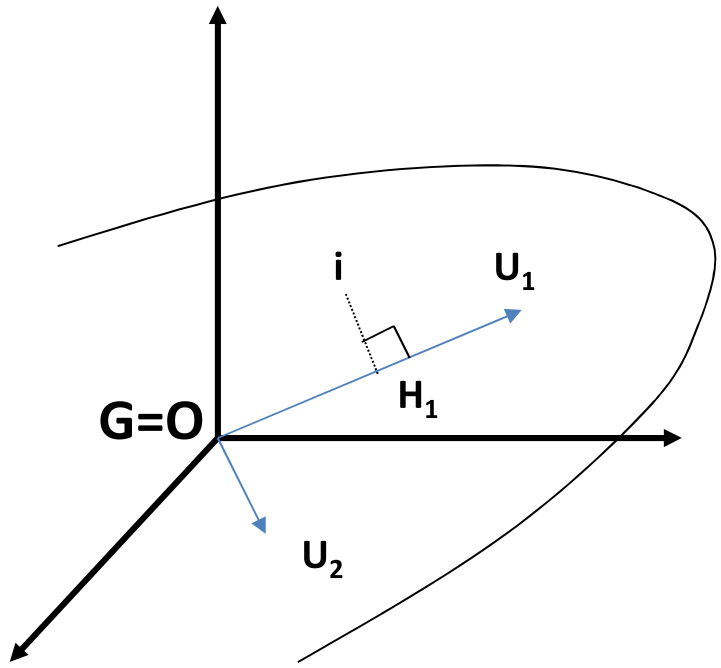
La Nube de Variables

Nube de Variables N_k asociada en el espacio R^n



Norma de la Variable X_k=1
(Hiperesfera)

El Ajuste de la Nube de Individuos



Critrerio

Máxima Inercia respecto al Centro de Gravedad G ⇒ Ejes Factoriales de Máximo Alargamiento de la Nube

El individuo “i” se proyecta sobre u_1 en H_i

Se busca en primer lugar u_i que maximiza ΣOH_i^2

Se busca u_2, ortogonal a u_1, que satisface el mismo criterio

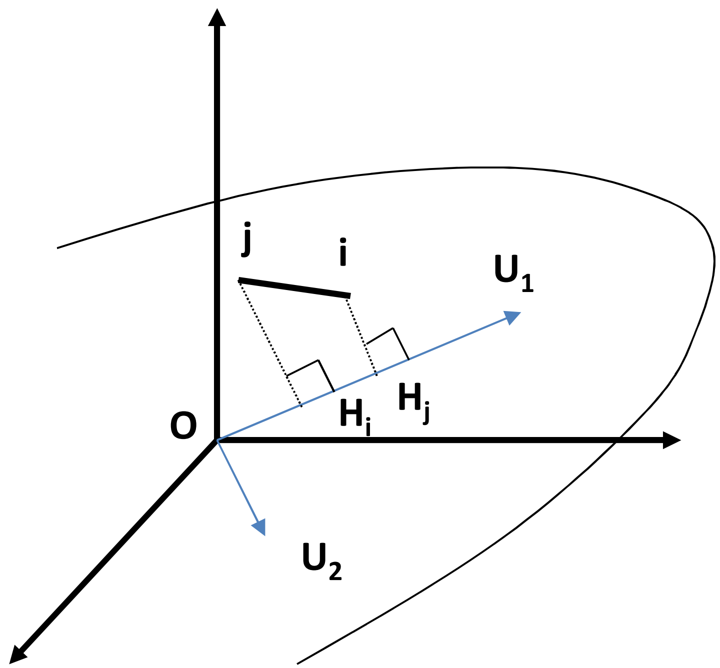
Si los individuos tienen pesos diferentes el criterio Consiste en maximizar $\Sigma_i p_i OH_i^2$

Distancias entre Individuos

El eje u_1 maximiza $\Sigma_i\Sigma_j\,(OH_i - OH_j)^2$ osea

$\Sigma_i\Sigma_j\, d^2(H_i, H_j)$ se acerca lo más posible a

$\Sigma_i\Sigma_j\, d^2(i,j)$



El Ajuste de la Nube de Variables

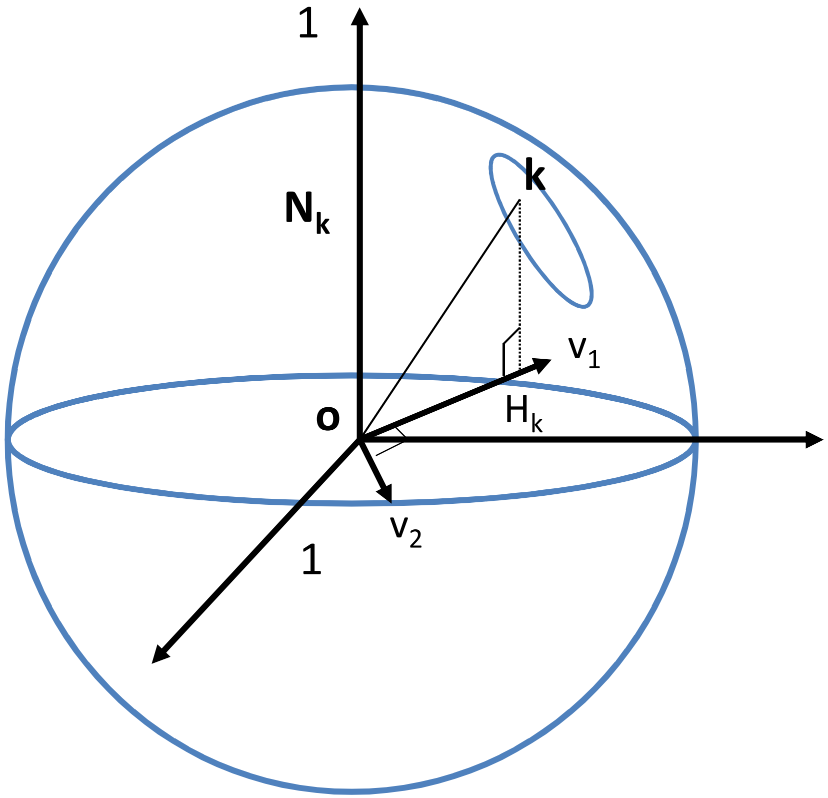
Critrerio

Máxima Inercia Proyectada → Ejes Factoriales Maximizan la Suma de Cosenos al Cuadrado

H_k: proyección sobre v_1 del punto que representa la variable k.

