

Estadística Bayesiana – R

Clase 01 de Abril

Componentes Principales

Introducción

El objetivo es pasar de n a p variables, con $n > p$. Al comprimir variables podemos hacer representaciones graficas en 2 o 3 dimensiones

Para calcular las componentes principales se descomponen nuestras variables originales en combinaciones lineales:

- Estandarizamos nuestras variables a media 0
- Bajo un proceso de optimización buscamos los valores de los factores que maximicen la varianza
- El cálculo de las demás componentes se hace de igual forma, pero teniendo presente que esta nueva dirección sea ortogonal al resto de componentes principales

Introducción

Componentes máximas = $\min(\text{Número de muestras}, \text{Numero de variables}) - 1$
Si calculamos el total de componentes su varianza sería igual a la varianza total del conjunto de datos original

Introducción

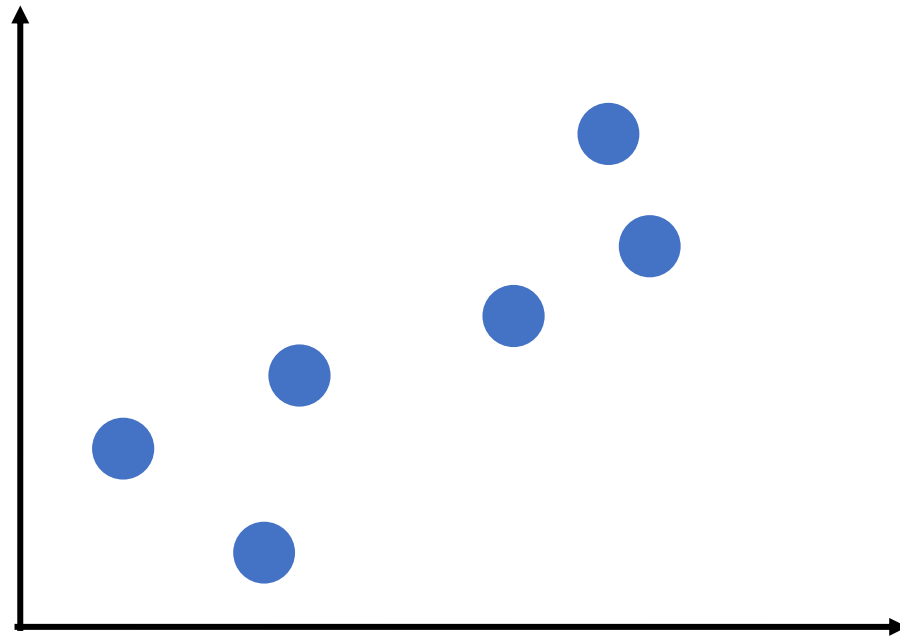
Si Tenemos k componentes, la varianza explicada será

$$Var_K = \frac{\sum_{p=1}^K (Q_p)}{\sum_{n=1}^N Var(X_n)}$$

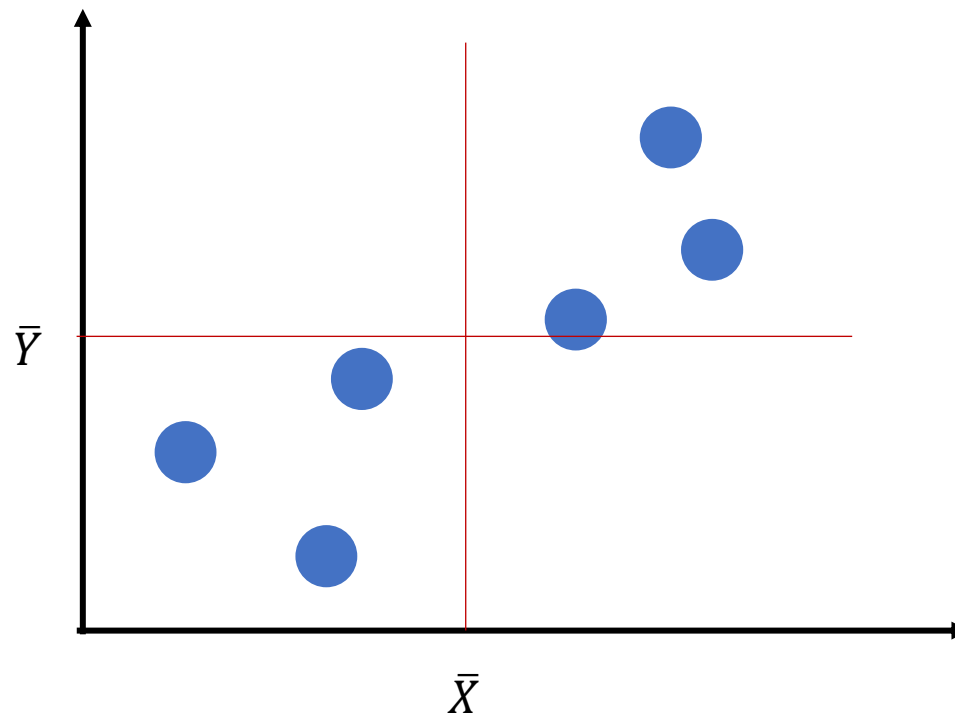
Intuición grafica del análisis de componentes principales

Imaginemos un dataset con 2 variables

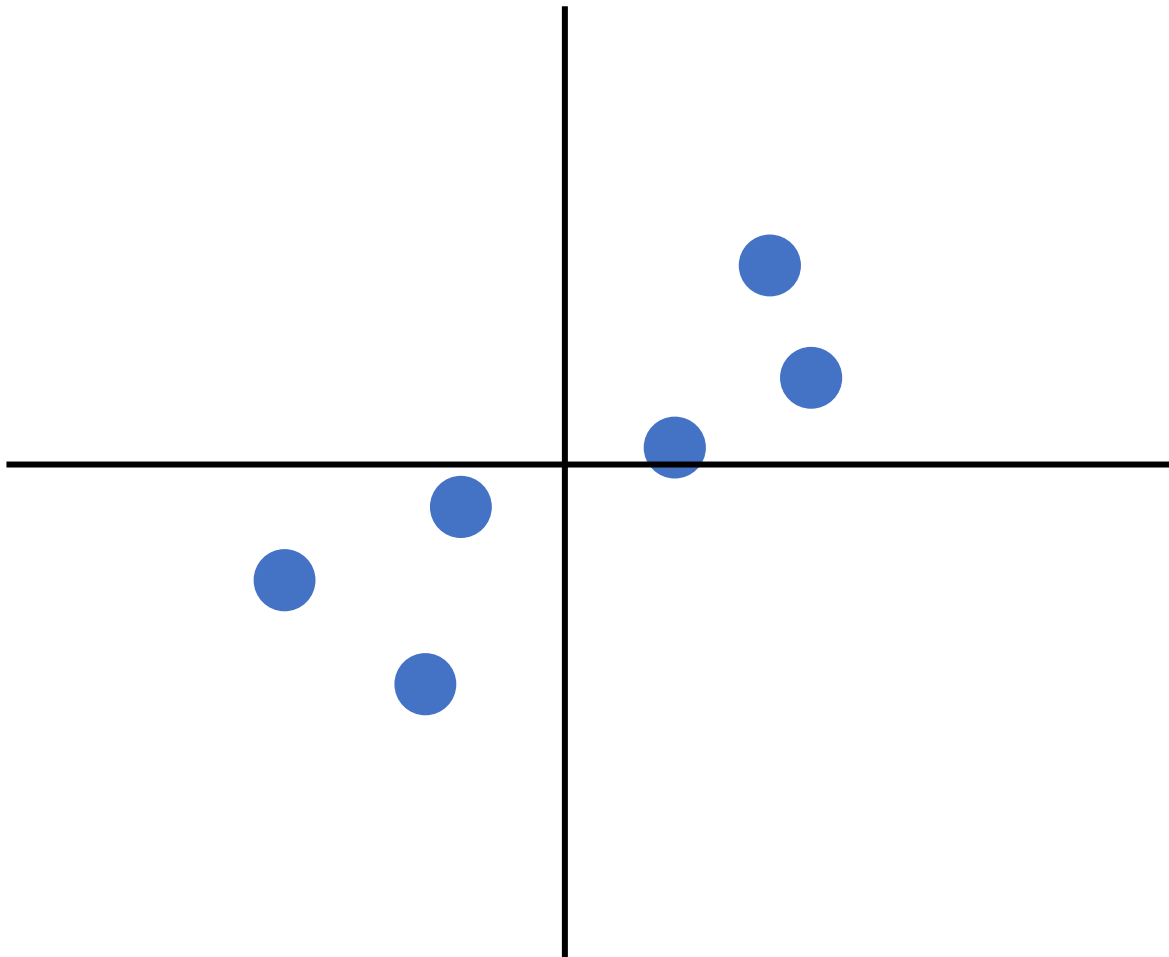
X	Y
X_1	Y_1
X_2	Y_2
X_3	Y_3
X_4	Y_4
X_5	Y_5
X_6	Y_6



