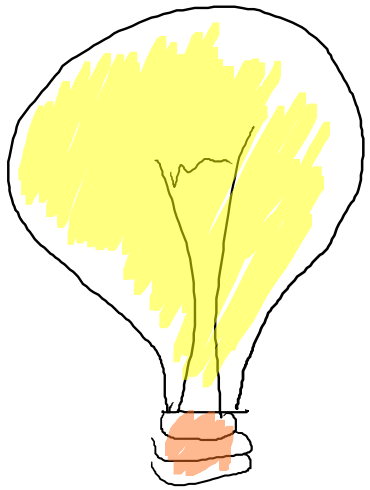
Análisis Multivariado en Ciencias de Datos y Estadística

Instituto Tecnológico Autónomo de México

Primavera 2017

Clase 3: PCA, FA, MCA y FAMD

¿De qué es esta clase?



En este tema vamos a aprender a REDUCIR DIMENSIONALIDAD (LINEALMENTE), que es un paso para resumir información

Usaremos Análisis de Componentes Principales (PCA), Análisis de Correspondencias Múltiples (MCA) y Análisis Factorial para Datos Mixtos (FAMD)

Trabajando con Datos Continuos

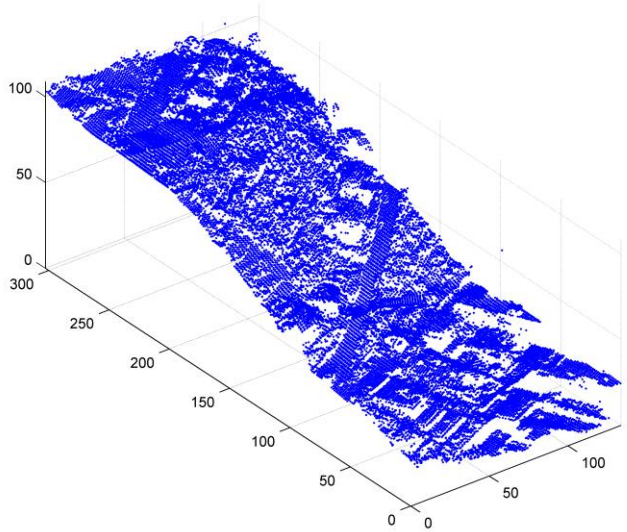
- El modelo con el que trabajamos con datos continuos será la base para trabajar con todo tipo de datos.
- Queremos entender como funcionan los datos **continuos** (de escala de intervalo o de razón)
- La forma de estudiar datos continuos es a través de las **nubes de puntos** (*point clouds)*

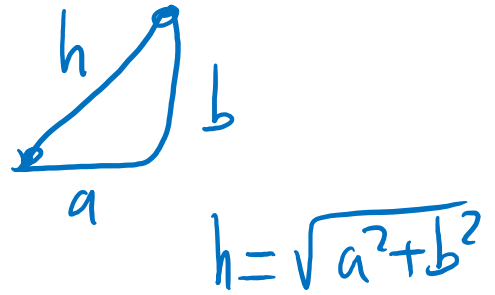
Distancias entre individuos y variables

- Las distancias **entre individuos** se miden usando una métrica, usualmente la **distancia euclidiana** de R^d

$$d(X_{i_1}, X_{i_2}) = \sqrt{\sum_{j=1}^{d} \left(X_{i_1}^j - X_{i_2}^j\right)^2} = \|X_{i_1} - X_{i_2}\|$$

- **Ojo:** medir las distancias **entre variables** como vectores de R^n no **nos va a llevar a nada**.
- Las variables son los **ejes**, lo que necesitamos más bien es el "ángulo" o "correlación" entre ejes