Se busca v_1 que maximice $\Sigma_k OH_k^2$

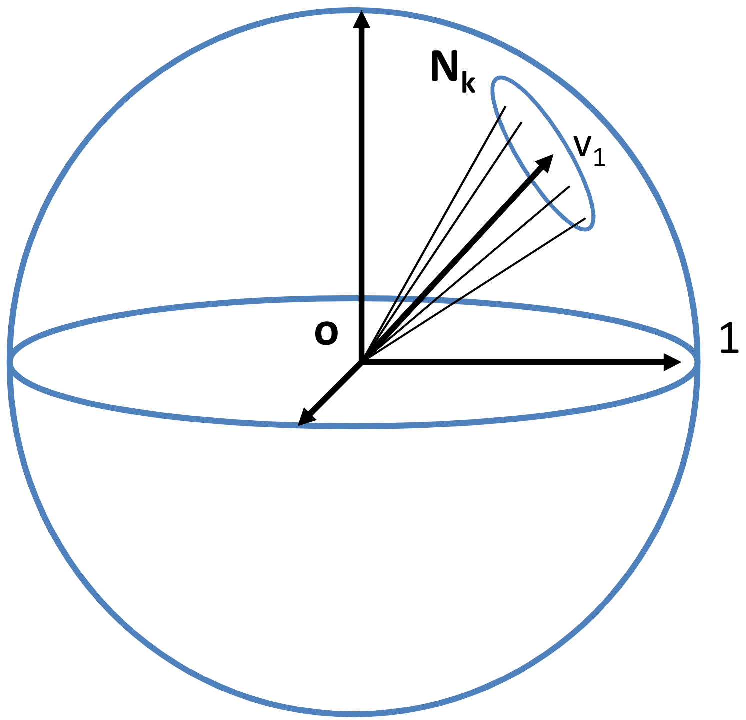
Después se busca v_2 ortogonal a v_1 que satisface el mismo criterio



El Efecto Talla en R^n

La nube N_k se concentra en un pequeño sector de la esfera.

La proyección de las variables sobre el primer eje factorial V_1 informa de la posición de N_k en relación a O.



Los Datos

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

$$
\begin{array}{cccccc}
 & & & & & \\
 & & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} & \begin{array}{l} \leftarrow \text{1ra observación} \\ \\ \\ \leftarrow \text{n-ésima observación} \end{array} \\
\text{Variables} \rightarrow & & \begin{array}{cccc} X_1 & X_2 & \quad & X_p \end{array} &
\end{array}
$$

$$X_{nxp} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Los Individuos

Lo que se evalúa es su semejanza. Dos individuos se asemejan más cuanto más próximos sean sus valores en el conjunto de las variables.

$$d^2(i,j) = \sum (x_{ik} - x_{jk})^2$$

Las Variables

Lo que se evalúa es su relación. La relación entre dos variables se mide por el coeficiente de correlación lineal, en otros casos, se utiliza la covarianza.

$$r(k,h) = \frac{Cov(k,h)}{\sqrt{var(k)\ var(h)}}$$

El Estudio Exploratorio

Semejanza entre Individuos

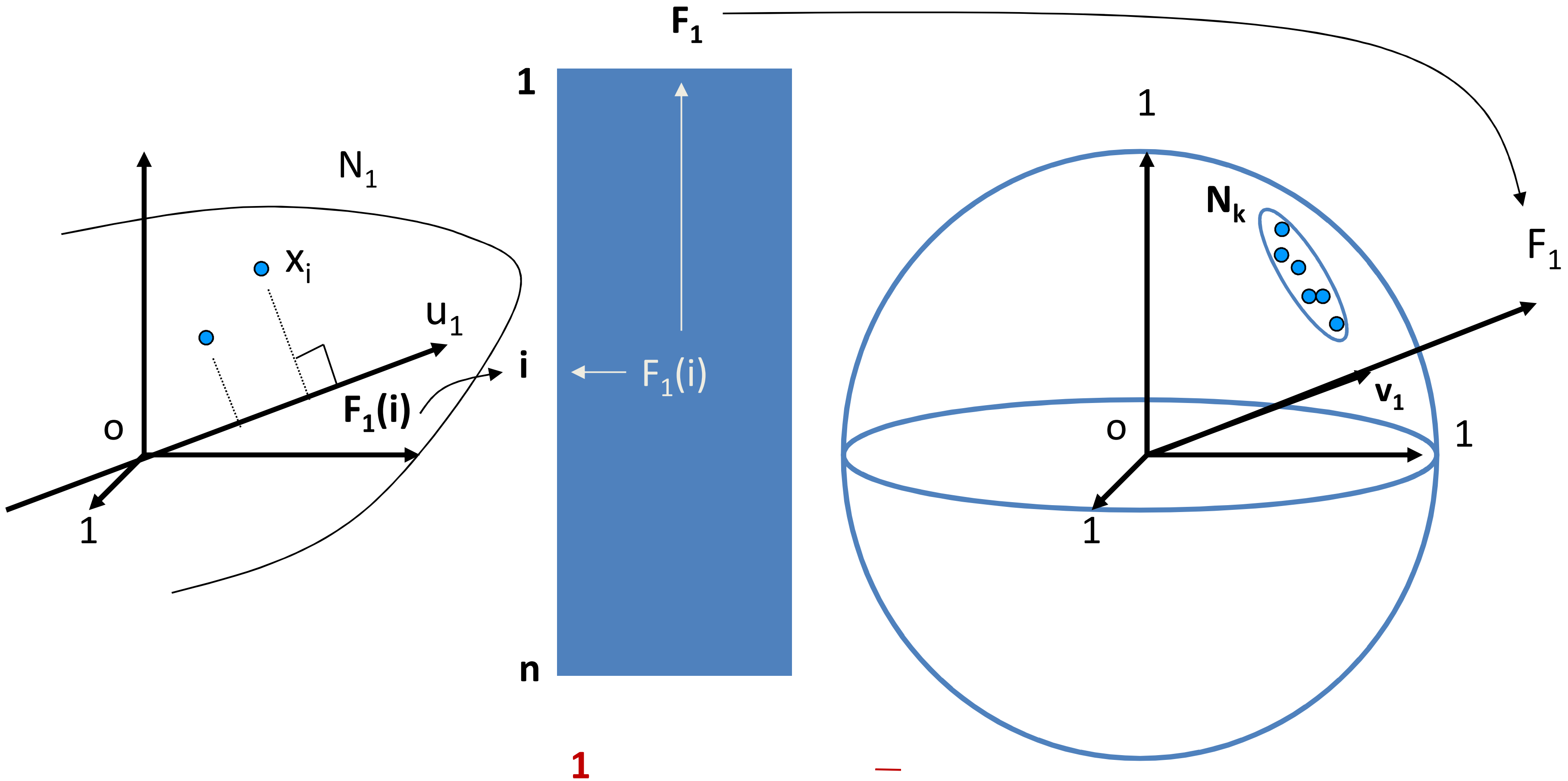
¿Qué Individuos se asemejan?
¿Qué Individuos se diferencian?
¿Existen grupos homogéneos de Individuos?
¿Se puede poner en evidencia una Tipología de Individuos?

