MACHINE LEARNING INMERSION

ANDRÉ OMAR CHÁVEZ PANDURO

Perú

« Si el **Plan** no funciona ,

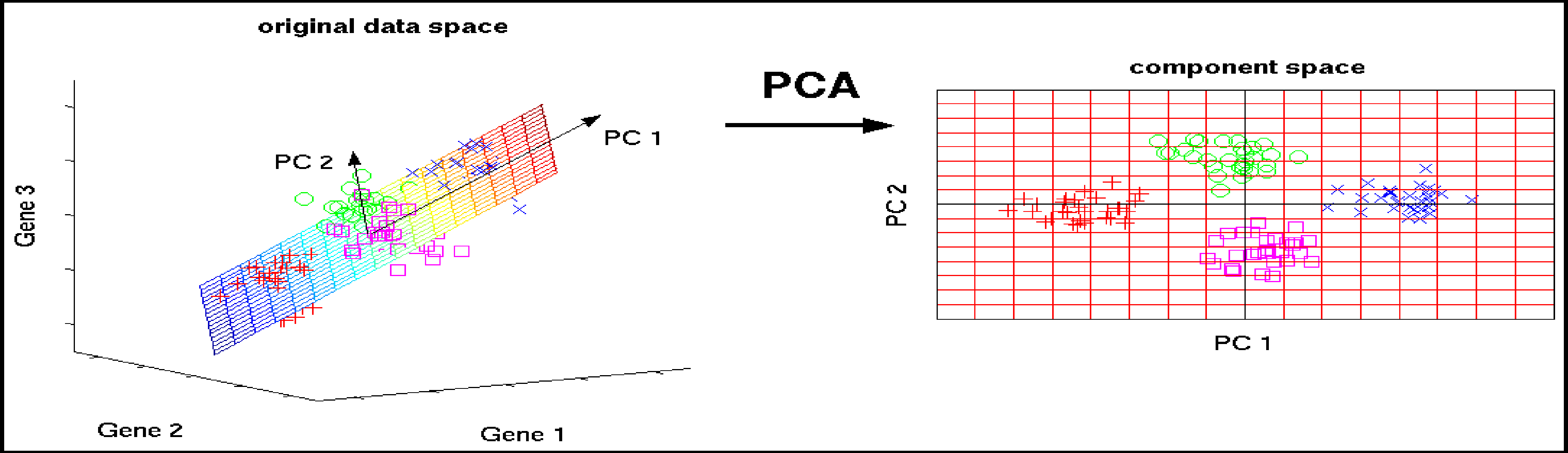
cambia el **Plan** pero **No Cambies** los

OBJETIVOS NI METAS»

AGENDA

- Reducción de Dimensionalidad
- Análisis de Componentes Principales vs Análisis Factorial.

