

Análisis de componentes principales

Haydé Martínez, Marco Mendoza, Ricardo Váldez

AEM 1 Noviembre 2022

Agenda

- Historia
- Definición
- Dimensionalidad
- Interpretación geométrica
- Ejemplos
- Algebra
- Explicación Matemática
- Ejercicios

Historia

Fue inventado en **1901** por **Karl Pearson** como una analogía al teorema de ejes de principales, que en **1930** fue ya desarrollada independientemente por **Harold Hotelling**.



Karl Pearson

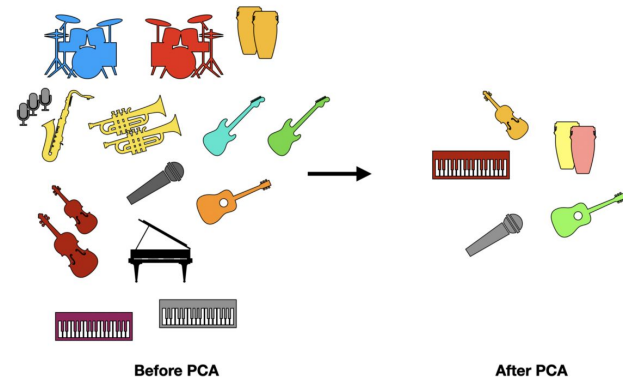


Harold Hotelling

Definición

Es un método estadístico que permite simplificar la complejidad de espacios muestras con muchas **dimensiones** a la vez que **conserva** su **información**.

Supóngase que existe una muestra con **n** individuos cada uno con **p** variables (**X1,X2,X3, ... Xp**), es decir, el espacio muestral tiene **p** dimensiones. Este análisis nos permite encontrar un número de factores subyacentes (**z** < **p**) que explican aproximadamente lo mismo que las **p** variables originales. Donde antes se necesitaban **p** valores para caracterizar a cada individuo, ahora bastan **z** valores. Cada una de estas **z** nuevas variables recibe el nombre de componente principal.



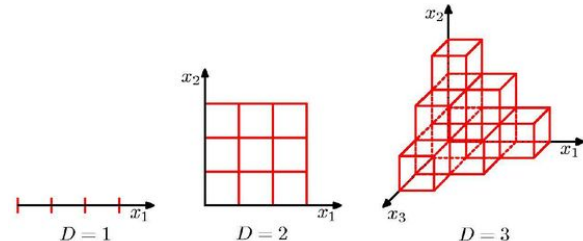
Dimensionalidad

La reducción de dimensionalidad o reducción de la dimensión es el proceso de reducir el número de variables aleatorias que se trate.

Hay distintos enfoques que se pueden usar para la selección de variables que tratan de encontrar un subconjunto de las variables originales.

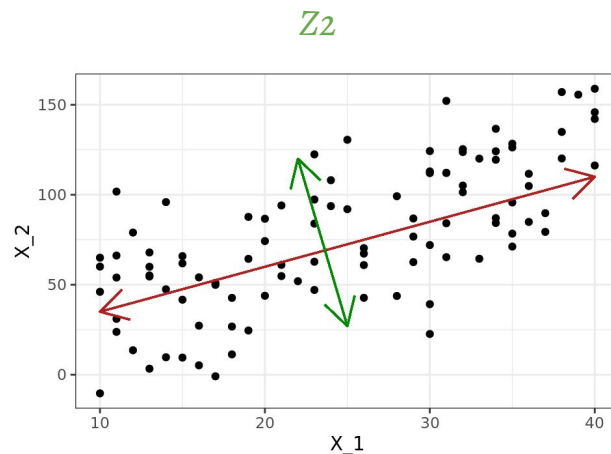
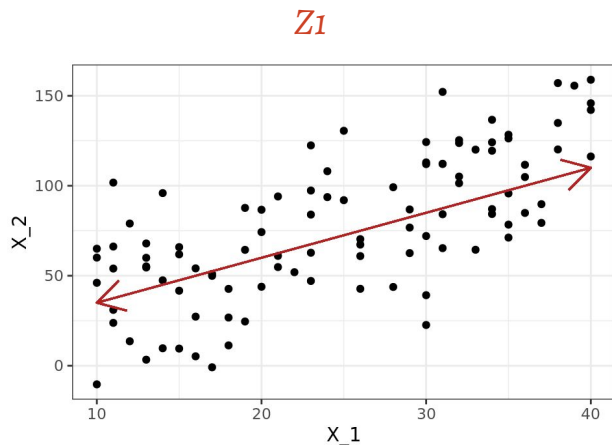
Ventajas:

- Reduce el espacio de tiempo y almacenamiento requerido.
- La eliminación de multicolinealidad mejora el rendimiento del modelo de aprendizaje automático.
- Se hace más fácil de visualizar los datos cuando se reduce a dimensiones muy bajas tales como 2D o 3D.




Interpretación Geométrica

Una forma intuitiva de entender el proceso de PCA consiste en interpretar las componentes principales desde un punto de vista geométrico.

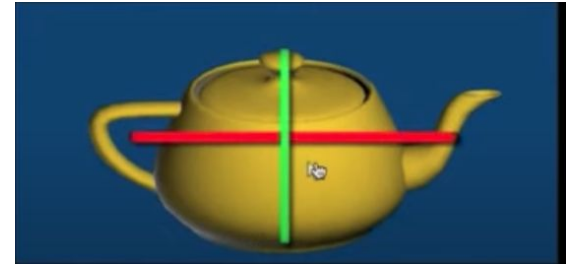
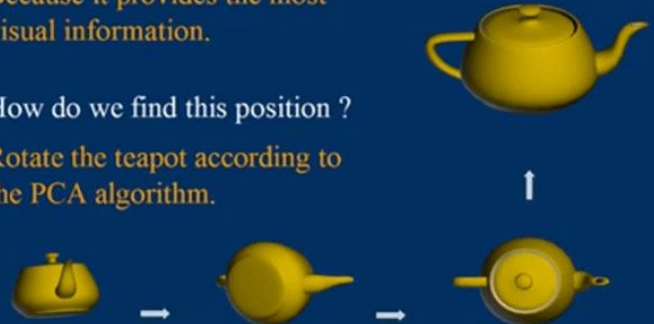


Ejemplos

 Best position for a teapot snapshot

Why this position?
Because it provides the most visual information.

How do we find this position ?
Rotate the teapot according to the PCA algorithm.