DUALIDAD

Relación entre Variables

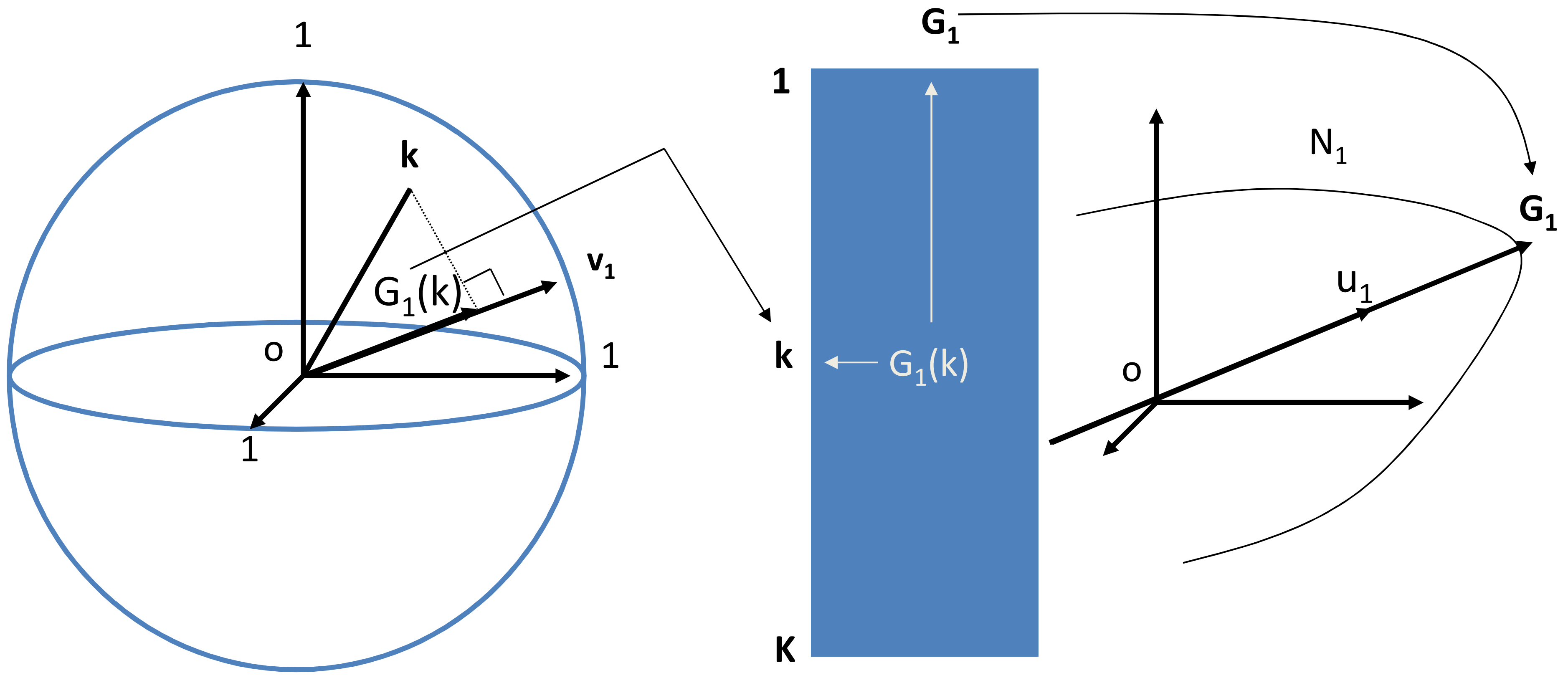
¿Qué variables se relacionan positivamente?
¿Qué variables se relacionan negativamente?
¿Existen grupos de variables correladas entre sí?
¿Se puede poner en evidencia una Tipología de Variables? Sintetizar (Componentes)

