Universidad Ricardo Palma

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

TALLER DE ESTADÍSTICA PARA LA CIENCIA DE DATOS

R + Python

MÓDULO 4 : ANALISIS MULTIVARIADO I

Análisis de Componentes Principales y Factorial

A nuestro recordado Maestro

Dr. Erwin Kraenau Espinal, Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos

TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

TALLER DE ESTADÍSTICA PARA CIENCIA DE DATOS

EXPOSITORES

José Antonio Cárdenas Garro
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricardo Palma

Predictive Modelling Specialist — Scotiabank

Portfolio and Consumption Analyst — Scotiabank

Data Analyst — ONP Oficina de Normalización Previsional

Correo : josecardenasgarro@gmail.com
LinkedIn : www.linkedin.com/in/jos%C3%A9-antonio-c%C3%A1rdenas-garro-599266b0

André Omar Chávez Panduro
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricardo Palma

Data Scientist — Interbank

Customer Intelligence Analyst — Claro

Data Analyst — INEI Instituto Nacional de Estadística e Informática

Correo : andrecp38@gmail.com / 09140205@unmsm.edu.pe
LinkedIn : www.linkedin.com/in/andré-chávez-a90078b9

TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

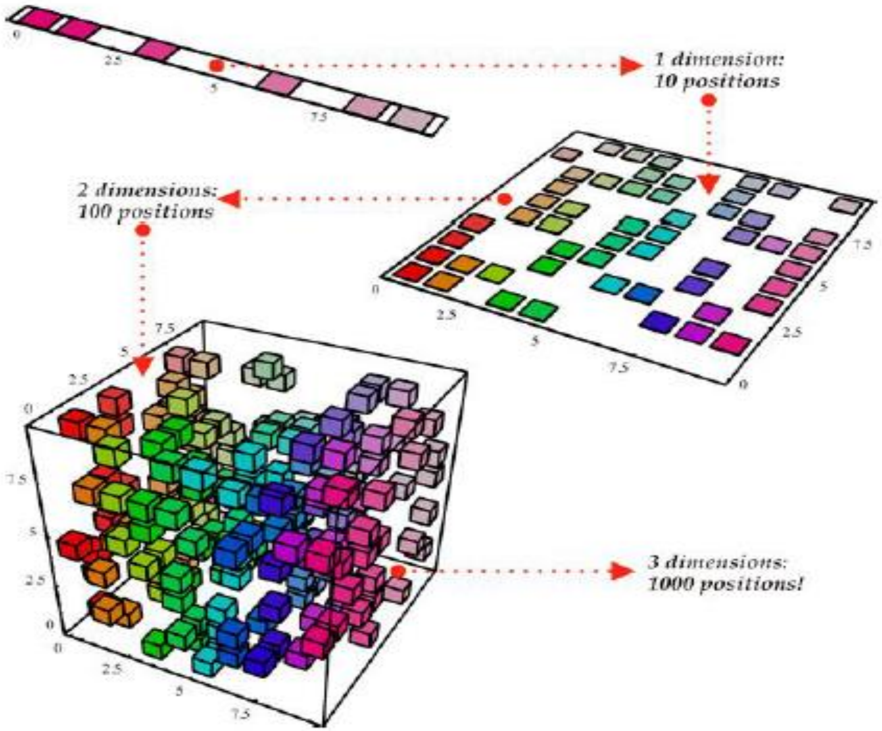
« Divide las dificultades que examinas en tantas partes como sea posible , para su mejor solución»

AGENDA

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

Reducción de Dimensionalidad

- La "Maldición de la Dimensionalidad".



Análisis factorial vs Componentes principales

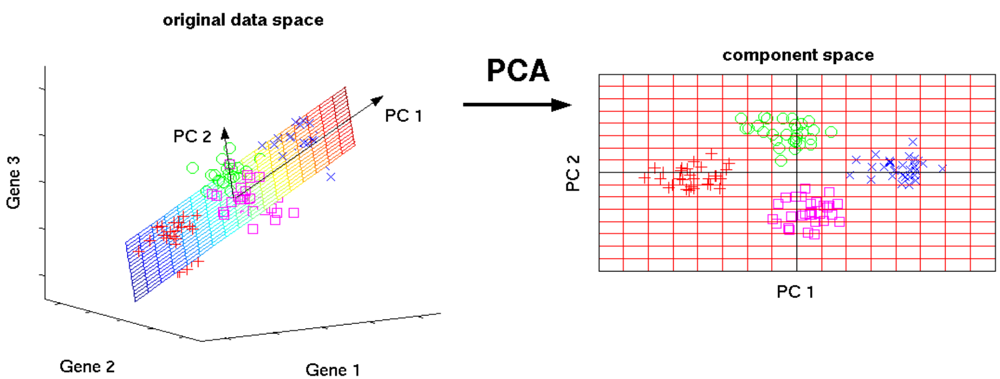
- *Factores latentes vs Reducción de variables*
- *Diferencia matemática basada en la inclusión de un componente aleatorio de variabilidad que es único por cada factor*
- *Componentes principales tiene solución analítica- Factorial no.*
- *En N grande la diferencia va desapareciendo*