Ejemplos

PAISES	AGR	MIN	MAN	ENER	CON	SER	FIN	SSP	TC
BELGICA	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2
DINAMARCA	9.2	0.1	21.8	0.6	8.3	14.6	6.5	32.2	7.1
FRANCIA	10.8	0.8	27.5	0.9	8.9	16.8	6.0	22.6	5.7
RFA	6.7	1.3	35.8	0.9	7.3	14.4	5.0	22.3	6.1
IRLANDA	23.2	1.0	20.7	1.3	7.5	16.8	2.8	20.8	6.1
ITALIA	15.9	0.6	27.6	0.5	10.0	18.1	1.6	20.1	5.7
LUXEMBURGO	7.7	3.1	30.8	0.8	9.2	18.5	4.6	19.2	6.2
HOLANDA	6.3	0.1	22.5	1.0	9.9	18.0	6.8	28.5	6.8
U.K.	2.7	1.4	30.2	1.4	6.9	16.9	5.7	28.3	6.4
AUSTRIA	12.7	1.1	30.2	1.4	9.0	16.8	4.9	16.8	7.0
FINLANDIA	13.0	0.4	25.9	1.3	7.4	14.7	5.5	24.3	7.6
GRECIA	41.4	0.6	17.6	0.6	8.1	11.5	2.4	11.0	6.7
NORUEGA	9.0	0.5	22.4	0.8	8.6	16.9	4.7	27.6	9.4
PORTUGAL	27.8	0.3	24.5	0.6	8.4	13.3	2.7	16.7	5.7
ESPAÑA	22.9	0.8	28.5	0.7	11.5	9.7	8.5	11.8	5.5
SUECIA	6.1	0.4	25.9	0.8	7.2	14.4	6.0	32.4	6.8
SUIZA	7.7	0.2	37.8	0.8	9.5	17.5	5.3	15.4	5.7
TURQUÍA	66.8	0.7	7.9	0.1	2.8	5.2	1.1	11.9	3.2
BULGARIA	23.6	1.9	32.3	0.6	7.9	8.0	0.7	18.2	6.7
CHECOSLOVAQUIA	16.5	2.9	35.5	1.2	8.7	9.2	0.9	17.9	7.0
RDA	4.2	2.9	41.2	1.3	7.6	11.2	1.2	22.1	8.4
HUNGRÍA	21.7	3.1	29.6	1.9	8.2	9.4	0.9	17.2	8.0
POLONIA	31.1	2.5	25.7	0.9	8.4	7.5	0.9	16.1	6.9
RUMANIA	34.7	2.1	30.1	0.6	8.7	5.9	1.3	11.7	5.0
URSS	23.7	1.4	25.8	0.6	9.2	6.1	0.5	23.6	9.3
YUGOSLAVIA	48.7	1.5	16.8	1.1	4.9	6.4	11.3	5.3	4.0



	Comp.1	Comp.2
AGR	-0.978122878	-0.07822089
MIN	-0.002471414	-0.90169648
MAN	0.648909478	-0.51820473
ENER	0.477522243	-0.38107264
CON	0.607237119	-0.07485607
SER	0.707591383	0.51108007
FIN	0.138884602	0.66217713
SSP	0.723443938	0.32331269
TC	0.685001641	-0.29568509

Math Background

PCA es un proceso que puede ser catalogado como aprendizaje no-supervisado que consta de seis pasos/partes:

- Obtención de un dataset n dimensional sin etiquetas.
- Calcular el promedio de todas las componentes.
- Calcular la matriz de covarianza del dataset completo.
- Calcular los eigenvectores con base en los eigenvalores del sistema.
- Crear una matriz de transformación con base en los eigenvalores y los eigenvectores encontrados.
- Siendo K eigenvalores, se procede a transformar los datos iniciales por medio de la matriz de transformación $(n \times k)$ obtenida.

Valores y vectores propios

Los eigenvalores o valores propios son definidos como las características principales de una transformación lineal (una matriz), mientras que los vectores propios conforman directamente el subespacio de las transformaciones que se obtendrán a través de dicha transformación lineal.

La forma matemática para obtener los eigenvalores parte de la suposición siguiente:

$$T : V \rightarrow V$$

$$\vec{x} \in V$$

$$T(\vec{x}) = \lambda \vec{x}$$

Valores y vectores propios

Si A es la matriz asociada a la transformación, podemos escribir:

$$A\vec{x} = \lambda\vec{x}$$

Con un poco de álgebra, podemos obtener el polinomio característico, del cual podemos obtener los valores de lambda, así, obtenemos:

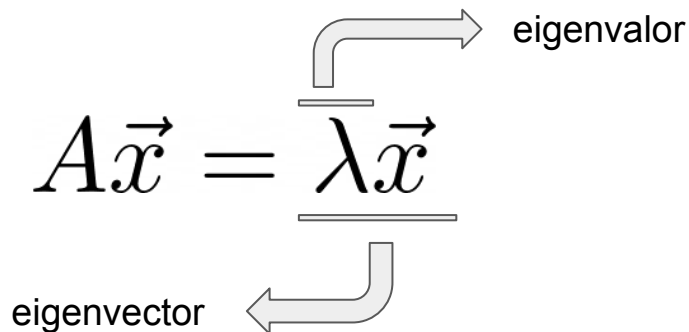
$$P(\lambda) = \det(I\lambda - A) = 0$$

Valores y vectores propios

Por su parte, los vectores propios se encuentran resolviendo la ecuación:

$$(I\lambda_i - A)\vec{x} = 0$$

donde λ_i es el i -ésimo valor de λ y \vec{x} es un vector columna de coordenadas cualesquiera que satisfacen la ecuación anterior.