Reducción de la Dimensionalidad

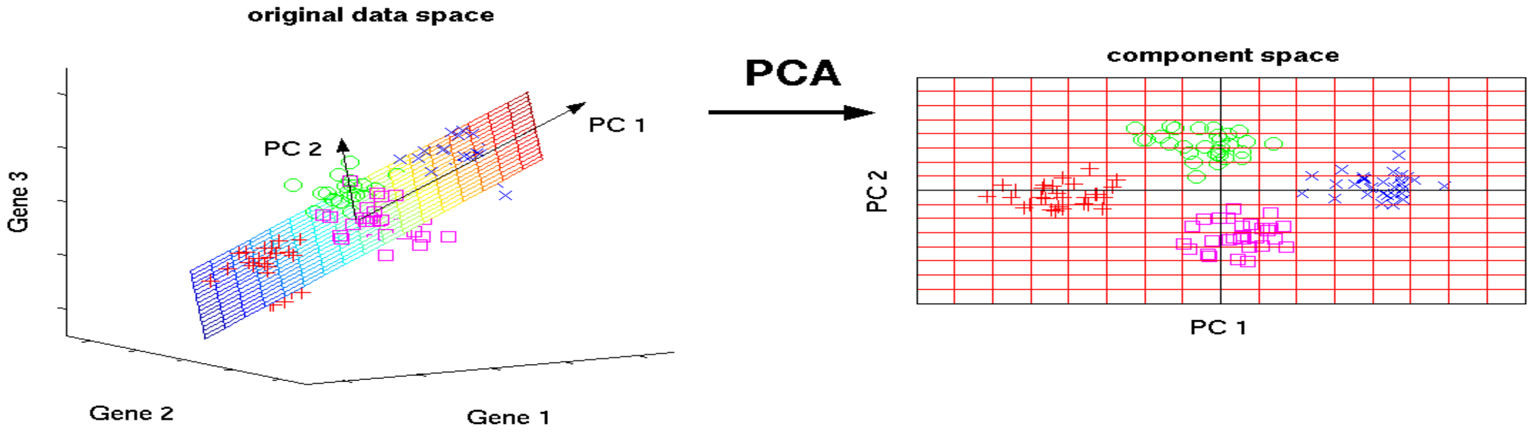


AED : REDUCCIÓN DE LA DIMENSIONALIDAD

- En el Análisis de Datos nos podemos encontrar con una gran cantidad de dimensiones, tanto en número de registros como en número de variables.

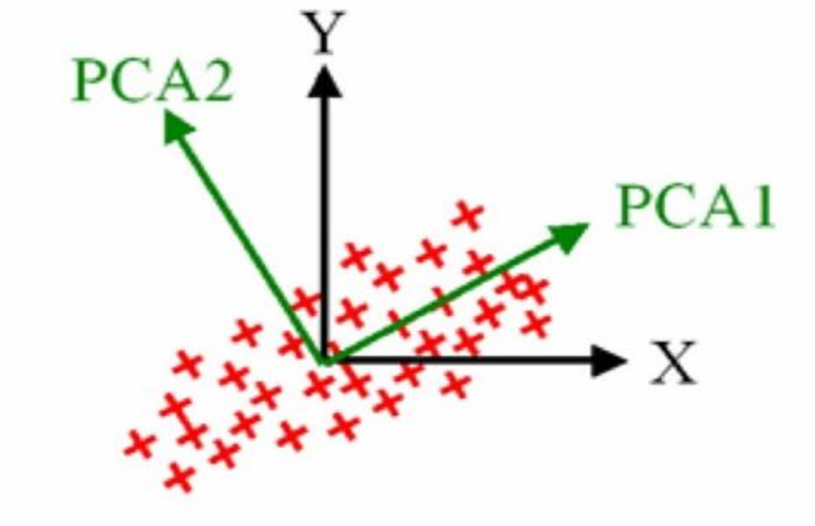
- La dimensionalidad de la información dificulta el procesamiento de algoritmos, nos quita interpretabilidad de la información y esconde relaciones existentes entre las variables.

- Una técnica que nos ayuda con lo antes descrito es el ANÁLIS DE COMPONENTES PRINCIPALES.