Primera Forma de Dualidad



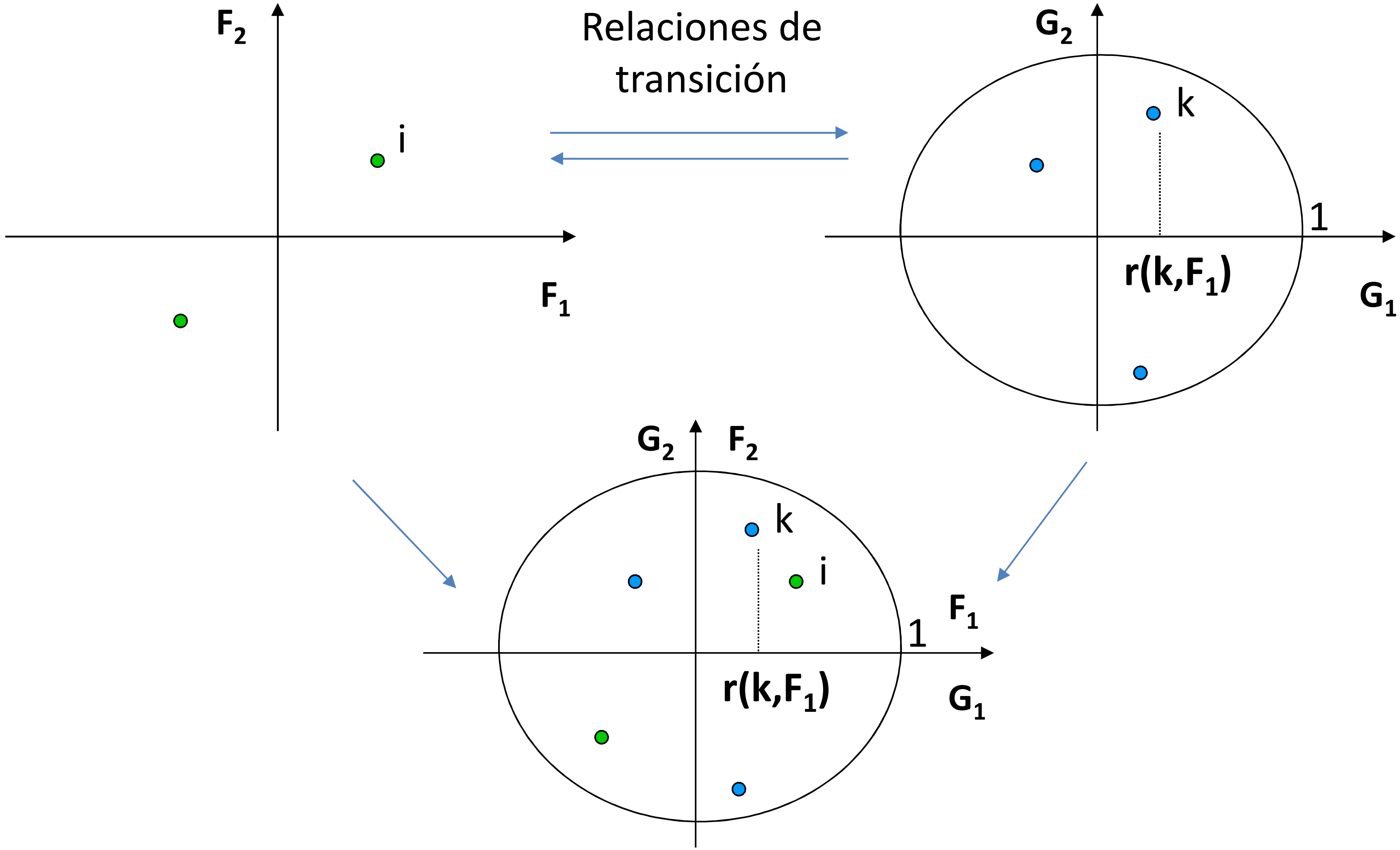
$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_S \frac{X_{ik} - \overline{x}_k}{S_k} G_s(k)$$

Segunda Forma de Dualidad



$$G_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{X_{ik} - \overline{x}_k}{S_k} F_s(k)$$

Análisis Dual



Cálculo de las Componentes Principales

Definición Operativa

La Componente Principal es una combinación lineal de 'p' variables aleatorias X_1, X_2, ..., X_p , representa un nuevo eje de coordenadas (rotación) cuya dirección es de máxima variabilidad.

Dado el vector aleatorio $X = [X_1, X_2, ..., X_p]$ y su covarianza Σ con valores característicos $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$

$$y_1 = L_1^t X = l_{11}X_1 + l_{21}X_2 + ... + l_{p1}X_p$$

$$y_2 = L_2^t X = l_{12}X_1 + l_{22}X_2 + ... + l_{p2}X_p$$

$$\vdots \qquad\qquad \vdots$$

$$y_p = L_p^t X = l_{1p}X_1 + l_{2p}X_2 + ... + l_{pp}X_p$$

$$Var(Y_i) = L_i^t \Sigma L_i \qquad i = 1, 2, ... ,p$$

$$Cov(Y_i, Y_k) = L_i^t \Sigma L_K \qquad i,k = 1, 2, ... , p$$

Proceso Iterativo de Maximización

Formulación Matemática K componentes principales

Primera iteración

$$\textbf{Max} \quad L_1^t S L_1$$
$$\textbf{s.a.} \quad L_1^t L_1 = 1$$

Segunda iteración

$$\textbf{Max} \quad L_2^t S L_2$$
$$\textbf{s.a.} \quad L_2^t L_2 = 1$$
$$L_2^t L_1 = 0$$

K-ésima iteración

$$\textbf{Max} \quad L_K^t S L_K$$
$$\textbf{s.a.} \quad L_k^t L_k = 1$$
$$L_k^t L_i = 0 \qquad \forall\ i$$

