Análisis de Componentens Principales

Objetivo: Reducir la dimensión de un conjunto de variables, conservando la mayor cantidad de información que sea posible.

Es la transformación a un nuevo conjunto de variables que son no correlacionadas y se ordenan de modo tal que unas pocas retengan la mayor cantidad de variación presente en el conjunto original de variables.

Análisis de Componentens Principales

Objetivo: Reducir la dimensión de un conjunto de variables, conservando la mayor cantidad de información que sea posible.

Es la transformación a un nuevo conjunto de variables que son no correlacionadas y se ordenan de modo tal que unas pocas retengan la mayor cantidad de variación presente en el conjunto original de variables.

En conjuntos de variables con alta dependencia es frecuente que un pequeño número de nuevas variables (menos del 20 % de las originales) expliquen la mayor parte (más del 80 % de la variabilidad original [6]).

Análisis de Componentens Principales

- Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas.
- Perdida de información con el fin de reducir la dimensión original.
- En PCA el concepto de información está asociado a mayor variabilidad, por lo tanto las direcciones seleccionadas para proyectar los datos son las de mayor varianza.
- La proyección corresponde a un plano que minimiza las distancias ortogonales de los puntos a él.

Análisis de Componentens Principales

Supongamos que contamos con un matriz $\mathbf{X}$ que contiene N muestras cada una de d variables.

Para el cálculo de las proyecciones de PCA es necesario que los elementos de la matriz $\mathbf{X}$ tengan media cero, así que suponemos que previamente hemos extraído la media a cada una de las variables.

Por lo tanto la matriz de covarianza de los datos está dada por $\frac{1}{N}\mathbf{X}^T\mathbf{X}$.

Supongamos entonces que inicialmente el objetivo es proyectar todas las muestras observadas sobre un subespacio de dimensión 1 (una recta), de tal forma que todos los puntos mantengan, en la medida de lo posible, sus posiciones relativas.

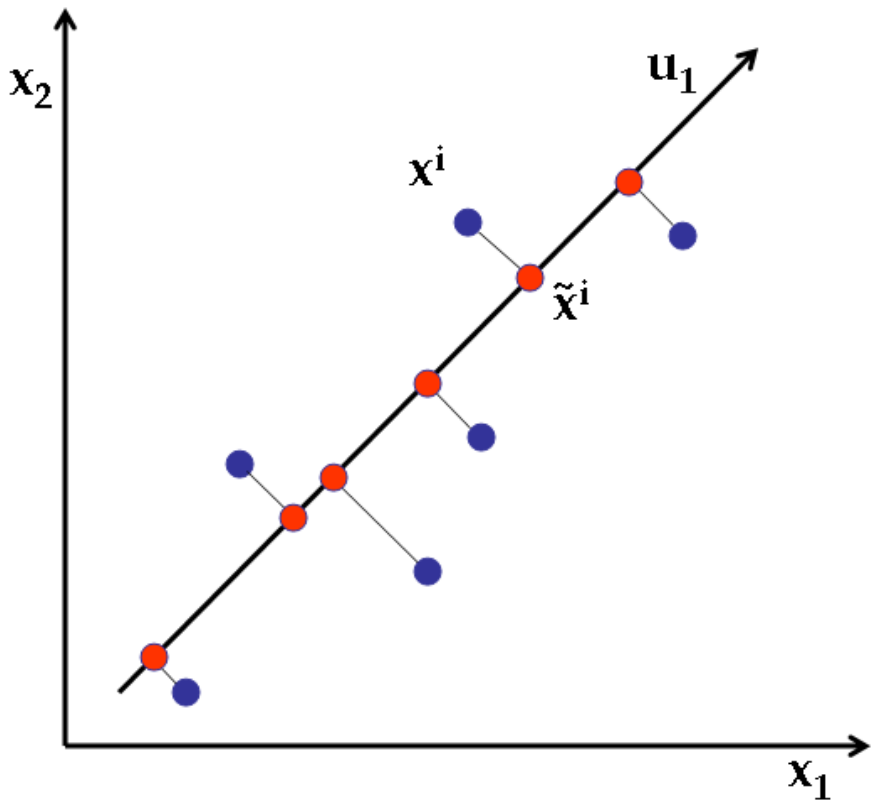
Análisis de Componentens Principales

Una condición a cumplir es que la distancia entre cada punto y su proyección sea mínima. Si representamos la dirección del plano por un vector unitario $\mathbf{u}_1 = (u_{11}, ..., u_{1d})^T$, la proyección de un punto cualquiera $\mathbf{x}_i$ sobre esta dirección es el escalar:

