Reducción de la dimensionalidad

Organización de Datos
Ing. Juan M. Rodríguez

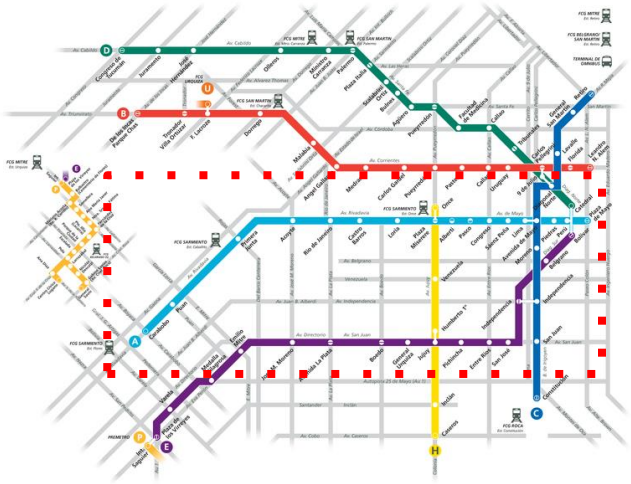
Reducción de la dimensionalidad

Visualización de datos en una dimensión menor (2D o 3D), preservando características importantes de estos como distancias, correlaciones, etc.

- Visualización de datos para entender su distribución
  - detección de patrones inherentes a simple vista
- Reducción del ruido
- Aceleración de los tiempos de entrenamiento de un modelo
- Compresión de la información
- Presentación de resultados a interesados (quienes no siempre conocen de ciencia de datos)

Reducción de la dimensionalidad

Reducción de la dimensionalidad



Reducción de la dimensionalidad

Hay varios algoritmos que resuelven este problema, con diversos usos, ventajas y debilidades

- PCA
- ISOMAP
- LLE
- Proyecciones aleatorias
- MDS
- t-SNE
- LDA
- UMAP

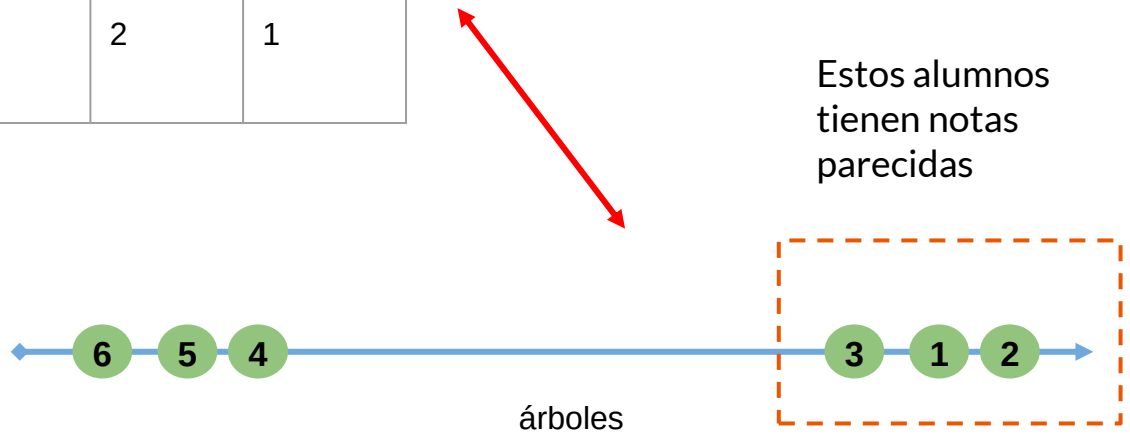
Estos son los algoritmos que veremos en clase