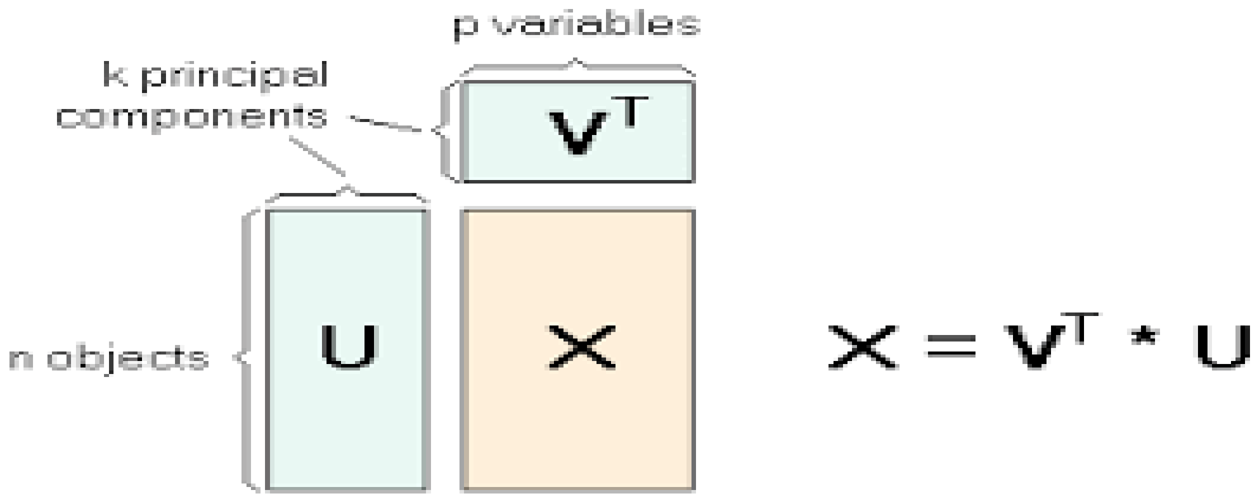
COMPONENTES PRINCIPALES

- **Karl Pearson**



OBJETIVO

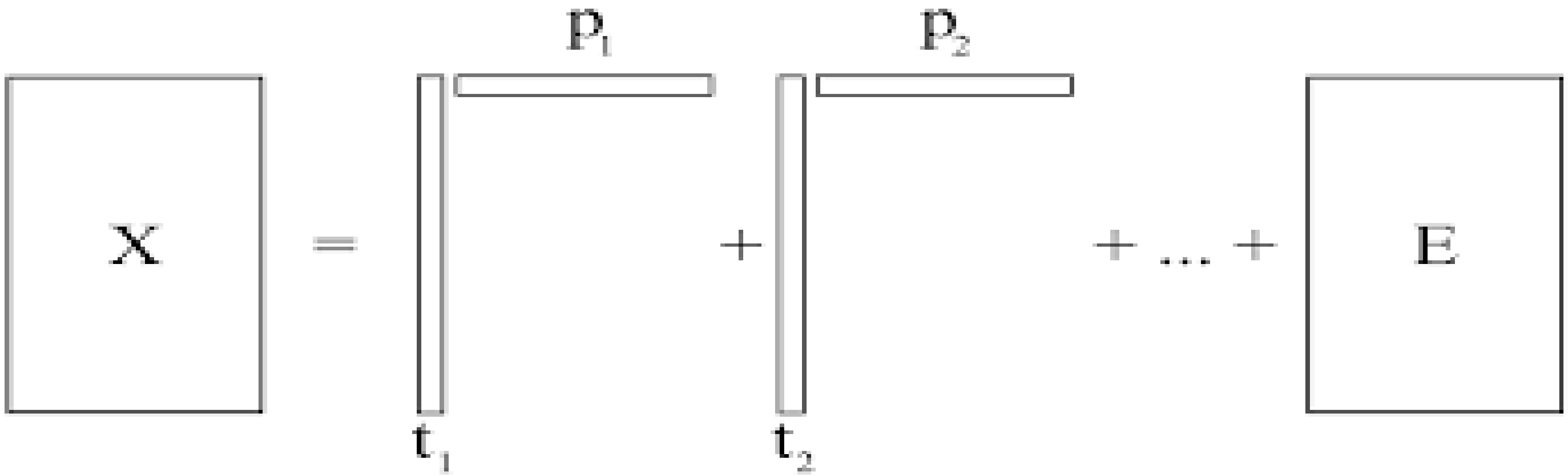
- Objetivo: Dada una matriz de datos de dimensiones nxp que representa los valores de p variables en n individuos, investigar si es posible representar los individuos mediante k variables (k<p) con poca (o ninguna si es posible) pérdida de información.



OBJETIVO

Nos gustaría encontrar nuevas variables Z, combinación lineal de las X originales, tales que:

- k de ellas contengan toda la información
- las restantes p-k fuesen irrelevantes



CUÁLES SON LAS RAZONES PARA UTILIZAR EL ANÁLISIS DE COMPONENTES PRINCIPALES?

- ✓ Es el método más útil para depurar datos multivariados, se recomienda como un paso previo antes de aplicar algoritmos de clasificación.

- ✓ Ayuda a localizar datos atípicos o discordantes.