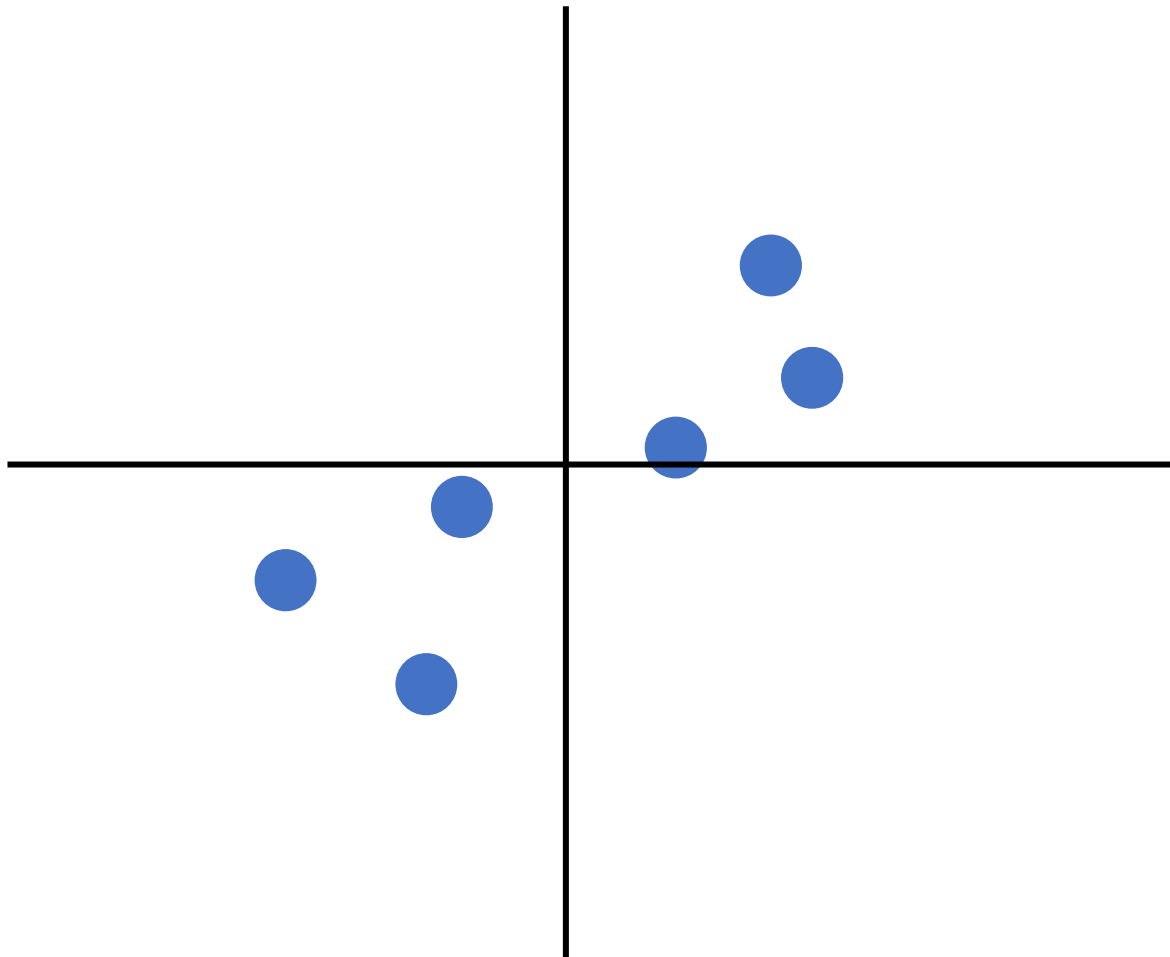
Primero se calcula la media promedio para cada variable, luego con el promedio de ambas variables se puede calcular el centro de los datos



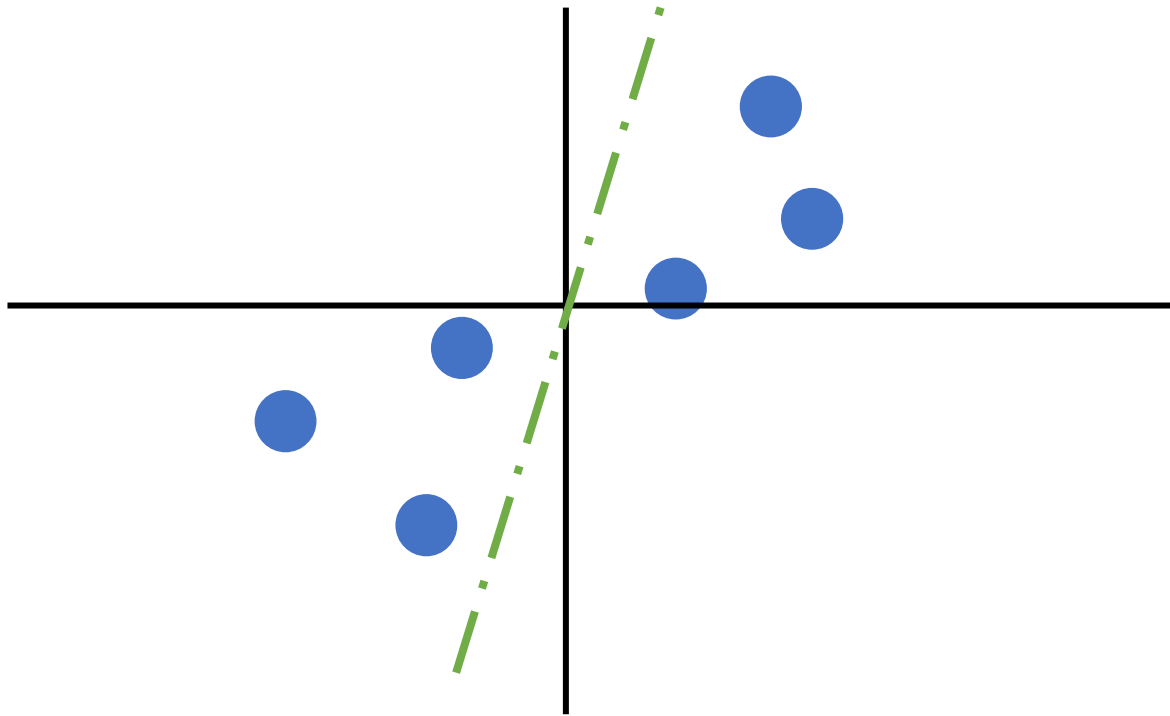
Ahora desplazamos los datos para que el centro quede en el origen del grafico