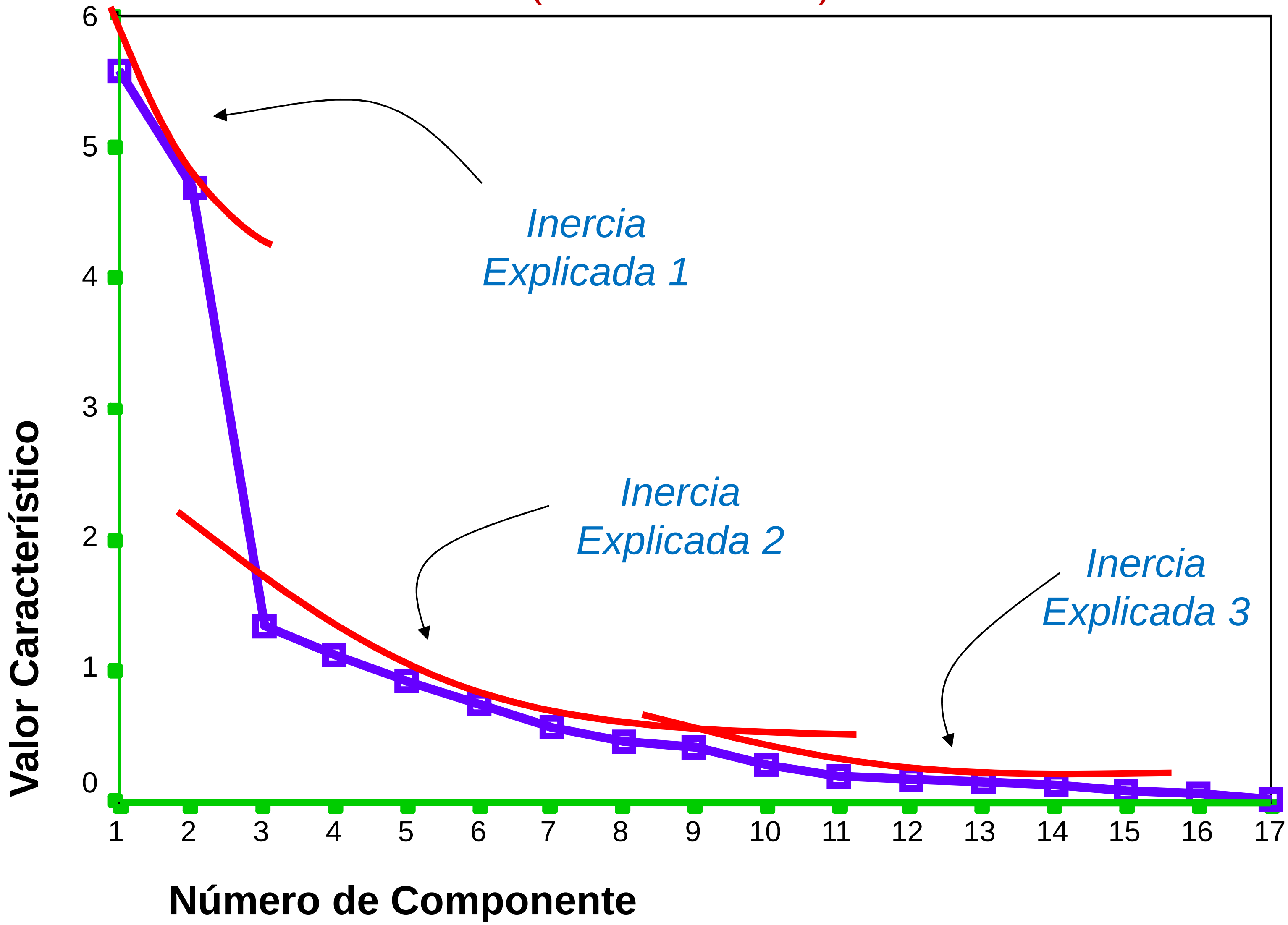
CRITERIO DE SELECCIÓN DE COMPONENTES PRINCIPALES

MÉTODO DE INERCIA TOTAL
(Coeficiente de Inercia)

$$r_k = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{Tr(\Lambda)} \qquad k \leq p$$

Se escogen las "p" componentes principales que explican el 100r% de la variabilidad total

MÉTODO DE ARCOS
(Método Gráfico)



MÉTODO DE INERCIA PROMEDIO

Componente Inercial "i-ésimo" i=1,...,p	$\geq$	**1**

Se selecciona la Componente Principal cuyo valor característico es mayor o igual a 1

$$\mathrm{Tr}(\Lambda) = p = \sum_{i} \lambda_i$$

"p" número de componentes principales totales

Formalización de las Formas de Dualidad

Relación entre los espacios $\Re^p$ y $\Re^n$

Definición

Los valores propios $\mu_1, \mu_2, ..., \mu_p$ asociado a los vectores propios $e_1, e_2, ..., e_p$ de XX^t son iguales respectivamente a los valores propios $\lambda_1, \lambda_2, ..., \lambda_p$ de X^tX

$$\mu_1 = \lambda_1, \mu_2 = \lambda_2, ..., \mu_p = \lambda_p$$

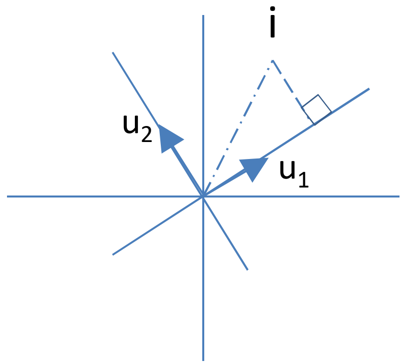
"La cantidad de información recogida por los ejes respectivos en ambos espacios es la misma"

ENFOQUE GEOMÉTRICO

Análisis en R^p:

$\underset{px1}{X_i}$ (observación i)

$$Comp_{\overline{u}_1}\, x_i = F_1(i) = x'_i\, \overline{u}_1$$



$$\underset{(nxp)(px1)}{F_\alpha() = X\, \overline{u}_\alpha}$$

$$\begin{array}{ll} \max & \overline{u}_\alpha' X' X \overline{u}_\alpha \\ s.a. & u_\alpha'\, u_\alpha = 1 \end{array}$$