$$z_i = \mathbf{u}_1^T \mathbf{x}_i$$

Sea r_i la distancia entre el punto $\mathbf{x}_i$ y su proyección sobre la dirección $\mathbf{u}_1$, el objetivo es:

$$\min_{\mathbf{u}_1} \sum_{i=1}^{N} r_i^2 = \min_{\mathbf{u}_1} \sum_{i=1}^{N} |\mathbf{x}_i - z_i \mathbf{u}_1|^2$$



Análisis de Componentens Principales

Lo que implica encontrar el vector $\mathbf{u}_1$ que maximice z_i^2, osea:

$$\max_{\mathbf{u}_1} \sum_{i=1}^{N} z_i^2 = \max_{\mathbf{u}_1} \sum_{i=1}^{N} \mathbf{u}_1^T \mathbf{x}_i^T \mathbf{x}_i \mathbf{u}_1$$

Teniendo en cuenta la proyección de toda la matriz $\mathbf{X}$,

$$\mathbf{z}_1 = \mathbf{X}\mathbf{u}_1$$

El problema de maximización anterior se puede reescribir como:

$$\frac{1}{N}\mathbf{z}_1^T \mathbf{z}_1 = \frac{1}{N}\mathbf{u}_1^T \mathbf{X}^T \mathbf{X}\mathbf{u}_1 = \mathbf{u}_1^T \mathbf{S}\mathbf{u}_1$$

donde $\mathbf{S}$ es la matriz de covarianza.

Análisis de Componentens Principales

Para que el valor de $\mathbf{u}_1$ no crezca indiscriminadamente durante la maximización, es necesario limitar su valor (Lo importante del vector $\mathbf{u}_1$ es su dirección). La restricción se introduce a través de multiplicadores de Lagrange:

$$M = \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 - \lambda(\mathbf{u}_1^T \mathbf{u}_1 - 1)$$

Derivando e igualando a cero:

$$\frac{\partial M}{\partial \mathbf{u}_1} = 2\mathbf{S}\mathbf{u}_1 - 2\lambda\mathbf{u}_1 = 0$$