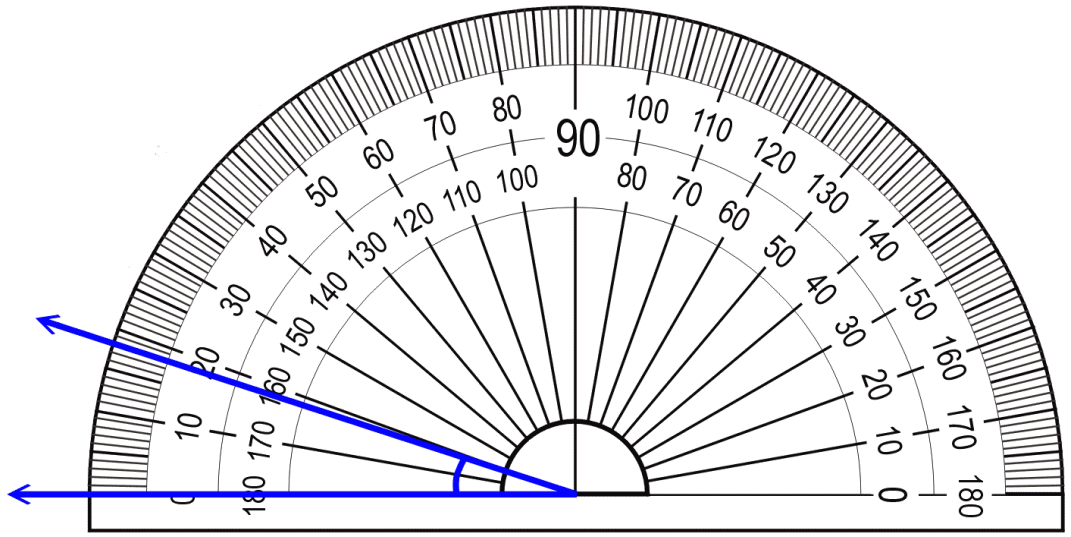


Ángulos

- ¿Cómo medimos ángulos en geometría?

$$\angle\left(X^{j_1}, X^{j_2}\right) = \arccos\left(\frac{\left(X^{j_1}\right)^{\top}\left(X^{j_2}\right)}{\|X^{j_1}\|\,\|X^{j_2}\|}\right)$$

- ¿Qué problemas le ven?

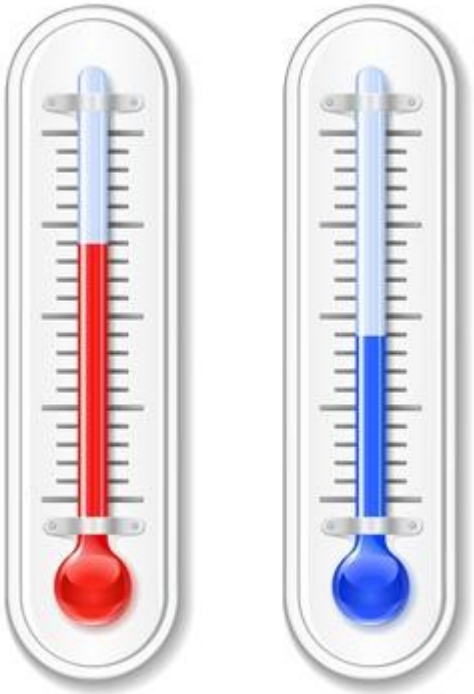


Problemas y ventajas de usar el ángulo

- Es insensible a multiplicar una variable por un escalas $\alpha x \rightarrow x$
- **NO** es insensible a transformaciones escalares $x + \beta \rightarrow x$
- Por ejemplo: no es insensible de pasar de Celsius a

$$Farenheit = 1.8 \times Celsius + 32$$

- ¿Solución?



De ángulos a correlaciones

- La solución es usar una **escala centrada** y restar a todas las variable su promedio; sea $\mu_k = \frac{1}{n}\sum_{i=1}^{n} X_i^{j_k}$ tenemos

$$\angle\left(X^{j_1} - \mu_1, X^{j_2} - \mu_2\right) = \arccos\left(\frac{\left(X^{j_1} - \mu_1\right)^{\top}\left(X^{j_2} - \mu_2\right)}{\|X^{j_1} - \mu_1\| \, \|X^{j_2} - \mu_2\|}\right)$$
$$= \arccos\left(\widehat{Corr}\left(X^{j_1}, X^{j_2}\right)\right)$$

- $\widehat{Corr}(X^{j_1}, X^{j_2})$ es el **estimador empírico** de las correlación si pensamos a los vectores variables como unas muestras independientes de variables aleatorias
- Normalmente nos olvidamos del arcocoseno (es una transformación monótona)
- Moraleja: **Medimos la similitud entre variables con la correlación**

Matriz de Covarianzas Empírica

- **Matriz de Covarianzas**

Es una matriz simétrica

$$Cov(X) = \begin{bmatrix} Var(X^1) & \cdots & Cov(X^1, X^q) \\ \vdots & \ddots & \\ Cov(X^q, X^1) & \cdots & Var(X^q) \end{bmatrix}$$