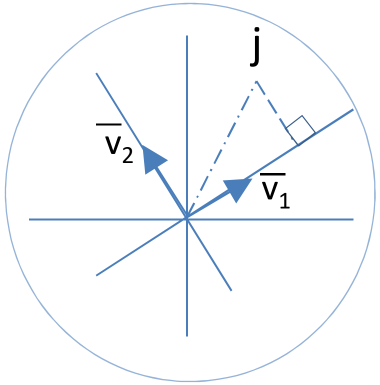
Sabiendo que: $X'X\, u_\alpha = \lambda_\alpha\, u_\alpha$ (a)

$u_\alpha\, 'X'X\, u_\alpha = \lambda_\alpha$ (máximo)

Análisis en R^n:

$\underset{nx1}{X_j}$ (variable j)

$Comp_{\bar{v}_1}\, x_j = G_1(j) = x'_j\, \bar{v}_1$



$G_\alpha() = \underset{(nxp)}{X'}\, \underset{(px1)}{\bar{v}_\alpha}$

$$\boxed{\begin{array}{ll} \max & \bar{v}_\alpha' XX' \bar{v}_\alpha \\ \text{s.a.} & \bar{v}_\alpha'\, \bar{v}_\alpha = 1 \end{array}}$$

Sabiendo que: $XX'\, \bar{v}_\alpha = \mu_\alpha\, \bar{v}_\alpha$ - - - - - - - - - - - - - - - - - (b)

$\bar{v}_\alpha' XX'\, \bar{v}_\alpha = \mu_\alpha$ (máximo)

Análisis entre R^p y R^n:

De (a): $X'X\,\overline{u}_\alpha = \lambda_\alpha \overline{u}_\alpha$ premultiplicando por X

$$(XX')X\,\overline{u}_\alpha = \lambda_\alpha X\overline{u}_\alpha \quad \Rightarrow \quad \overline{v}_\alpha \propto X\overline{u}_\alpha$$

Como μ_α es máximo y asociado a $\overline{v}_\alpha$ $\Rightarrow$ $\mu_\alpha \geq \lambda_\alpha$ (v)

De (b): $XX'\,\overline{v}_\alpha = \mu_\alpha \overline{v}_\alpha$ premultiplicando por X'

$$(X'X)X'\,\overline{v}_\alpha = \mu_\alpha X'\overline{v}_\alpha \quad \Rightarrow \quad \overline{u}_\alpha \propto X'\overline{v}_\alpha$$ de (b)

Como λ_α es máximo y asociado a $\overline{u}_\alpha$ $\Rightarrow$ $\lambda_\alpha \geq \mu_\alpha$ (vi)

De (v) y (vi): $\boxed{\lambda_\alpha = \mu_\alpha}$

Sabemos que: $\bar{u}_\alpha \propto X'\bar{v}_\alpha$

$$\bar{u}_\alpha = kX'\bar{v}_\alpha \longrightarrow \bar{u}_\alpha'\bar{u}_\alpha = k^2\,\bar{v}'_\alpha XX'\bar{v}_\alpha$$

$$\bar{u}_\alpha'\bar{u}_\alpha = k^2\,\mu_\alpha$$

$$\longrightarrow K = 1/\sqrt{\mu_\alpha}$$

Luego,

$$\sqrt{\mu_\alpha}\,\bar{u}_\alpha = X'\bar{v}_\alpha$$

$$\sqrt{\mu_\alpha}\,\bar{u}_\alpha\bar{v}'_\alpha = X'\,\bar{v}_\alpha\bar{v}'_\alpha$$

$$\sum_\alpha \sqrt{\mu_\alpha}\,\bar{u}_\alpha\bar{v}'_\alpha = \sum_\alpha X'\,\bar{v}_\alpha\bar{v}'_\alpha$$

$$\boxed{X' \cong \sum_\alpha \sqrt{\mu_\alpha}\,\bar{u}_\alpha\bar{v}'_\alpha}$$ - - - - - - - - - - comparar con (iv)

Definición

Conocido los vectores propios de un sub espacio se pueden obtener los del otro sin necesidad de una nueva factorización. Así tenemos,

$$\boxed{v_\alpha = (1 / \lambda_\alpha)X^t u_a}$$

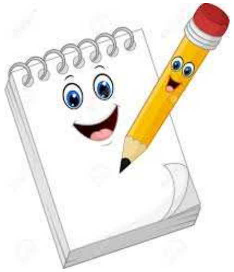
$$\boxed{u_\alpha = (1 / \lambda_\alpha)X v_a}$$

Definición

Existe una proporcionalidad entre las coordenadas de los puntos individuos sobre el eje factorial α en $\Re^p$, $x\nu_a$ y las componentes del vector unitario director del eje α en el otro espacio ν_α

$$G_\alpha = X^t U_\alpha = \sqrt{\lambda_\alpha}\,\nu_\alpha$$

$$G_\alpha(j) = \sum_i x_{ij} U_{\alpha i} = \sqrt{\lambda_\alpha}\,\nu_{\alpha i}$$



Análisis sobre las variables y sobre los individuos.