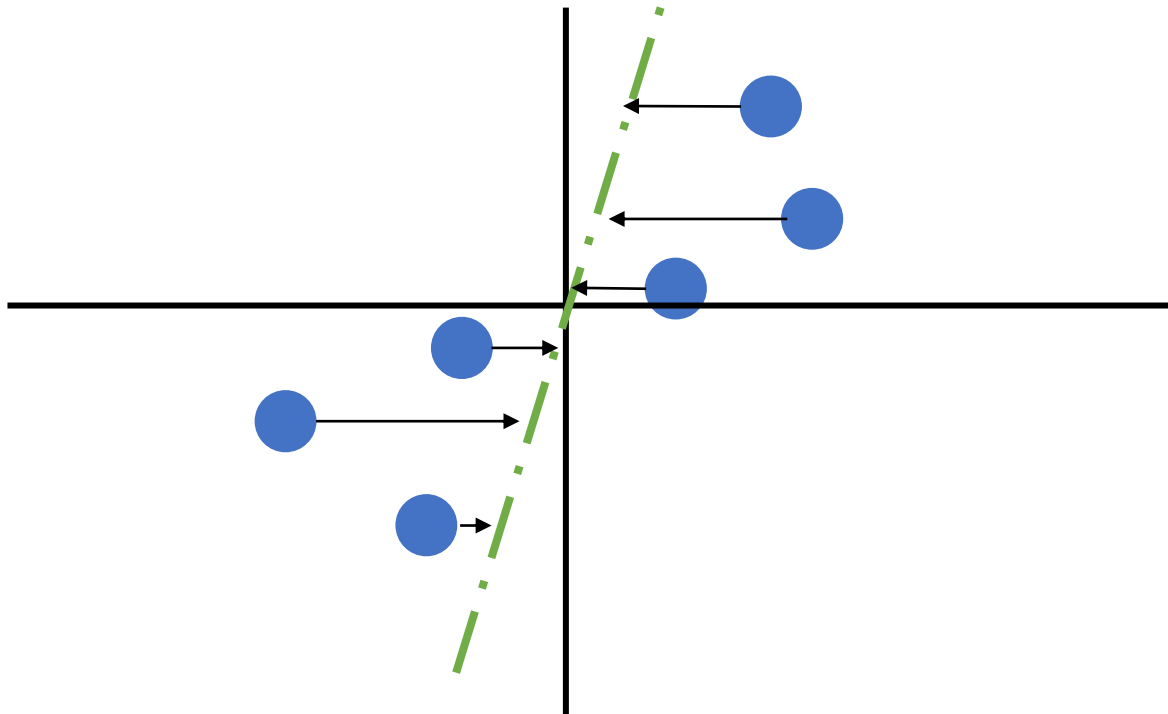
Es importante notar que, si bien hemos cambiado el centro de los datos, no hemos cambiado la forma en que los datos se posicionan relativos unos de otros



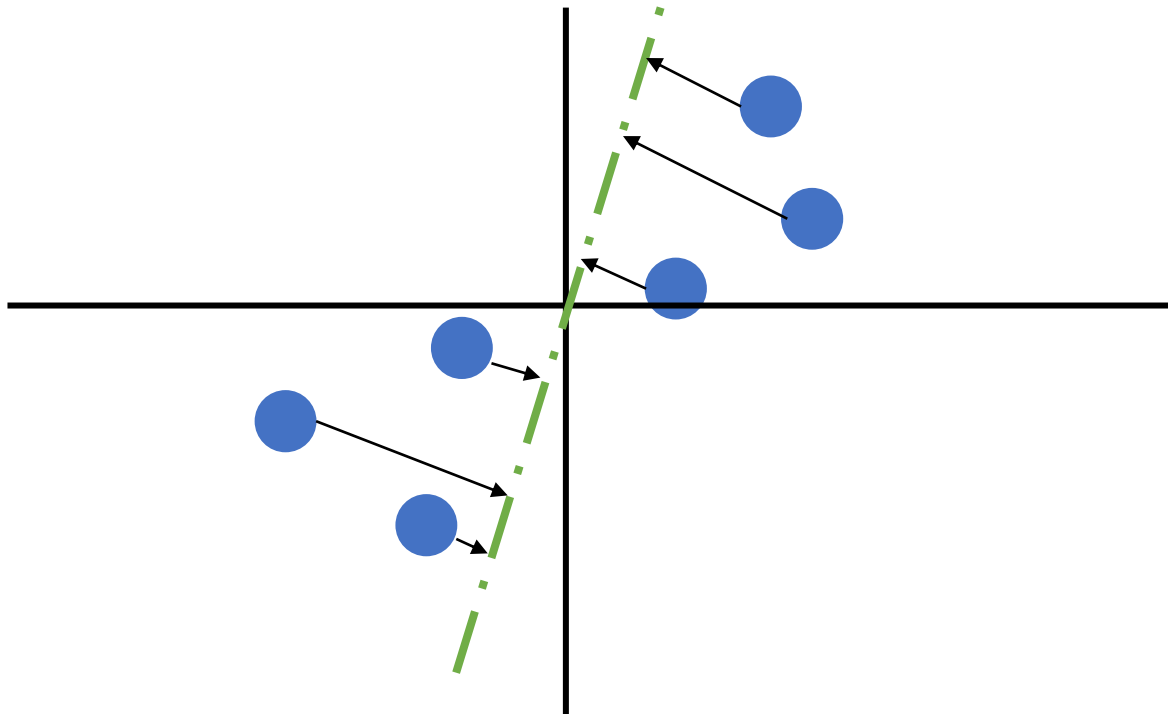
Ahora que los datos están centrados buscamos una línea que se ajuste a los datos



