MACHINE LEARNING: ALGORITMOS NO SUPERVISADOS

Análisis de Componentes Principales

Instructor: José Nelson Zepeda

San Salvador, Diciembre 2018

ACP

Concepto

Objetivo

Metodología

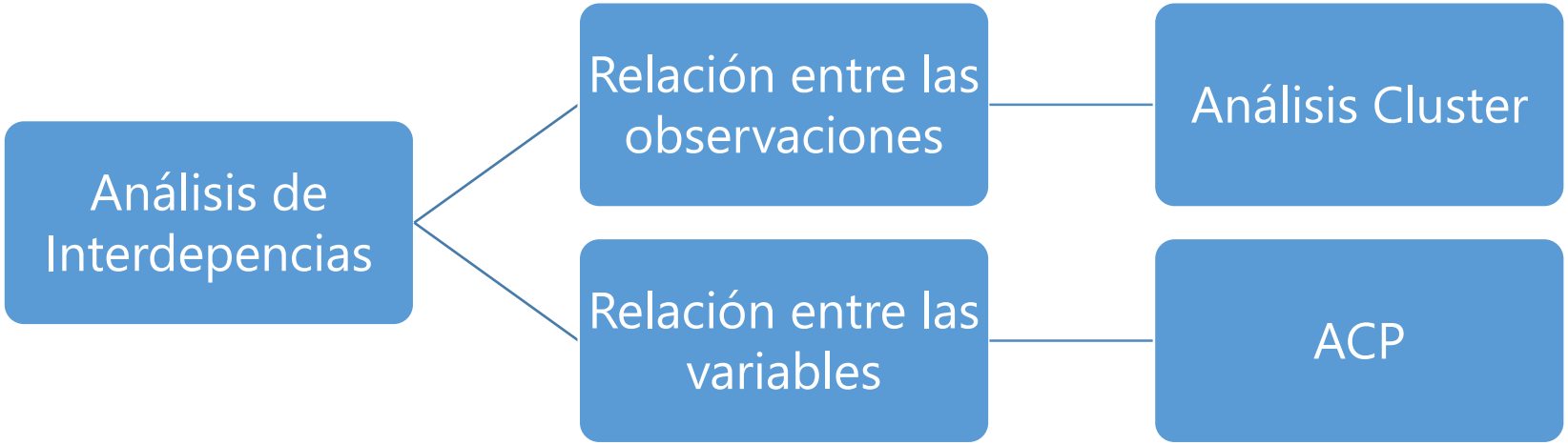
Definiciones

Gráfica de Sedimentación

Conceptos Básicos

¿Qué es ACP?

El análisis de componentes principales se ha convertido en una poderosa herramienta para la investigación científica, construcción de modelos, revisión de escenarios y sintetización debido a que permite analizar conjuntamente un número grande de variables.

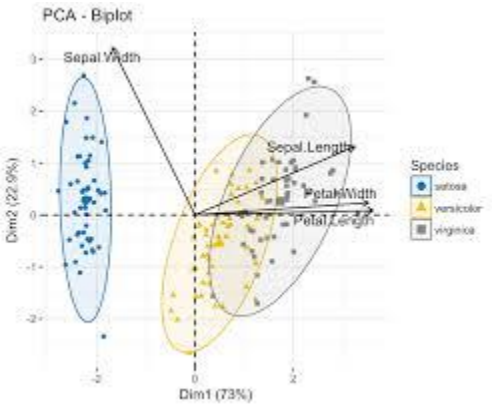


Objetivo del ACP

El objetivo del ACP es transformar un conjunto de variables (variables originales) en un nuevo conjunto de variables (componentes principales), in-correlacionadas entre sí.

La ventaja principal de este tipo de análisis es su capacidad de "acomodar" las variables utilizadas en una investigación, con el fin de comprender las relaciones complejas.

- La explicación de fenómenos cuya información se cifra en muchas variables más o menos correlacionadas.
- Reducir la dimensión del número de variables inicialmente consideradas en el análisis.
- Se pueden ordenar las nuevas variables de acuerdo a la información que llevan.



Existen 2 principios fundamentales para el análisis PCA:

- ACP es para datos cuantitativos y no es necesario establecer jerarquías ni comprobar la normalidad.
- Si las variables originales no están correlacionadas, el análisis tendrá muy poco valor.

Metodología del ACP

La metodología de los Componentes Principales busca unas pocas combinaciones lineales de las variables observables, que puedan utilizarse para resumir los datos, perdiendo la menor cantidad de información posible.

1. Se parte de una matriz de datos

Ind	X_1	X_2	X_p
1	x_{11}	x_{12}	x_{1p}
2	x_{21}	x_{22}	x_{2p}
...			
n	x_{n1}	x_{n2}	x_{np}

2. Se calcula la matriz de convarianza

$$C=\begin{bmatrix} c_{11} & c_{12} & c_{1p} \\ c_{21} & c_{22} & c_{2p} \\ & & \\ c_{n1} & c_{n2} & c_{np} \end{bmatrix}$$

3. Se calculan los componentes principales

$$Z_i = a_{i1}\,x_1 + a_{i2}\,x_2 + \ldots + a_{ip}\,x_p$$

el (ACP) presenta múltiples ventajas pues es una técnica que reduce la dimensionalidad de un conjunto de datos multivariados, remueve las interrelaciones existentes entre variables y organiza los datos en forma de vectores ortogonales en donde cada una de las variables dentro del vector se comporta en forma similar en base a sus correlaciones; a cada uno de estos vectores se le llama **componente principal**.