Ejemplo

Matriz de dispersión sobre las variables $\longrightarrow$ $X^{T}X$ (pxp) $\quad \dfrac{X^{T}X}{n-1}$

Matriz de dispersión sobre los individuos $\longrightarrow$ XX^{T} (nxn) $\quad \dfrac{XX^{T}}{p-1}$

$$\begin{array}{cc} & \textbf{variables} \\ \mathbf{X} = & \begin{bmatrix} -2 & -2 & -1 \\ 0 & 1 & 0 \\ 2 & 2 & 2 \\ 0 & -1 & -1 \end{bmatrix} \end{array} \ \textbf{observaciones}$$

variables

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 8 & 8 & 6 \\ 8 & 10 & 7 \\ 6 & 7 & 6 \end{bmatrix}$$

Tr(**X’X**)=24

Raíces características de **X’X**:

$$l_1 = 22.2819$$
$$l_2 = 1.0000$$
$$l_3 = .7181$$

X’X: es definida positiva de rango 3

observaciones

$$\mathbf{X}\mathbf{X}' = \begin{bmatrix} 9 & -2 & -10 & 3 \\ -2 & 1 & 2 & -1 \\ -10 & 2 & 12 & -4 \\ 3 & -1 & -4 & 2 \end{bmatrix}$$

Tr(**XX’**)=24

Raíces características de **XX’**:

$$l_1 = 22.2819$$
$$l_2 = 1.0000$$
$$l_3 = .7181$$
$$l_4 = 0$$

XX’: es semidefinida positiva de rango 3

Vectores característicos de **X'X**:

$$\mathbf{U} = \begin{bmatrix} -.574 & .816 & -.066 \\ -.654 & -.408 & .636 \\ -.493 & -.408 & -.768 \end{bmatrix}$$

U'U=I

Vectores característicos de **XX'**:

$$\mathbf{U}^* = \begin{bmatrix} .625 & -.408 & -.439 \\ -.139 & -.408 & .751 \\ -.729 & 0 & -.467 \\ .243 & .816 & .156 \end{bmatrix}$$

U*'U*=I

Valores característicos **L** de **X'X y XX'**:

$$\mathbf{L} = \begin{bmatrix} 22.2819 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & .7181 \end{bmatrix} \qquad \mathbf{L}^{1/2} = \begin{bmatrix} 4.7204 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & .8474 \end{bmatrix}$$

$$\mathbf{V} = \mathbf{U}\mathbf{L}^{1/2} = \begin{bmatrix} -2.707 & .816 & .056 \\ -3.089 & -.408 & -.539 \\ -2.326 & -.408 & .651 \end{bmatrix}$$

$$\mathbf{V}^* = \mathbf{U}^*\mathbf{L}^{1/2} = \begin{bmatrix} 2.9549 & -.408 & -.372 \\ -.656 & -.408 & .636 \\ -3.441 & 0 & -.396 \\ 1.147 & .816 & .132 \end{bmatrix}$$

Escalamiento de los vectores característicos por $\mathbf{L}^{1/2}$

Se verifica que:

$$\mathbf{VV'=X'X} \text{ y } \mathbf{V^*V^{*\prime}=XX'} \qquad \text{(i)}$$

Debido a la descomposición espectral:

$$\mathbf{X'X=ULU'} \text{ y } \mathbf{XX'=U^*LU^{*\prime}} \qquad \text{(ii)}$$

$$\underset{(pxp)}{\mathbf{W}} = \mathbf{U}\mathbf{L}^{-1/2} = \begin{bmatrix} -.122 & .816 & -.078 \\ -.139 & -.408 & .751 \\ -.104 & -.408 & -.906 \end{bmatrix}$$

$$\underset{(nxp)}{\mathbf{W}^*} = \mathbf{U}^*\mathbf{L}^{-1/2} = \begin{bmatrix} .132 & -.408 & -.518 \\ -.029 & -.408 & .886 \\ -.154 & 0 & -.552 \\ .051 & .816 & .184 \end{bmatrix}$$

Escalamiento de los vectores característicos por $\mathbf{L}^{-1/2}$

Se verifica (i) y (ii) para **W** y **W***

Scores de las observaciones (CP):

$$\underset{(nxp)}{\mathbf{Y}} = \mathbf{XW} = \mathbf{U}^*$$

*"Los scores de las **n** observaciones (CP) son iguales a los vectores característicos obtenidos de las matriz suma de cuadrados de observaciones – ver (ii)"*

Scores de las variables (CP):

$$\underset{(pxp)}{\mathbf{Y}^*} = \mathbf{X}'\mathbf{W}^* = \mathbf{U}$$

*"Los scores de las **p** variables (CP) son iguales a los vectores característicos obtenidos de la matriz suma de cuadrados de variables – ver (ii)"*

DESCOMPOSICIÓN POR VALOR SINGULAR (DVS)

Los datos originales se pueden obtener de las CP:

$$\mathbf{X} = \mathbf{YV}' \qquad \text{(iii)}$$

O también en términos de las transformaciones:

Descomposición por valor singular (DVS)

$$\boxed{\mathbf{X} = \mathbf{U}^*\mathbf{L}^{1/2}\mathbf{U}'} \qquad \text{(iv)}$$

En DVS una matriz de datos X se descompone como el producto de vectores característicos de **X'X**, los vectores característicos de **XX'** y una función de sus raíces características **L**.

$$\mathbf{X} = \mathbf{YL}^{1/2}\mathbf{U}' = \mathbf{U}^*\mathbf{L}^{1/2}\mathbf{Y}^*$$

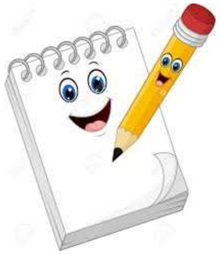
RESULTADOS IMPORTANTES

1 *Ninguna combinación lineal estandarizada de $x_{(px1)}$ tiene varianza mayor que λ_1, la varianza del primer componente principal*

2 *Si la matriz de covarianzas de **x** tiene rango "**r**" (r<p), entonces, la variación de x puede ser totalmente explicada por las primeras "**r**" componentes principales.*

3 *Si $Y_1=e_1X$, $Y_2=e_2X$, …, $Y_p=e_pX$ son los componentes principales obtenidos a partir de Σ, entonces,*

$$\rho_{Y_i,X_k} = \frac{e_{ki}\sqrt{\lambda i}}{\sqrt{\sigma_{kk}}}$$



Ejemplo

Se realiza un estudio cuyo objetivo es poner de relieve los factores que diferencian al máximo las marcas entre si y determinar aquellas marcas que el conjunto de encuestados considera semejantes.