- ✓ Se utiliza para ayudar a formar grupos de unidades experimentales o individuos con características similares.

- ✓ Sí, en un caso de aplicación regresión múltiple, las variables regresoras están intensamente correlacionadas (multicolinealidad), el ACP puede ayudar a resolver este problema.

SUPUESTOS

- Variables métricas (escala intervalo o razón)
- Las Variables deben estar correlacionadas.
- Linealidad entre las variables.
- No presencia de datos discordantes en la medida de lo posible.
- Normalidad multivariada, sí se desea aplicar inferencia multivariada.

BASE MATEMÁTICA

- Sí $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$, **A** es ortogonal, es decir $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, la distancia al origen no cambia:
  - $\mathbf{y}^T_i\,\mathbf{y}_i = (\mathbf{A}\mathbf{x}_i)^T(\mathbf{A}\mathbf{x}_i) = \mathbf{x}_i^T\mathbf{x}_i$

- Así, una matriz ortogonal transforma $\mathbf{x}_i$ en $\mathbf{y}_i$ a un punto que tiene la misma distancia desde el origen, y los ejes son efectivamente rotados.

- Encontrar los ejes del hiperelipsoide es equivalente a encontrar la matriz ortogonal **A** que rota los ejes como la extensión natural de la nube de puntos para las nuevas variables incorrelacionadas. La media muestral de la matriz de covarianzas de $\mathbf{y}_i$:

BASE MATEMÁTICA

$$\Sigma_{\mathbf{y}} = \mathbf{A}\Sigma\mathbf{A}^T = \begin{bmatrix} \sigma_{y_1}^2 & 0 & \ldots. & 0 \\ 0 & \sigma_{y_2}^2 & 0\ldots. & 0 \\ . & . & \ldots\ldots & . \\ 0 & 0 & \ldots\ldots & \sigma_{y_p}^2 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \ldots. & 0 \\ 0 & \lambda_2 & 0\ldots. & 0 \\ . & . & \ldots\ldots & . \\ 0 & 0 & \ldots\ldots & \lambda_p \end{bmatrix}$$

Donde λ_i son los autovalores de $\mathbf{\Sigma}$:

- La matriz ortogonal **A** que diagonaliza a **Σ:**

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ . & . & \cdots & . \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}$$

Las componentes principales son combinación lineal de las variables originales:

$$Y_1 = a_{11}\overline{X}_1 + a_{12}\overline{X}_2 + \cdots + a_{1k}\overline{X}_k$$

$$Y_2 = a_{21}\overline{X}_1 + a_{22}\overline{X}_2 + \cdots + a_{2k}\overline{X}_k$$

$$\cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots$$

$$Y_m = a_{m1}\overline{X}_1 + a_{m2}\overline{X}_2 + \cdots + a_{mk}\overline{X}_k$$

- Una aproximación algebraica de las componentes principales se puede describir brevemente como la búsqueda de una combinación lineal con una varianza máxima.

$$\lambda = \frac{\mathbf{a}^T \Sigma \mathbf{a}}{\mathbf{a}^T \mathbf{a}}$$

- El máximo valor de λ, está dado por el mayor autovalor, sí:

$$(\Sigma - \lambda \mathbf{I}_p)\mathbf{a} = 0, \quad \mathbf{a}^T \mathbf{a} = 1$$

- El autovector $\mathbf{a}_1$ corresponde al mayor autovalor y así sucesivamente.

Número Optimo de Componentes Principales

- Criterio de **Kaiser**, indica que se retendrán aquellas componentes principales, cuyos **autovalores** son mayores que el promedio de todos los **autovalores** de la matriz de covarianza o de la matriz de correlaciones.

- El número de componentes elegidas son aquellas cuya calidad global de la representación (**Proporción de la variación total explicada** por las "r" primeras CP) sea aproximadamente al menos el 80% .

- Criterio de Catell, construir un gráfico del número de CP y sus correspondientes autovalores . El punto de inflexión en el gráfico, indicará el número de componentes principales a ser retenidas. Se denomina gráfico de **sedimentación.**

GRACIAS
POR SU ATENCIÓN