PCA: Análisis de componentes principales

PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

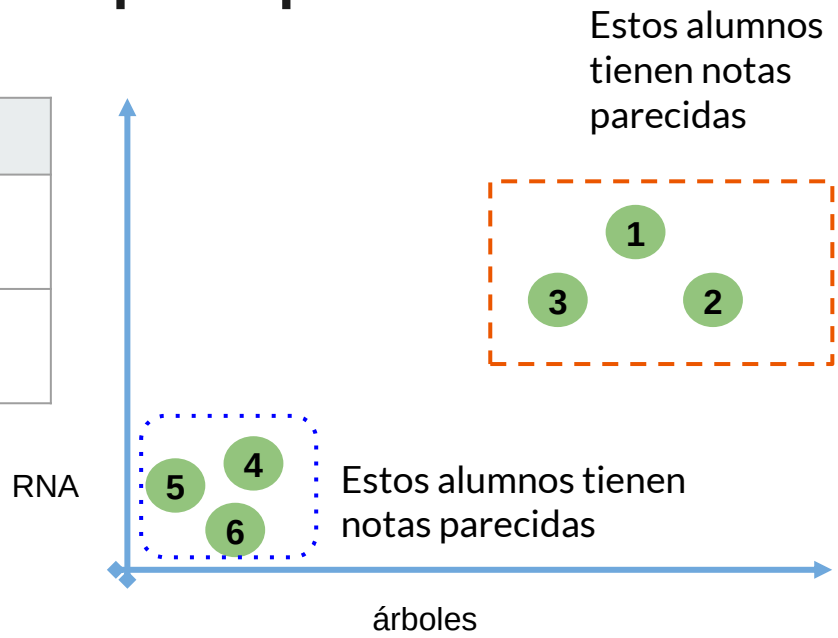
Puntos	Alumno 1	Alumno 2	Alumno 3	Alumno 4	Alumno 5	Alumno 6
árboles	9	10	8	3	2	1



PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

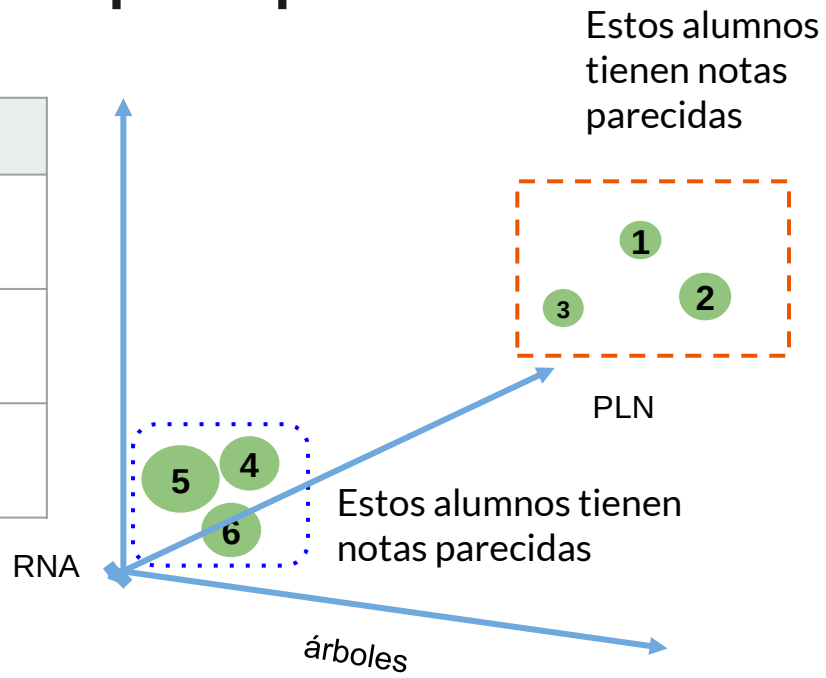
Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	8	3	2	1
RNA	6	4	5	3	2.8	1



PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	8	3	2	1
RNA	6	4	5	3	2.8	1
PLN	12	9	10	2.5	1.3	2