Marcas	Características		
	Elegancia	Comodidad	Deportivo
A	2	3	6
B	3	2	4
C	4	5	4
D	5	5	4
E	8	9	6
F	9	7	7

Paso 1.- Tipificado de datos

$$X_{ij} = \frac{z_{ij} - \overline{z_j}}{S_j \sqrt{n}}$$

i = 1, ...,6

j = 1, ...,3

n = 6

Así obtenemos,

$$X = \begin{pmatrix} -0.51 & -0.38 & 0.28 \\ -0.35 & -0.55 & -0.39 \\ -0.19 & -0.03 & -0.39 \\ -0.03 & -0.03 & -0.39 \\ -0.46 & 0.67 & 0.28 \\ 0.62 & 0.03 & 0.63 \end{pmatrix}$$

Paso 2.- Cálculo de la Matriz de Correlación

$$\mathbf{C = X^tX =} \begin{pmatrix} 1 & 0.89 & 0.58 \\ 0.89 & 1 & 0.52 \\ 0.58 & 0.52 & 1 \end{pmatrix}$$

Paso 3.- Cálculo de la Inercia Explicada

$$\begin{vmatrix} 1 - \lambda & 0.89 & 0.58 \\ 0.89 & 1 - \lambda & 0.52 \\ 0.58 & 0.52 & 1 - \lambda \end{vmatrix} = 0 \quad \Rightarrow \quad \begin{array}{l} \lambda_1 = 2.334 \\ \lambda_2 = 0.560 \\ \lambda_3 = 0.106 \end{array}$$

Se verifica que: $Tr(C) = \lambda_1 + \lambda_2 + \lambda_3 = 3$

Componentes Principales e Inercia

Componente i-ésimo	Inercia Total (λ_i)	Inercia % (λ_i / p)*100%	r_p
1	2.334	78.16%	78.16
2	0.560	13.36%	91.52
3	0.106	8.48%	100.00

Paso 4.- Obtención de las Representaciones Gráficas

Dirección del Primer Eje

Sea v_α el vector que sigue la dirección del primer eje, entonces,

$$v_\alpha = \begin{pmatrix} v_{11} \\ v_{12} \\ v_{13} \end{pmatrix} \quad \text{donde} \quad v_{11}^2 + v_{12}^2 + v_{13}^2 = 1 \qquad \textbf{(i)}$$

y cumple que: $$(C - \lambda_1 I)\, v_\alpha = 0 \qquad \textbf{(ii)}$$

De (i) y (ii) se obtiene que: $$v_\alpha = \begin{pmatrix} 0.622 \\ 0.603 \\ 0.505 \end{pmatrix}$$

Paso 5.- Proyección de las Marcas sobre el Primer Eje:

$$F_1 = \sqrt{\frac{n}{p}} \times \mathbf{v_1} \qquad \text{donde} \qquad \sqrt{\frac{n}{p}} = \sqrt{\frac{6}{3}}$$

$$F_1 = \sqrt{\frac{6}{3}} \begin{pmatrix} -0.51 & -0.38 & 0.28 \\ -0.35 & -0.55 & -0.39 \\ -0.19 & -0.03 & -0.39 \\ -0.03 & 0.03 & 0.39 \\ 0.46 & 0.67 & 0.28 \\ 0.62 & 0.03 & 0.63 \end{pmatrix} \begin{pmatrix} 0.622 \\ 0.603 \\ 0.505 \end{pmatrix} = \begin{pmatrix} -0.57 \\ -0.06 \\ -0.47 \\ -0.33 \\ 1.17 \\ 1.25 \end{pmatrix}$$

$$F_2 = \sqrt{\frac{n}{p}} \times \mathbf{v_2} = \begin{pmatrix} 0.77 \\ -0.01 \\ -0.38 \\ -0.45 \\ -0.24 \\ 0.31 \end{pmatrix}$$

Para obtener las coordenadas de los puntos variables (características), haremos uso de la relación de ambos espacios:

$$\mathbf{G_1} = \mathbf{X}^{t}\,\mathbf{e_1} = \sqrt{\lambda_1}\,\mathbf{v_1}$$

De donde,

$$G_1 = \begin{pmatrix} G_1(\text{elegancia}) \\ G_1(\text{comodidad}) \\ G_1(\text{deportivo}) \end{pmatrix} = \sqrt{2.347}\begin{pmatrix} 0.622 \\ 0.603 \\ 0.505 \end{pmatrix} = \begin{pmatrix} 0.95 \\ 0.92 \\ 0.77 \end{pmatrix}$$

$$G_2 = \begin{pmatrix} G_2(\text{elegancia}) \\ G_2(\text{comodidad}) \\ G_2(\text{deportivo}) \end{pmatrix} = \begin{pmatrix} -0.22 \\ -0.31 \\ 0.64 \end{pmatrix}$$

CASO: CONTAMINACIÓN AMBIENTAL

Datos: Airpollution.txt

Tomado de: Multivariate Analysis of Variance. Johnson&Wichern

Variables

TMR: Tasa de mortalidad total
SMIN: lectura de sulfato quincenal más pequeña
SMEAN: lectura de sulfato quincenal promedio
SMAX: lectura de sulfato quincenal máximo
PMIN: lecturas de partículas suspendidas cada dos semanas - mínimas.
PMEAN: lecturas de partículas suspendidas cada dos semanas - promedio.
PMAX: lecturas de partículas suspendidas cada dos semanas - máximas.
PM2: Densidad de población por milla cuadrada x 0.1
GE65: Porcentaje de POBLACIÓN al menos 65 x 10
PERCWH: Porcentaje de blancos en población
NONPOOR: Porcentaje de familias con ingresos por encima del nivel de pobreza
LPOP: Logaritmo (base 10) de la población x 10.