Curso

3

MACHINE LEARNING & BIG DATA

Unsupervised Learning: Principal Components Analysis

Algoritmos No-Supervisados

JOSÉ NELSON ZEPEDA DOÑO

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.
Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.

Nelson Zepeda
MIP • V 1.0
San Salvador El Salvador
Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

Capítulo 1

Análisis de Componentes Principales

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.

El análisis multivariado se refiere al análisis simultaneo de dos o más variables medidas en cada individuo u objeto de análisis.

Uno de los objetivos del cálculo de componentes principales es la identificación de los mismos, es decir, averiguar qué información de la muestra resumen. Sin embargo este es un problema difícil que a menudo resulta subjetivo. Habitualmente, se conservan sólo aquellos componentes que recogen la mayor parte de la variabilidad, hecho que permite representar los datos según dos o tres dimensiones si se conservan dos o tres ejes factoriales, pudiéndose identificar entonces grupos naturales entre las observaciones.

Historia y Concepto

Estas técnicas fueron inicialmente desarrolladas por Pearson a finales del siglo XIX y posteriormente fueron estudiadas por Hotelling en los años 30 del siglo XX. Sin embargo, fue hasta la aparición de los ordenadores que se empezaron a popularizar.

Recientemente, la estadística multivariada experimentó una utilización creciente en todas las áreas de investigación. La causa es que, conforme aumenta el número de variables necesarias, existe una necesidad mayor de conocer en profundidad la estructura y las interrelaciones entre ellas.

Luego de los primeros párrafos, vale la pena mencionar que el ACP da respuesta a muchas interrogantes por medio del descubrimiento de la verdadera dimensionalidad de los datos.

Cuando se determina la dimensionalidad y es menor que p dimensiones, las p variables originales se pueden remplazar por un número menor de variables subyacentes, sin que se pierda información.

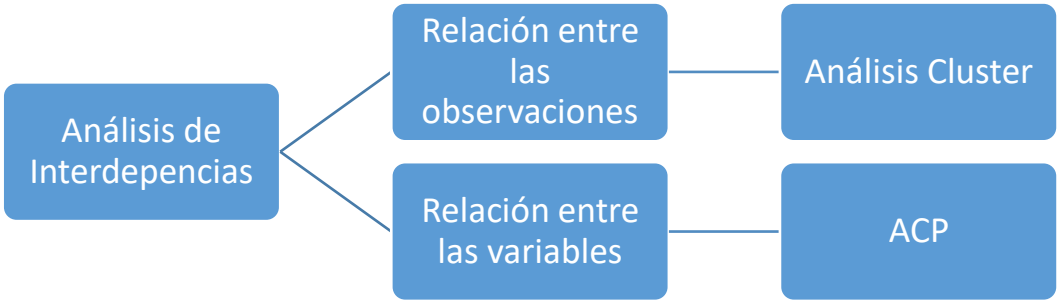

Figura 1-1 Análisis Interdependencias

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.

A continuación se presentan los principales puntos de interés:

- 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.

Ahora bien, 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.

Las nuevas variables generadas a partir del ACP son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.

$$var(Z_1) \geq var(Z_2) \geq \ldots \geq var(Z_p)$$

De modo ideal, se buscan m<p variables que sean combinaciones lineales de las p originales y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.

Si las variables originales están incorreladas de partida, entonces no tiene sentido realizar un análisis de componentes principales ya que los factores coincidirán con las variables originales, pero si las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se podrá explicar con muy pocos componentes (factores).

Es vital aclarar que el análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes

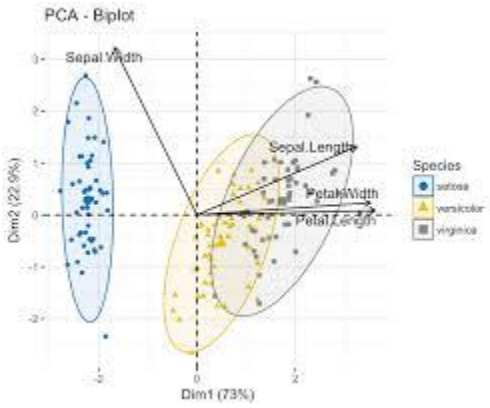

Figura 1-2 Análisis grafico ACP

Por otro lado, si las variables originales están incorreladas de partida, entonces no tiene sentido realizar un análisis de componentes principales.