- **Estimador en Caso centrado**

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & & \\ x_{n1} & \cdots & x_{nq} \end{pmatrix} = (X^1 | \cdots | X^q)$$

Suponer $\overline{X^j} = \frac{1}{n}\sum_{i=1}^{n} x_i^j = 0$ caso centrado.

$$Cov(X) \simeq \frac{1}{n} X^T X = \widehat{Cov}(X)$$

$$= \frac{1}{n}\begin{bmatrix} X^1 \\ \vdots \\ X^q \end{bmatrix}[X^1 | \cdots | X^q] = \frac{1}{n}\begin{bmatrix} (X^1)^T X^1 & (X^1)^T X^q \\ & \ddots \\ (X^q)^T X^1 & (X^q)^T X^q \end{bmatrix}$$

Objetivo: Encontrar K tal que $KX = \left[X^1 - \bar{X}^1 \mid \dots \mid X^d - \bar{X}^d \right]$

Observación $\bar{X}^j = \frac{1}{n}\sum_{i=1}^{n} X_i^j = \frac{1}{n}\mathbb{1}^T X^j$ con $\mathbb{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

$Id = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

- **Matriz centradora**

$$K = \left(Id - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right)$$

$$KX = \left(Id - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right)X = X - \frac{1}{n}\mathbb{1}\mathbb{1}^T X$$

$$= X - \mathbb{1}\left[\frac{1}{n}\mathbb{1}^T X^1 \mid \dots \mid \frac{1}{n}\mathbb{1}^T X^d\right]$$

$$= \left[X^1 - \bar{X}^1 \mid \dots \mid X^d - \bar{X}^d\right]$$

- **Propiedades: Simetría e Idempotencia**

Simetría: $K^T = \left(Id - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right)^T = \left(Id\right)^T - \frac{1}{n}\mathbb{1}^{TT}\mathbb{1}^T = K$

$$K^2 = \left(Id - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right)\left(Id - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right) = K$$

- **Fórmula matricial general para la varianza**

matriz de covarianzas

$$\widehat{Cov}(X) = \frac{1}{n}(KX)^T(KX) = \frac{1}{n}\begin{bmatrix} (KX^1)^T(KX^1) & \cdots & \\ \vdots & \ddots & \\ & & (KX^q)^T(KX^q) \end{bmatrix}$$

$$\frac{1}{n} X^T(K^TK)X$$

K simétrica

$$\frac{1}{n} X^T KKX$$

$$\frac{1}{n} X^T KX$$

Matriz de covarianzas y transformaciones lineales

- ¿Cuántas formas hay de medir?
- **Transformaciones lineales de variables**

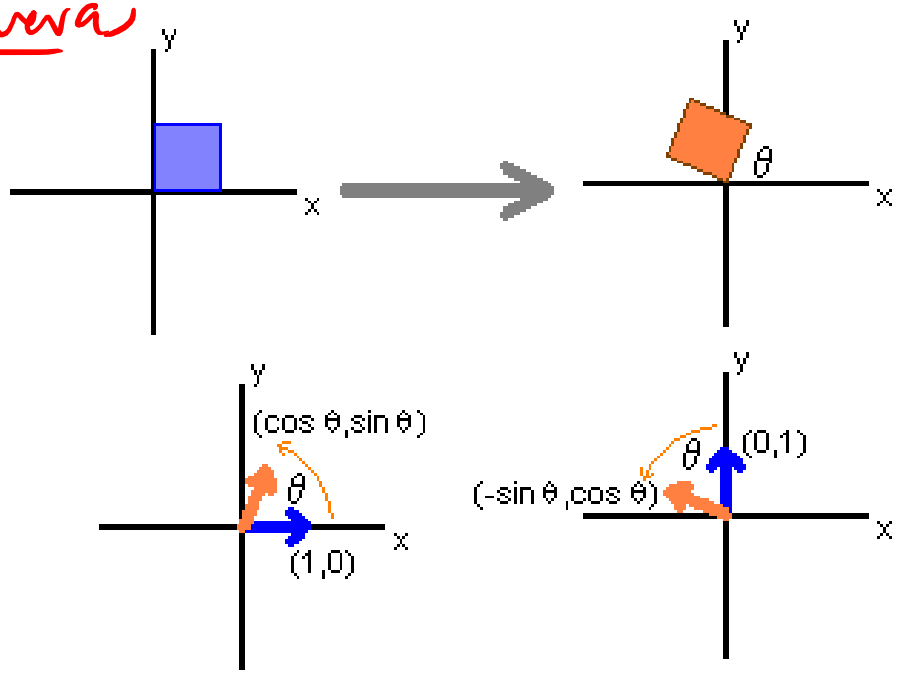
$$Y^j = c_{11}X^1 + \cdots + c_{d1}X^d$$