Grado de Correlación

Para identificar el grado de correlación entre variables es posible utilizar el Test de Esfericidad de Barlett.

En este test, para comprobar que las correlaciones entre las variables son distintas de cero de modo significativo, se comprueba si el determinante de la matriz es distinto de 1, es decir si la matriz de correlaciones es distinta de la matriz unidad.

Si las variables están correlacionadas, habrán muchos valores altos en valor absoluto fuera de la diagonal principal de la matriz de correlaciones, además, el determinante será menor que 1 (El valor máximo del determinante es 1 si las variables están incorreladas).

El **test de Barlett** realiza el contraste: $\begin{cases} H_0 : |R| = 1 \\ H_1 : |R| \neq 1 \end{cases}$

Conceptos Adicionales

•Comunalidad: Es la cantidad de varianza que una variable comparte con las demás variables consideradas.

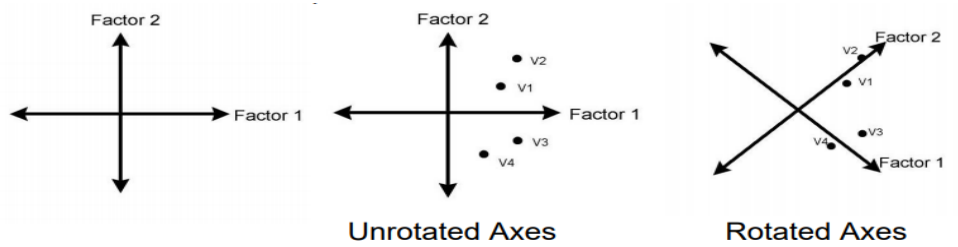
$$h_k^2 = F_{1j}^2 + F_{2j}^2 + \ldots + F_{kj}^2 = \sum F_{kj}^2$$

•Eigenvalores: Indican la proporción de la varianza total de una variable explicada por ese factor.

•Factor: combinación lineal de las variables originales

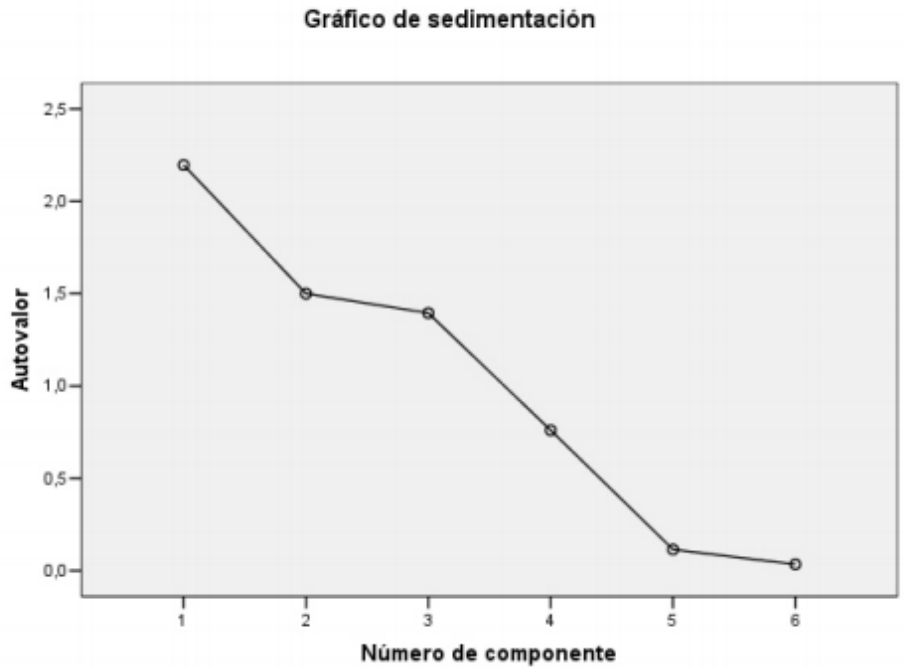
•Loadings: Correlación entre las variables originales y los factores.

•Rotación de Factores: proceso de ajuste de los ejes de los factores con el fin de obtener un factor de mayor significancia.



Conceptos Adicionales

- Varimax: Método de rotación ortogonal que minimiza el número de variables que tiene saturaciones altas en cada factor.
- Quartimax: Método de rotación ortogonal que minimiza el número de factores necesarios para explicar cada variable.
- Equamax: Método de rotación que combina el método Varimax y Quartimax.
- Gráfico de sedimentación: Muestra la representación gráfica de la tendencia que sirve de regla para la determinación del número de factores óptimos.



Bibliografía

A Tutorial on Principal Component Analysis

By Lindsay I Smith, 2002

Componentes Principales

By Santiago de la Fuente Fernández, 2011

R Data Analysis Cookbook

by Kuntal Ganguly,2017