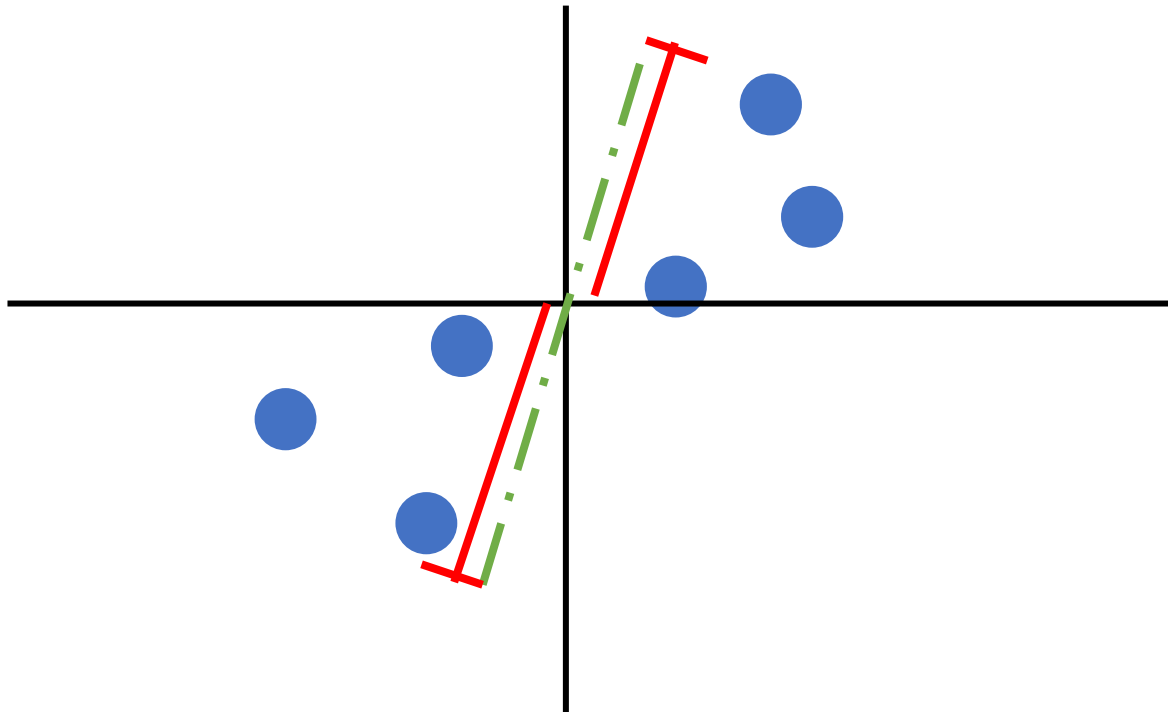
Para cuantificar que tan bueno es el ajuste de esta línea, proyectamos los datos sobre la misma, para encontrar la distancia entre los datos y la recta y así buscar la recta con menor distancia .



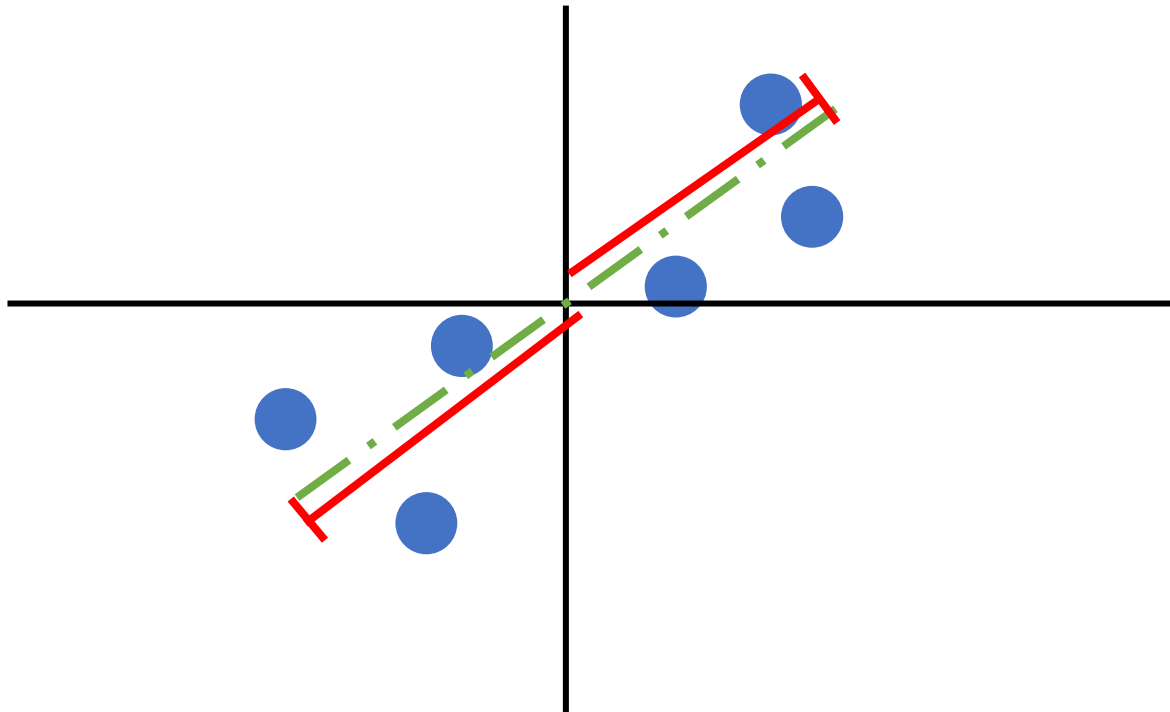
La mejor recta será aquella que minimice las distancias entre los puntos y sus proyecciones en la recta:



O buscar una recta que maximice la distancia desde los puntos proyectados hasta el origen



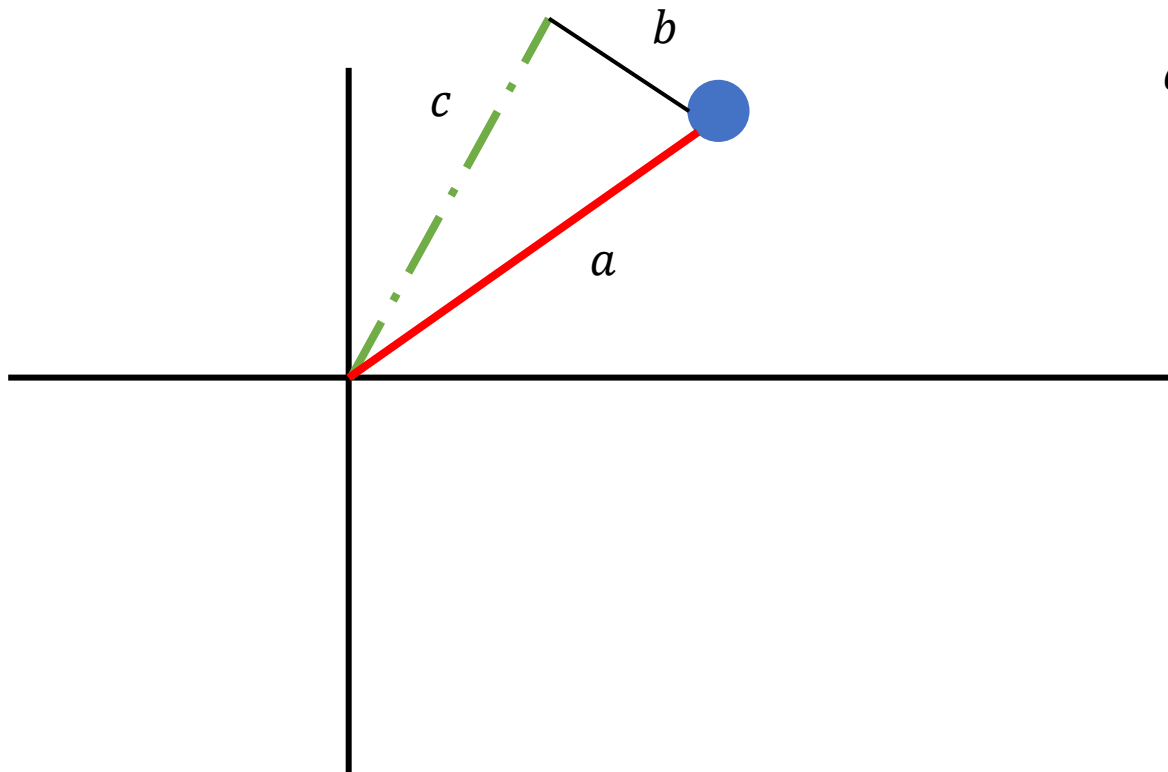
La varianza se maximiza cuando las distancias se minimizan