d variables nueva

Matricialmente

$$Y = XC$$

Donde $C \in R_{d\times d}$ y X la matriz de datos.



$Y = (Y^1 | \dots | Y^d)$

$\| \quad$

- ¿Cuál es la **matriz de covarianzas empírica** de XC?

$$Cov(XC) = \frac{1}{n}(XC)^T K (XC)$$

$\uparrow$ centradora

$$= \frac{1}{n} C^T X^T K X C = C^T Cov(X) C$$

Medidas de disimilitud

¿Cómo medimos distancias?

similitud opuesto disimilitud ≈ distancia

La disimilitud **generaliza** la noción de distancia

Tres propiedades

1. $d(x,x) = 0$
2. $d(x,y) \geq 0 \quad \forall x, y$
3. $d(x,y) = d(y,x)$

} disimilitud

¿Cuáles falta?

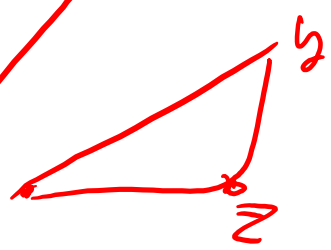
4. (triángulo)

métrica

$d(x,y) \leq d(x,z) + d(z,y)$

ultramétrica

$d(x,y) \leq \max\{d(x,z), d(y,z)\}$



relación

Covarianza y disimilitud

$$disimilitud = 1 - covarianza$$

relación

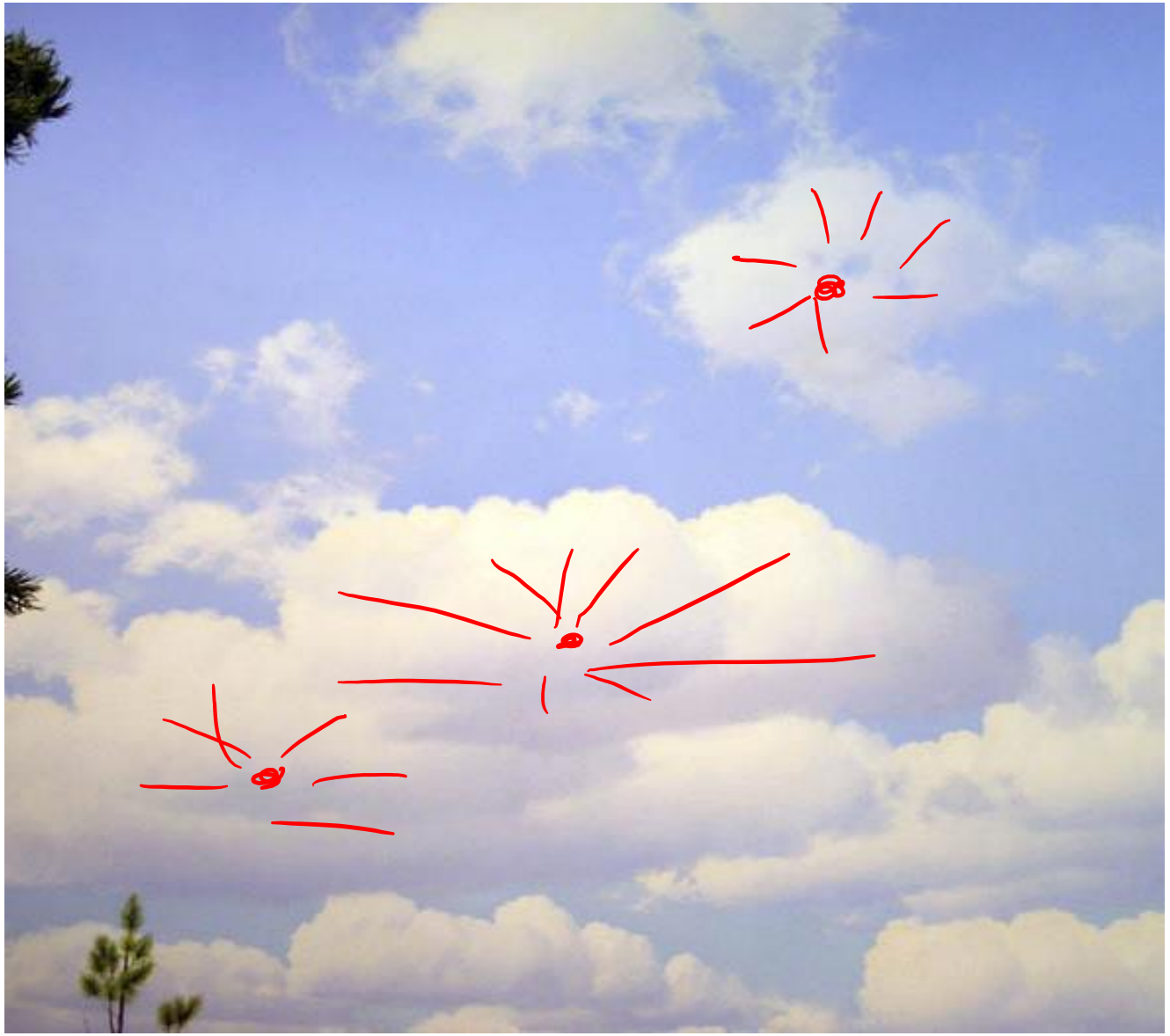
Otras opciones

$$\text{disimilitud} = \sqrt{1 - \text{correlación}^2}$$

La disimilitud de correlación sirve para distancias entre variables