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.

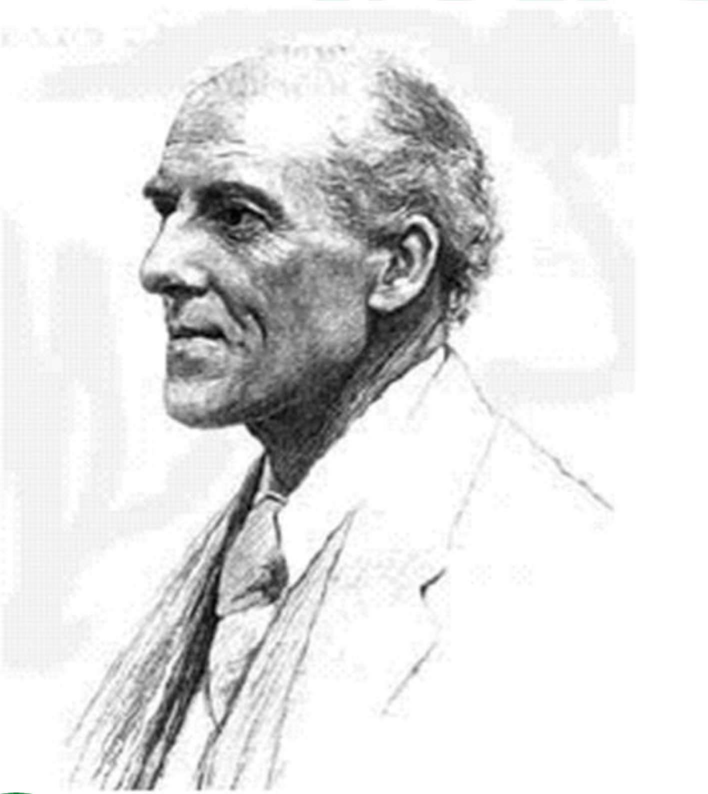
- La dimensionalidad de la información dificulta el procesamiento de algoritmos, nos quita interpretabilidad de la información y esconde relaciones existentes entre las variables.

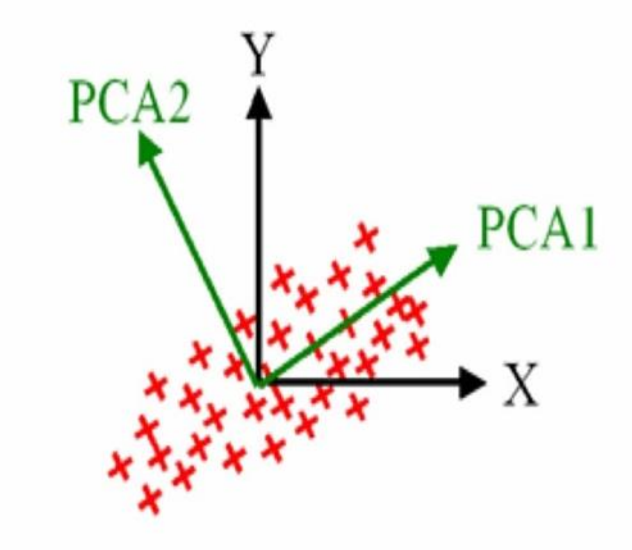
- Una técnica que nos ayuda con lo antes descrito es el ANÁLIS DE COMPONENTES PRINCIPALES.