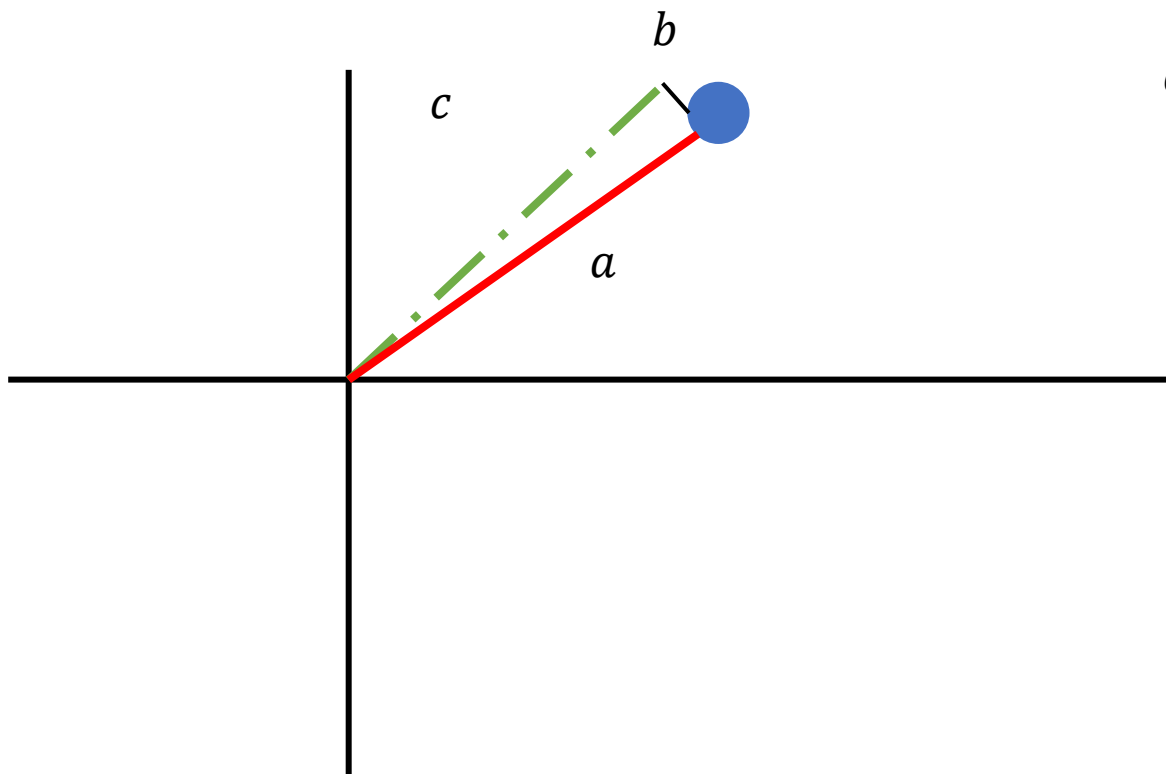
Intuición

Para entender que pasa de forma matemática tomemos solo un punto. Dicho punto es fijo, y conocemos su distancia al origen. Esto implica que esa distancia al origen es independiente de la recta de ajuste que tracemos.

Si proyectamos el punto sobre la recta de ajuste se forma un triángulo rectángulo. Con lo cual podemos ver como la varianza de la recta está inversamente relacionada con la distancia de la recta al punto



$$a^2 = b^2 + c^2$$



$$a^2 = b^2 + c^2$$

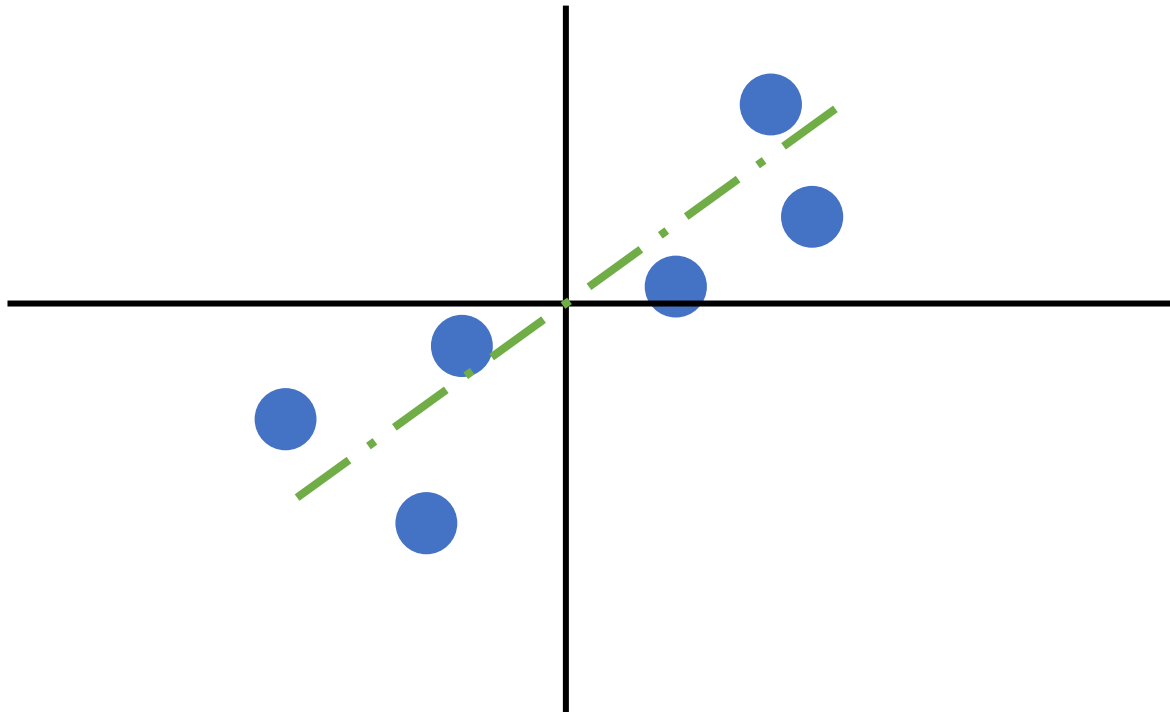
Intuición

“a” cuadrado está fija, por lo tanto, para que “b” decrezca, “c” debe crecer. PCA puede minimizar la distancia de proyección del punto o maximizar la distancia desde el punto de proyección al origen.