Centro de gravedad e inercia

- ¿Qué tan dispersa es una nube de datos?
- Necesitamos medir la distancia en **CADA DIRECCIÓN** al **CENTRO DE GRAVEDAD**



- El **centro de gravedad** de una nube es el punto

$$G = (\mu_1, \dots, \mu_d)$$

donde $\mu_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ es el promedio de la j-ésima variable

- La **inercia** o **variación total** es

$$\sum_{i=1}^{n} \|X_i - G\|^2 = \sum_{j=1}^{d} \left\|X^j - \mu_j\right\|^2 = \sum_{i=1}^{d} \widehat{Var}\left(X^j\right)$$

$$= \text{"}\textbf{traza}\text{ de matriz de covarianzas"}$$

- De nuevo regresamos a conceptos estadísticos. La inercia es la suma de las varianzas.

$$\|X_i - G\|^2 = \sum_{j=1}^{d} \left(x_i^j - \mu_j\right)^2$$

Discusión

Sea $Y = XC$

- ¿Cómo influye C en la nueva varianza?
- ¿Cuándo crece la inercia?
- ¿Cuándo se contrae?
- ¿Cuándo se mantiene?

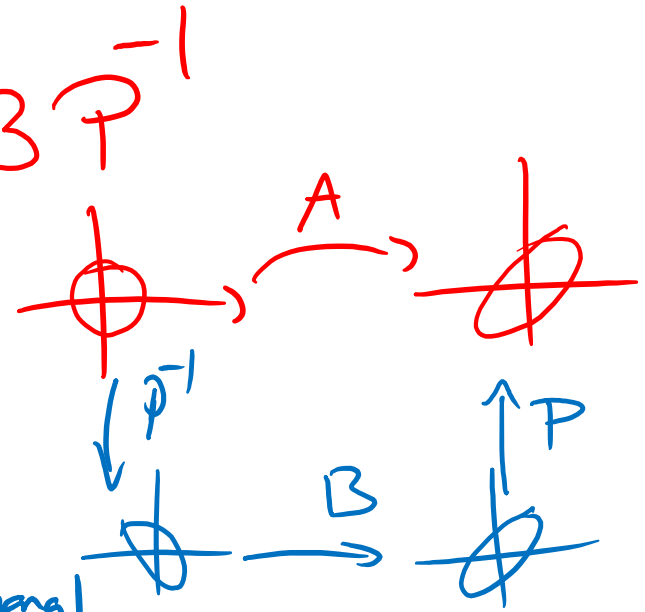
¿Inercia y cambios de coordenadas?