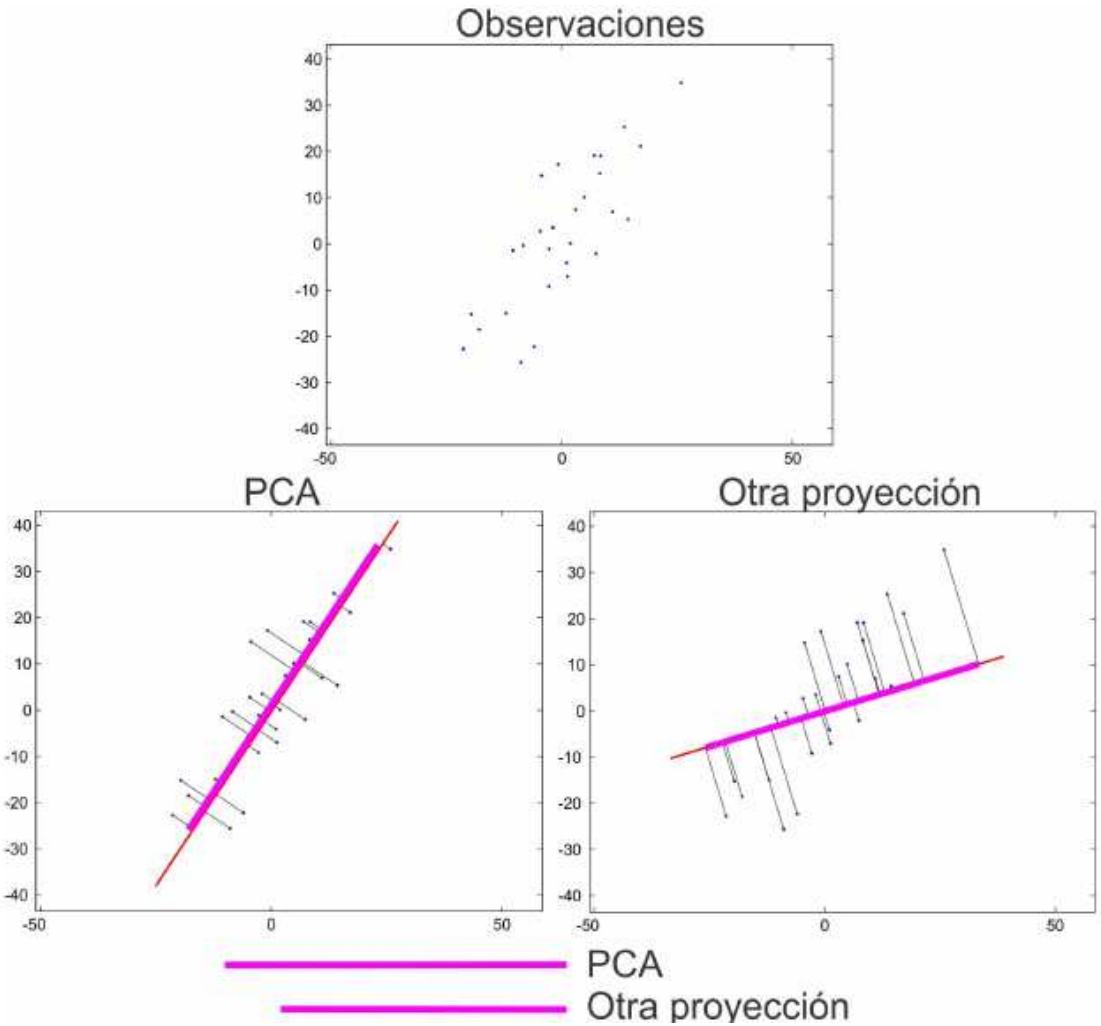
Da como resultado

$$\mathbf{S}\mathbf{u}_1 = \lambda\mathbf{u}_1$$

Esto significa que $\mathbf{u}_1$ es un vector propio de la matriz $\mathbf{S}$ y está asociado al valor propio λ.

Análisis de Componentens Principales

En resumen, el vector propio asociado al mayor valor propio **S** corresponde al primer componente principal.



Análisis de Componentens Principales

En general, es posible hallar el espacio de dimensión $p < d$ que mejor represente los datos, el cual está dado por los vectores propios asociados a los p mayores valores propios de $\mathbf{S}$. Estas nuevas direcciones se denominan direcciones principales de los datos y las proyecciones de los datos originales sobre estas direcciones se conocen como **componentes principales**.

Análisis de Componentens Principales

Algoritmo - Cálculo de los componentes principales

1: Centralizar la matrix de datos $\mathbf{X}$ (Hacer que cada variable tenga media cero)
2: Obtener la matriz de covarianza $\mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$
3: Calcular los valores propios de la matriz $\mathbf{S}$ y sus respectivos vectores propios
4: Ordenar de forma descendente los valores propios
5: Proyectar los datos sobre las direcciones principales luego del ordenamiento de los valores propios.

Selección del número de componentes

- Realizar un gráfico de λ_i contra i.
- Seleccionar los primeros componentes hasta cubrir una proporción determinada de varianza, como por ejemplo el 80 %, 90 % o 95 %.
- Desechar aquellos componentes asociados a valores propios inferiores a una cota, la cual es usualmente la varianza media.
- Tomar como valor óptimo el número de componentes correspondiente al codo la curva λ_i contra i.
- Evaluar a partir de un criterio wrapper el mejor número de componentes (costoso computacionalmente pero de mejore resultados).

Referencias

[1] Velten K., *Mathematical Modeling and Simulation*, WILEY-VCH, 2009.

[2] Murphy K.P., *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[3] Duda R.O., Hart P.E., Stork D.G., *Pattern Classification*. 2ed, WILEY-INTERSCIENCE, 2001.

[4] Bishop, C.M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] Webb, A.R. *Statistical Pattern Recognition*, 2nd Revised edition, John Wiley & Sons Ltd, 2002.

[6]Peña, D. *Análisis de datos multivariantes*. Mc Graw Hill, Madrid, España, 2002.