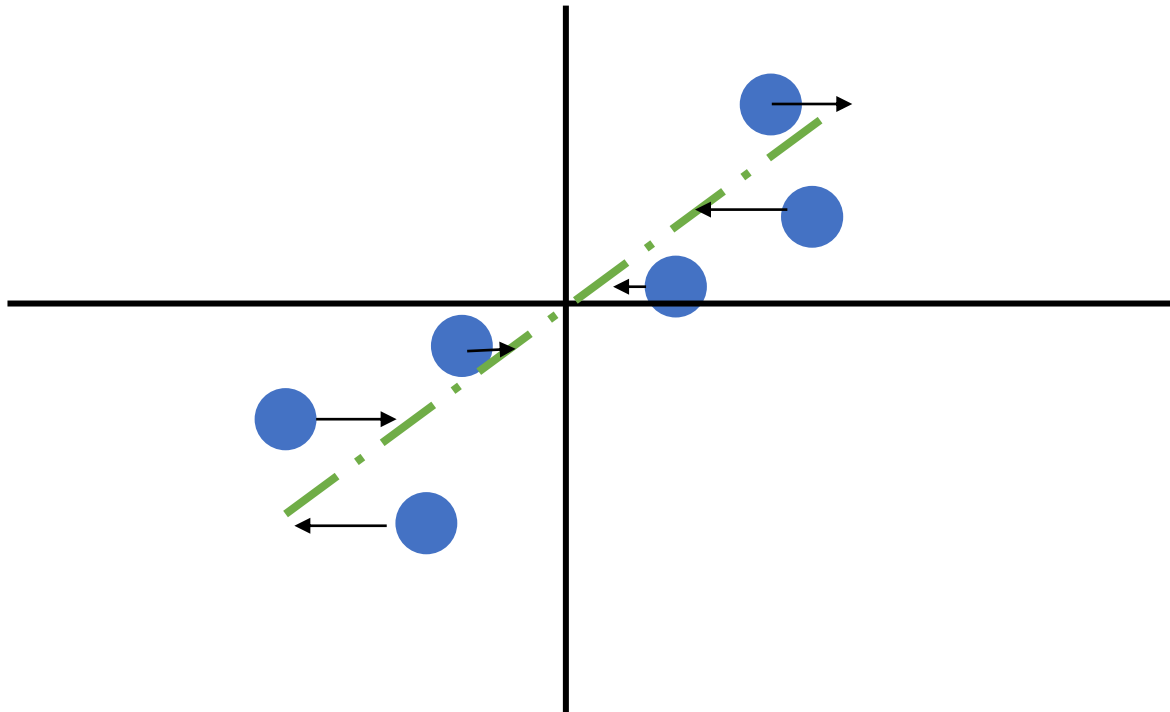
Si bien es más intuitivo seguir el camino de la minimización es más fácil metodológicamente maximizar.

PCA encuentra la mejor recta ajuste que maximice el cuadrado de las distancias del punto proyectado hacia el origen

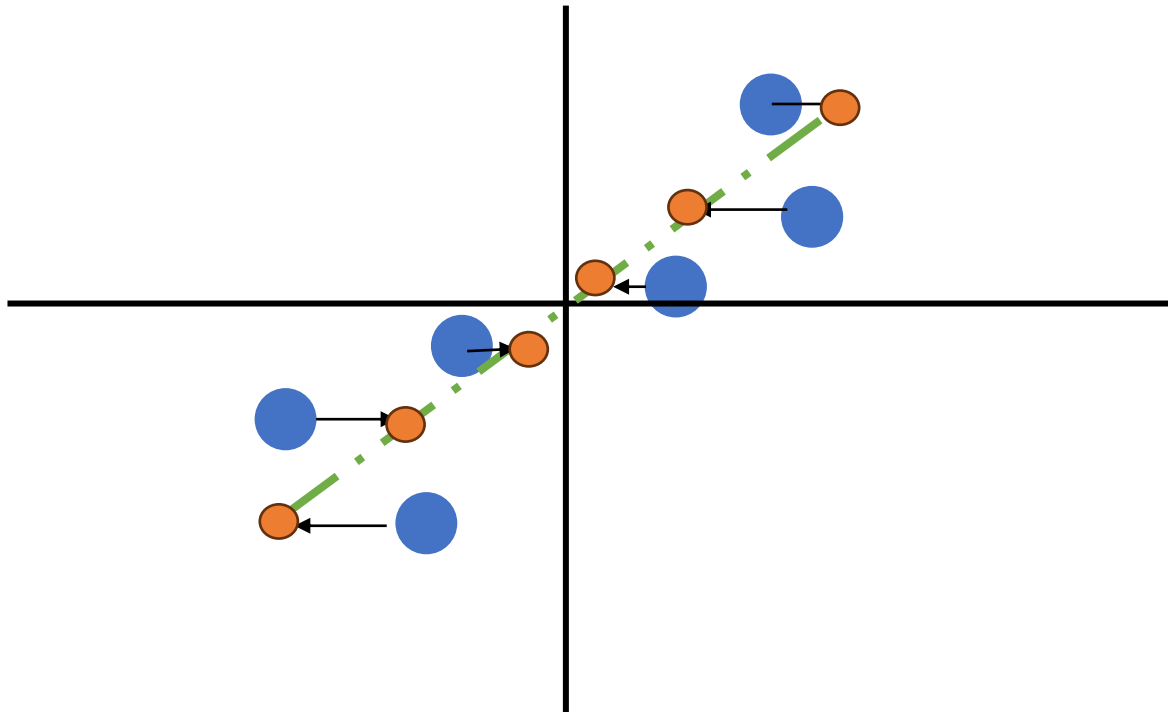
Esta recta es llamada primer componente principal.



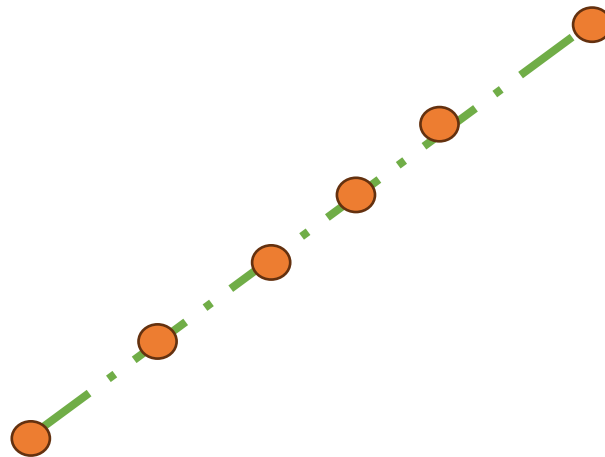
Esta recta es llamada primer componente principal.



Esta recta es llamada primer componente principal.