$$A\vec{x} = \lambda\vec{x}$$


eigenvalor

eigenvector

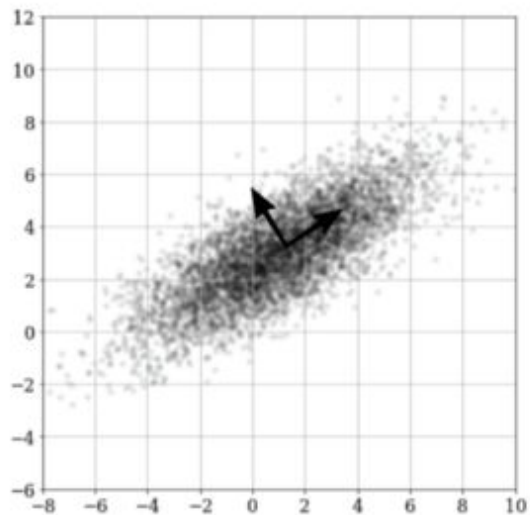
Ejemplo

$$R(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}, \quad \text{where } 0 \leq \theta < 2\pi,$$

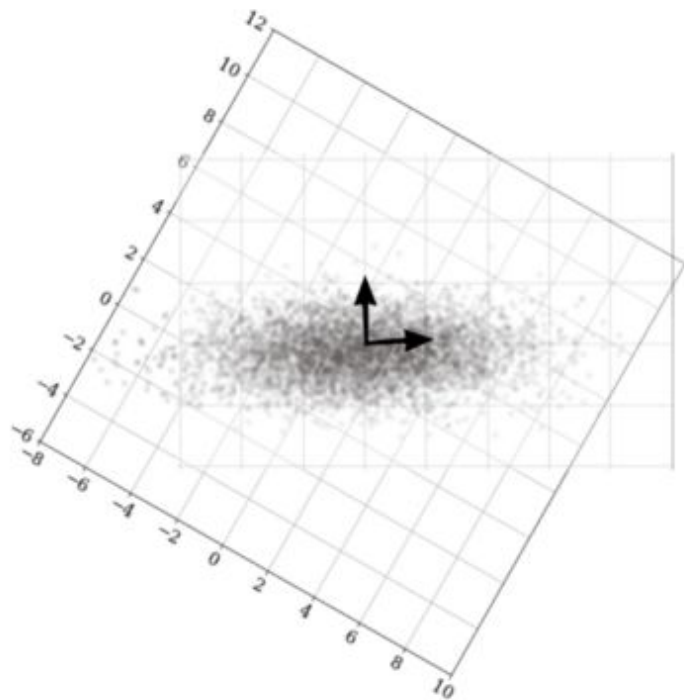
$$\det(R(\theta) - \lambda \mathbf{I}) = 0 \quad \implies \quad \det \begin{pmatrix} \cos \theta - \lambda & -\sin \theta \\ \sin \theta & \cos \theta - \lambda \end{pmatrix} = 0,$$

$$(\cos \theta - \lambda)^2 + \sin^2 \theta = 0.$$

$$\lambda = \cos \theta \pm \sqrt{\cos^2 \theta - 1} = \cos \theta \pm i \sin \theta = e^{\pm i\theta}.$$



what our eigenvectors look like



roughly “what our eigenbases
see” - superimposed on cartesian
plane

Hands-on

Ámonos pal R...