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; es decir, que expliquen las diferencias entre los individuos, casi con la misma efectividad que toda la base de datos, y sean no correlacionadas, para no reiterar información.

Actualmente hay una fuerte tendencia entre los investigadores a dar significado a las variables componentes principales recién creadas. Si las interpretaciones son obvias, entonces se debe seguir adelante y usarlas, sin embargo, esos pocos casos en donde a las componentes principales se les puede dar una interpretación pueden considerarse como un premio, porque lo común es no esperar que se puedan interpretar las variables componentes principales.

Con respecto al número de componentes que se deben retener, existen varios criterios. Uno de éstos es el de la media aritmética, que plantea seleccionar aquellas componentes cuya raíz característica excedan la media de las raíces características.

Cuando las variables están tipificadas, se seleccionan aquellas componentes que tienen raíz característica mayor que 1.

Pasos del ACP

Los pasos principales son:

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

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

donde c_{ii} es la varianza de X_i

y c_{ij} es la covarianza de $X_i X_j$

3. Se calculan los componentes principales

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

Previamente mencionamos que 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.

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: $\begin{cases} H_0 : |R| = 1 \\ H_1 : |R| \neq 1 \end{cases}$

Definiciones 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.

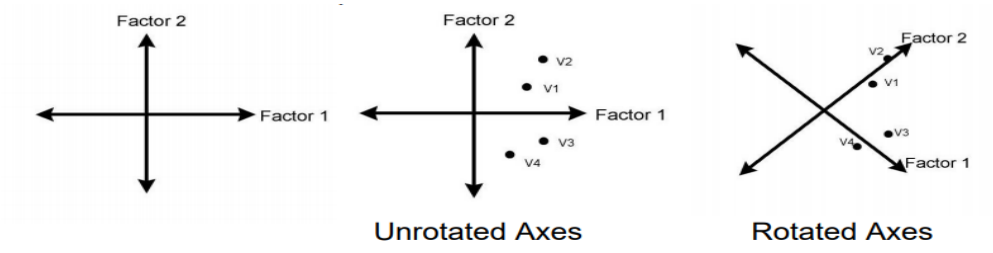

Figura 1-3 Rotación de Factores en ACP

- Gráfico de sedimentación: Muestra la representación gráfica de la tendencia que sirve de regla para la determinación del número de factores óptimos.

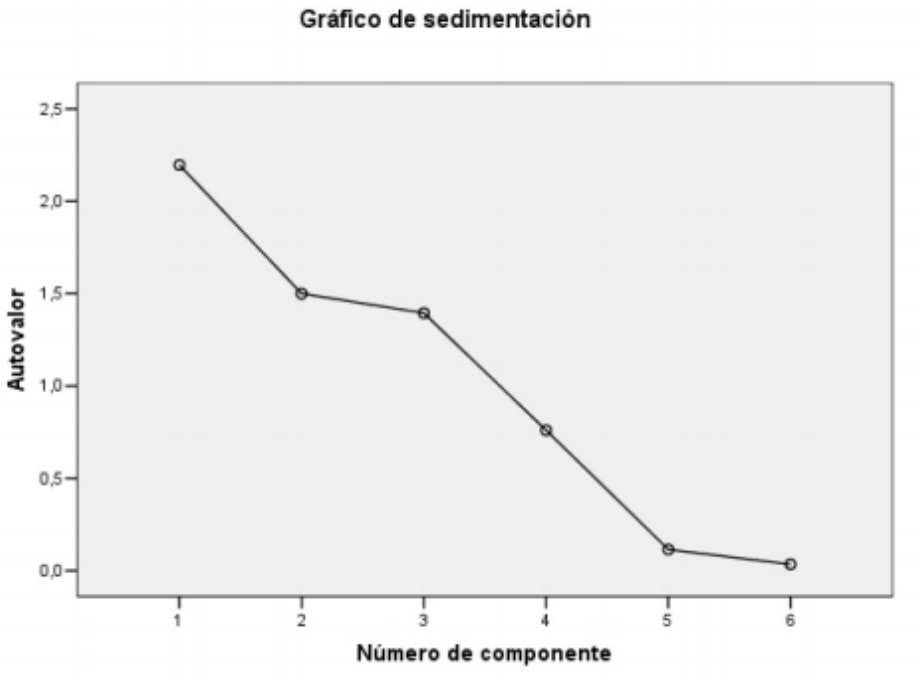

Figura 1-4 Gráfico de sedimentación

- 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.

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