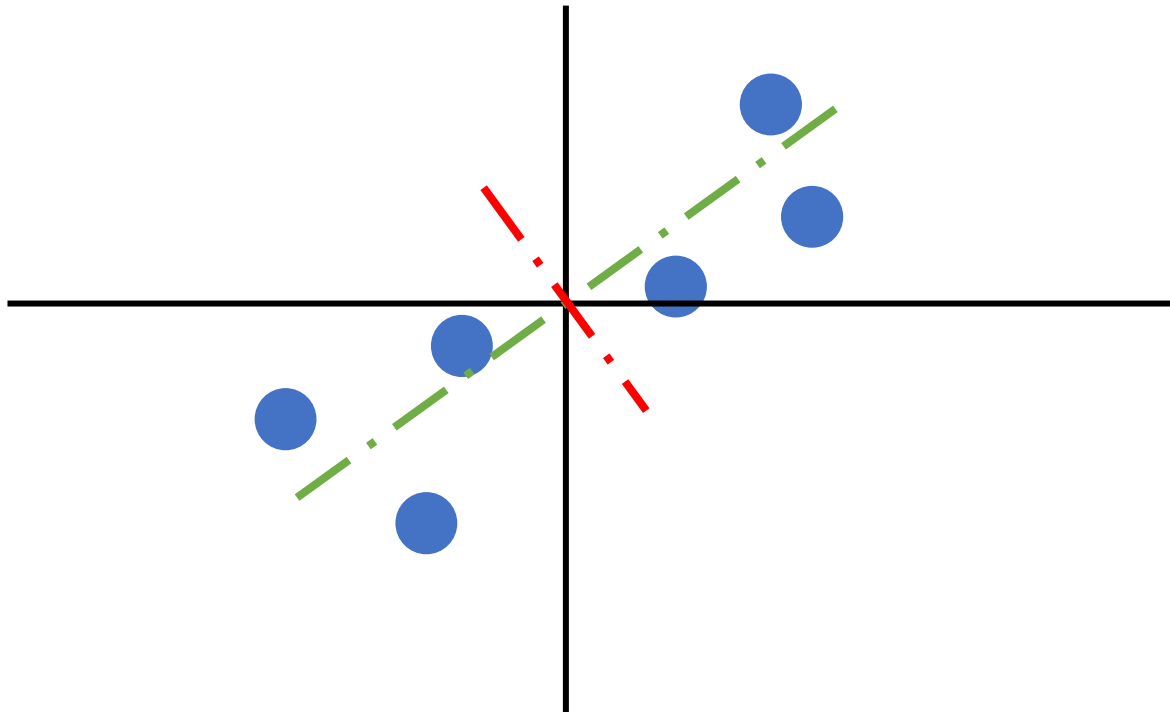
Esta recta es llamada primer componente principal.

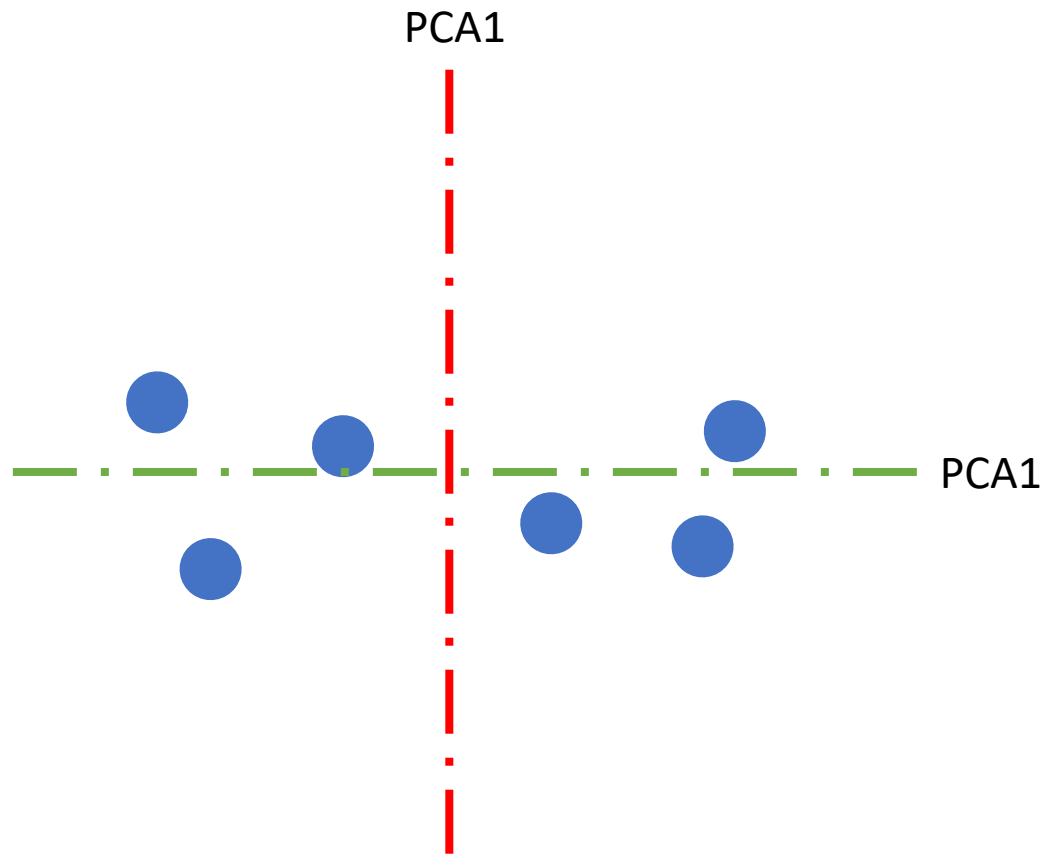


Esta recta es llamada primer componente principal.



La segunda componente será ortogonal a la primera





Cálculo de Componentes Principales

Análisis de componentes principales tiene como objetivo explicar la estructura de varianzas y covarianzas de un conjunto de variables a través de unas pocas combinaciones lineales de estas variables.

- Reducción de la dimensionalidad de los datos.
- Permite transformar variables correlacionadas, en un conjunto de variables
- No correlacionadas.
- Ayuda a la interpretación de los valores que toman las variables.
- No requiere supuesto de normalidad multivariante.
- Trabaja con variables continuas.

Cálculo de Componentes Principales

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{np} \end{bmatrix}$$

- n individuos
- p variables

Cálculo de Componentes Principales

El primer componente principal será la combinación lineal de las variables originales que tenga varianza máxima. Los valores de este primer componente en los n individuos se representarán por un vector z_1 , dado por

$$z_1 = Xa$$

Cálculo de Componentes Principales

Para cada variable calculamos su media y configuramos la matriz X como la matriz de las desviaciones de las observaciones respecto al promedio. Como las variables tienen media cero también z_1 tendrá media nula. Su varianza será:

$$Var(z_1) = \frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X' X a_1 = a_1' S a_1$$

donde S es la matriz de varianzas y covarianzas de las observaciones. Es obvio que podemos maximizar la varianza sin límite aumentando el módulo del vector a_1 .

Para que la maximización de la varianza tenga solución debemos imponer una restricción al módulo del vector a_1 , y, sin pérdida de generalidad, impondremos que $a_1' a_1 = 1$.

Cálculo de Componentes Principales

Introduciremos esta restricción mediante el multiplicador de Lagrange:

$$M = a'_1 S a_1 - \lambda(a'_1 a_1 - 1)$$

y maximizaremos esta expresión de la forma habitual derivando respecto a los componentes de a_1 e igualando a cero. Entonces

$$\frac{\partial M}{\partial a_1} = 2S a_1 - 2\lambda a_1 = 0$$

Cálculo de Componentes Principales

Cuya solución es:

$$Sa_1 = \lambda a_1$$

que implica que a_1 es un vector propio de la matriz S , y λ su correspondiente valor propio.