- **ADVERTENCIA:** Álgebra lineal…

La inercia no depende de las coordenadas

- **Matrices similares**

$$A \sim B \iff A = P B P^{-1}$$



- **Diagonalización: eigenvectores y eigenvalores**

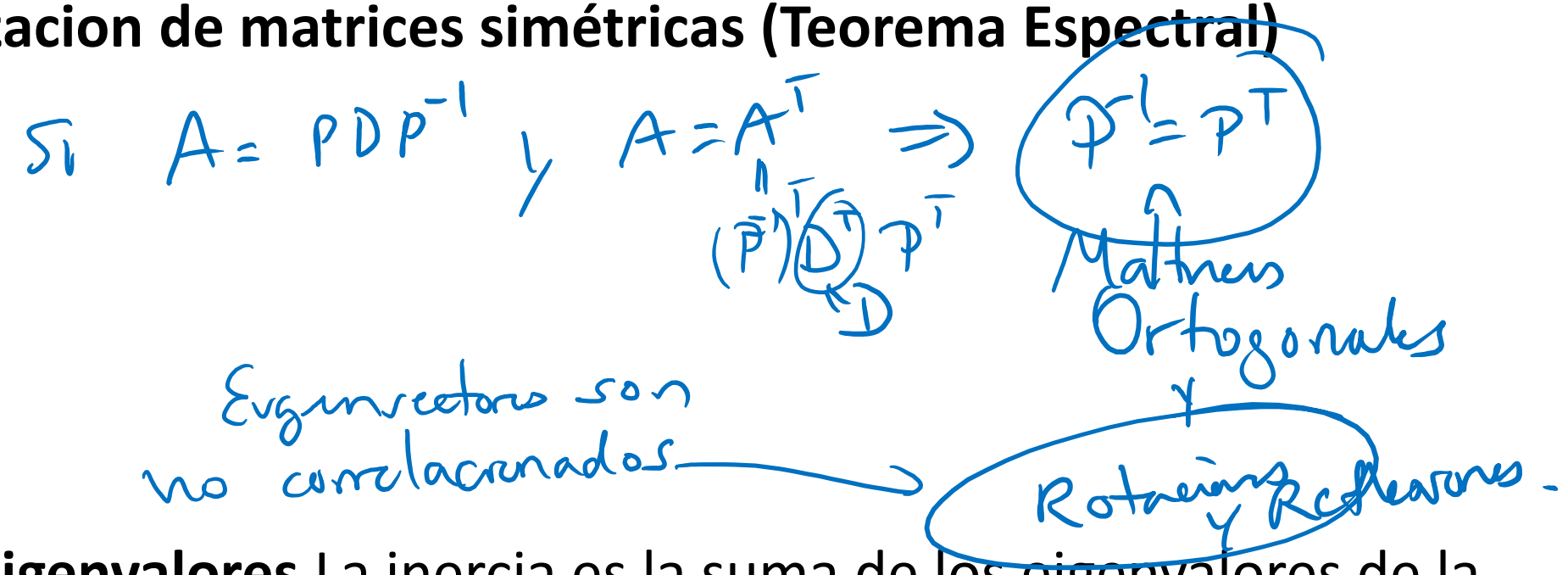
A es diagonalizable $\iff A \sim D$ con D diagonal

$$A = PDP^{-1} = P \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} P^{-1}$$

eigenvectores $[\vec{v}_1 | \dots | \vec{v}_n]$

eigenvalores

- **Diagonalizacion de matrices simétricas (Teorema Espectral)**



- **Inercia y eigenvalores** La inercia es la suma de los eigenvalores de la matriz de covarianzas

Tarea individual: preguntas de investigación y reflexión

- Interpreta la diagonalización como una factorización de tres transformaciones lineales, ¿cómo son estas transformaciones?
- Interpreta el Teorema Espectral para matrices simétricas usando lo anterior
- ¿Cómo visualizas los eigenvectores y eigenvalores de una matriz simétrica? ¿Qué tiene que ver con las formas cuadráticas?
- ¿Cómo visualizas una matriz de covarianzas en una nube de puntos y cómo visualizas los eigenvectores?

  - Escribir 1 a 2 cuartillas para entregar **impreso** en clase.

a mano.