COMPONENTES PRINCIPALES

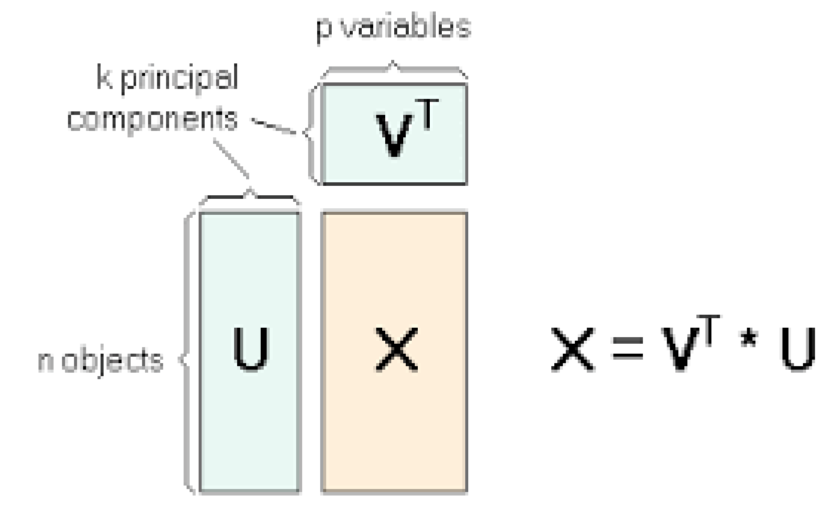
- **Karl Pearson**





OBJETIVO

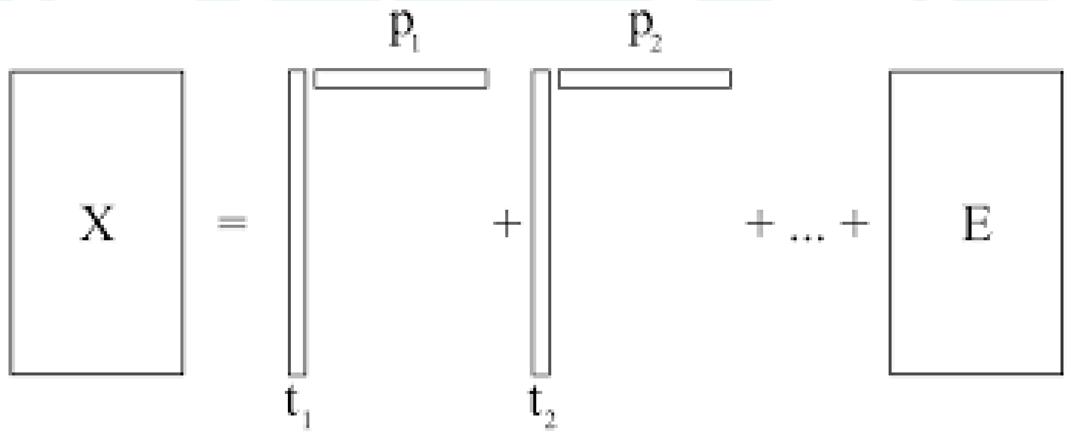
- Objetivo: Dada una matriz de datos de dimensiones nxp que representa los valores de p variables en n individuos, investigar si es posible representar los individuos mediante k variables (k<p) con poca (o ninguna si es posible) pérdida de información.



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 $y_i = Ax_i$, **A** es ortogonal, es decir $A^T A = I$, la distancia al origen no cambia:
- $y^T_i\, y_i = (Ax_i)^T (Ax_i) = x_i^T x_i$
- Así, una matriz ortogonal transforma x_i en 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 y_i :

BASE MATEMÁTICA

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

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

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

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

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

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 de Componentes Principales a retener

- 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.**

Utilidad

- Reducción de dimensiones (trabajar con menos variables, específicamente de p a k, donde k<p).
- Los Componentes Principales son ortogonales entre sí, es decir que son variables incorrelacionadas entre sí.
- Los componentes son una combinación lineal de las variables originales.
- La transformación se realiza multiplicando la matriz original de datos por la matriz de cargas, lo que da origen a los componentes.
- Se acostumbra graficar las dos o tres primeras componentes para tener una mejor visualización de los datos.

Sugerencias

- Es una buena práctica aplicar el ACP antes de la elaboración de un modelo de segmentación o un modelo predictivo. Permite saber si existe mucha correlación entre las variables originales.
- Si la varianza explicada acumulada crece lentamente, significa que la correlación entre las variables no es muy alta, por lo tanto no conviene aplicar el ACP.
- El ACP nace orientado al tratamiento de variables cuantitativas continuas, sin embargo muchos investigadores transforman escalas Likert en números de 1 a 5 para poder aplicar el ACP e identificar preguntas relacionadas en el análisis de encuestas. Esto no es recomendable, en estos casos se sugiere realizar el Análisis Factorial de Correspondencias.

¡Gracias!

TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”