Cálculo de Componentes Principales

cuya solución es:

$$Sa_1 = \lambda a_1$$

que implica que a_1 es un vector propio de la matriz S , y λ su correspondiente valor propio.

Para determinar qué valor propio de S es la solución de la ecuación tendremos en cuenta que:

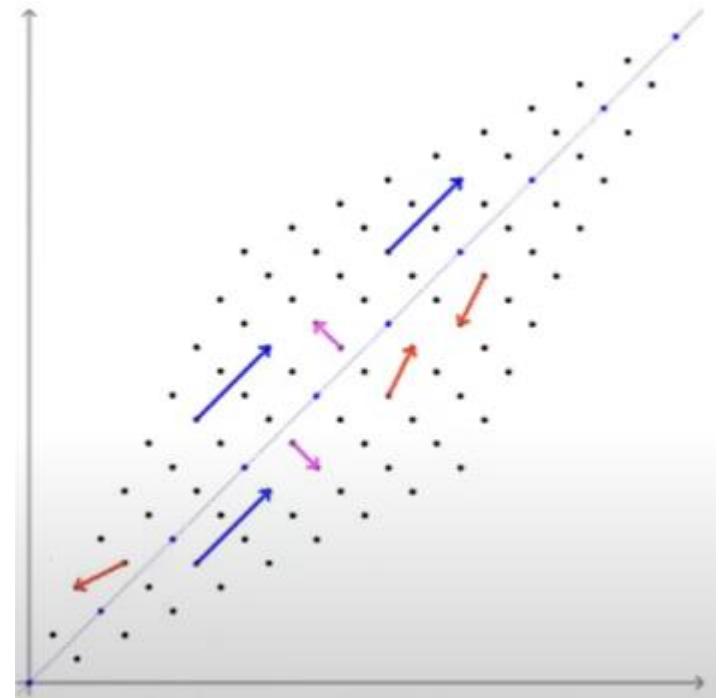
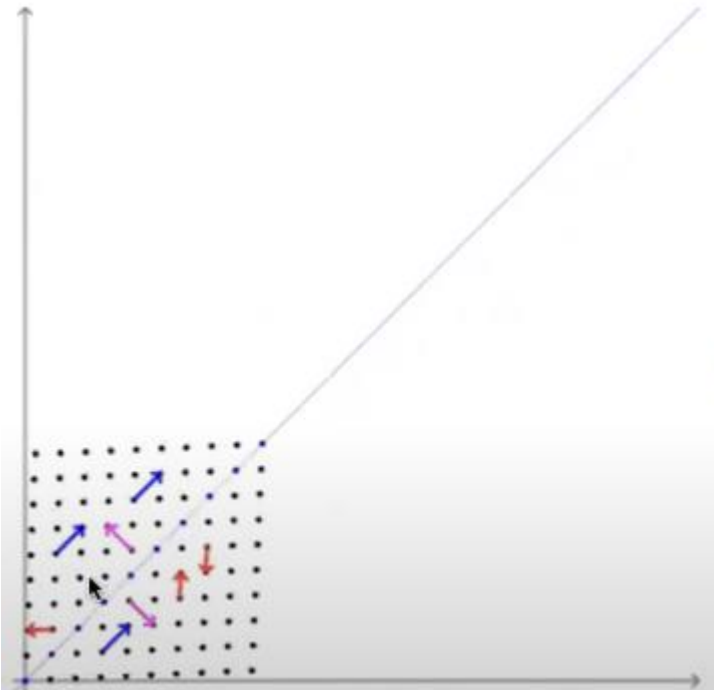
$$Sa_1 = \lambda a_1$$

$$a_1' a_1 = 1$$

$$\begin{aligned} a_1' Sa_1 &= \lambda a_1' a_1 \\ &= \lambda \end{aligned}$$

Concluimos que λ es la varianza de $z_1 \rightarrow a_1' Sa_1 = Var(z_1) = \lambda$

¿Qué son los valores y vectores propios?



Ejes de varianza. Nos importa cuantas transformaciones les aplique a los datos, estos vectores (ejes) no van a cambiar

¿Qué son los valores y vectores propios?

Si quisiéramos hacer una transformación a los datos, nos interesa conocer estos vectores inmutables, ya que van a preservar la estructura de los datos

¿Qué son los valores y vectores propios?

En una matriz A , un valor propio λ su correspondiente vector propio \vec{v} , deben satisfacer la condición:

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

Cada vector, no nulo, que cumple la condición descrita arriba es un vector propio, asociado a un valor propio.

$$Sa_1 = \lambda a_1$$

$$\mathbf{a}'_1 S \mathbf{a}_1 = \lambda \mathbf{a}'_1 \mathbf{a}_1 = \lambda$$

¿Qué son los valores y vectores propios?

$$\begin{aligned} A\vec{v} &= \lambda\vec{v} \\ A\vec{v} - \lambda\vec{v} &= 0 \\ A\vec{v} - \lambda I_n \vec{v} &= 0 \\ (A - \lambda I_n)\vec{v} &= 0 \end{aligned}$$

Por propiedades, $A - \lambda I_n$ no es invertible, por lo tanto, su determinante es 0
De ese modo, para calcular los valores propios de A basta solucionar la ecuación:

$$\det(A - \lambda I_n) = 0$$

Cálculo de Componentes Principales

Solución es:

$$Sa_1 = \lambda a_1$$

que implica que a_1 es un vector propio de la matriz S , y λ su correspondiente valor propio.

Para determinar qué valor propio de S es la solución de la ecuación tendremos en cuenta que:

$$Sa_1 = \lambda a_1$$

$$a_1' a_1 = 1$$

$$\begin{aligned} a_1' Sa_1 &= \lambda a_1' a_1 \\ &= \lambda \end{aligned}$$

Concluimos que λ es la varianza de $z_1 \rightarrow a_1' Sa_1 = Var(z_1) = \lambda$

Segunda componente principal

En este caso la suma de las varianzas de $z_1 = Xa_1$ y $z_2 = Xa_2$ debe ser máxima, y a_1 y a_2 definen el plano de proyección de las variables \mathbf{X} . La función objetivo será:

$$\phi = a_1' S a_1 + a_2' S a_2 - \lambda_1 (a_1' a_1 - 1) - \lambda_2 (a_2' a_2 - 1)$$

$$\frac{\partial \phi}{\partial a_1} = 2S a_1 - 2\lambda_1 a_1 = 0$$

$$\frac{\partial \phi}{\partial a_2} = 2S a_2 - 2\lambda_2 a_2 = 0$$

Segunda componente principal

La solución de este sistema es:

$$Sa_1 = \lambda_1 a_1$$

$$Sa_2 = \lambda_2 a_2$$

que indica que a_1 y a_2 deben ser vectores propios de \mathbf{S} . En el máximo la función objetivo es:

$$\emptyset = \lambda_1 + \lambda_2$$

λ_1 y λ_2 deben ser los dos valores propios mayores de la matriz \mathbf{S} y u_1 y u_2 sus correspondientes vectores propios.

IMPORTANTE $a_1' a_2 = 0$, luego las variables y_1 y y_2 están incorreladas