Traza

$$tr(A) = \sum_{i=1}^{d} a_{ii} \qquad A \in \mathbb{R}^{d \times d}$$

$$\text{Inercia}(X) = tr(\widehat{Cov}(X)) = \sum_{j=1}^{d} Var(X^j)$$

Propiedades de la traza

$$tr(ABC) = tr(BCA) = tr(CAB)$$

En tanto

Si $A = PDP^{-1}$

$$tr(A) = tr(PDP^{-1}) = tr(DP^{-1}P) = tr(D)$$

En tanto

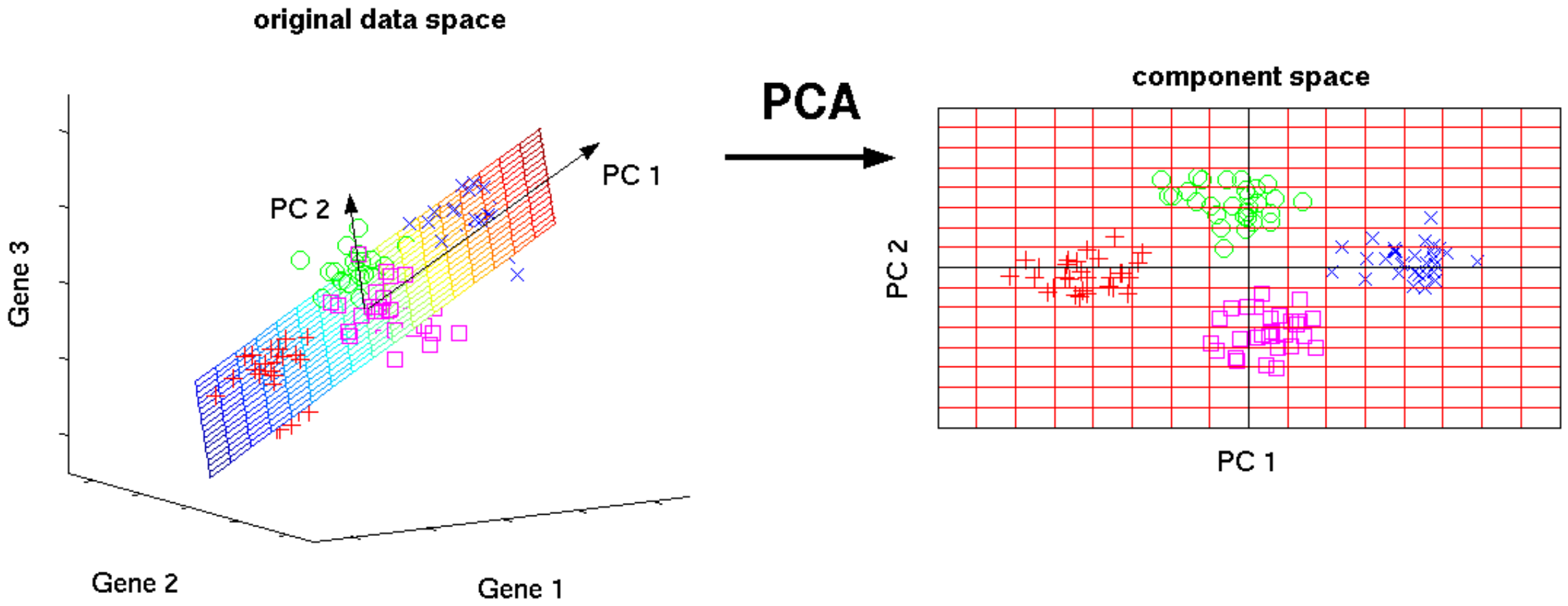
$$tr(A) = \sum \lambda_i$$

Conclusión:

$$\text{Inercia}(X) = tr(\widehat{Cov}(X)) = \sum \lambda_i$$

λ_i eigen de $Cov(X)$

II. PCA: Análisis de Componentes Principales



Motivación

- Datos redundantes
- Duplicación de la información entre variables

Objetivo del PCA

- Proporcionar un conjunto de variables no correlacionadas que contengan casi la misma información que todas las variables redundantes

Varianza como medida de información

- Una analogía de nubes: ¿Qué nube es mas importante? Más varianza es más información

Subespacios y Proyecciones lineales

I. Viviendo en las nubes…

Los datos numéricos se representan como **NUBES DE DATOS,** que son matrices $X \in R_{n \times d}$ con n individuos y d variables donde:

- Cada individuo X_i es un vector en R^d
- Cada individuo es un punto de la nube