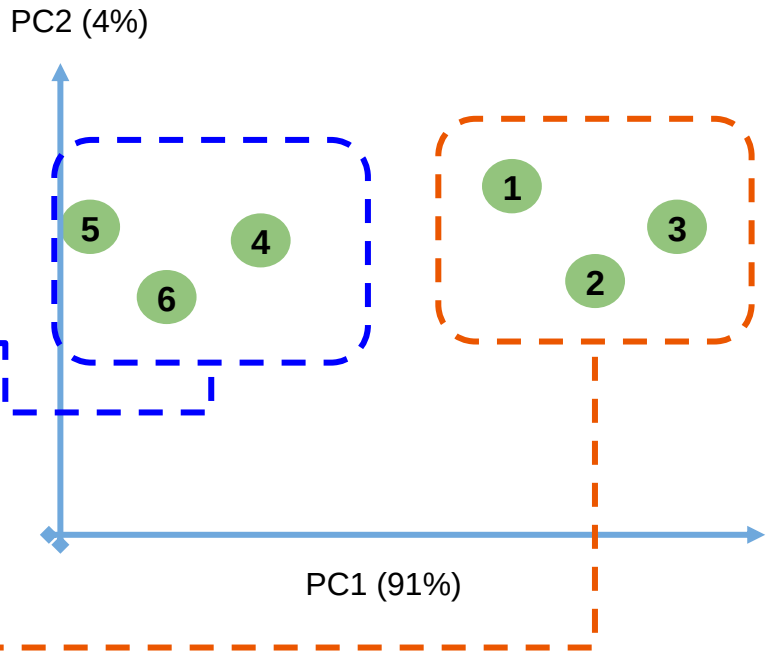


¿Qué pasa si tengo 4 o más variables?

PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

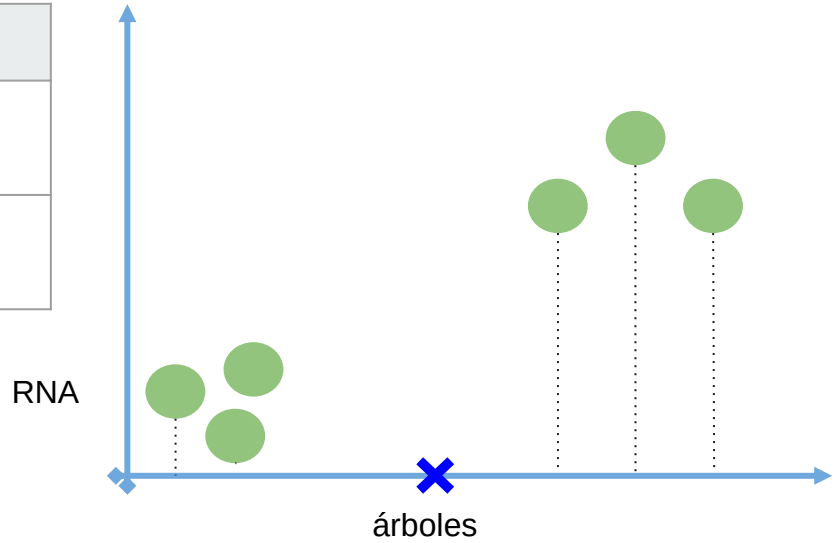
Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	8	3	2	1
RNA	6	4	5	3	2.8	1
PLN	12	9	10	2.5	1.3	2
SVM	5	7	6	2	4	7



PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	8	3	2	1
RNA	6	4	5	3	2.8	1

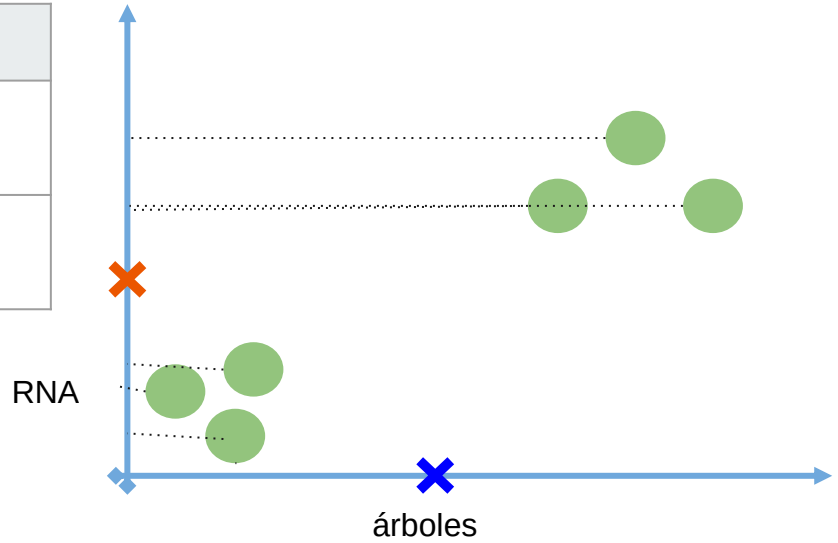


Calculamos el valor promedio para árboles (variable 1)

PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	8	3	2	1
RNA	6	4	5	3	2.8	1



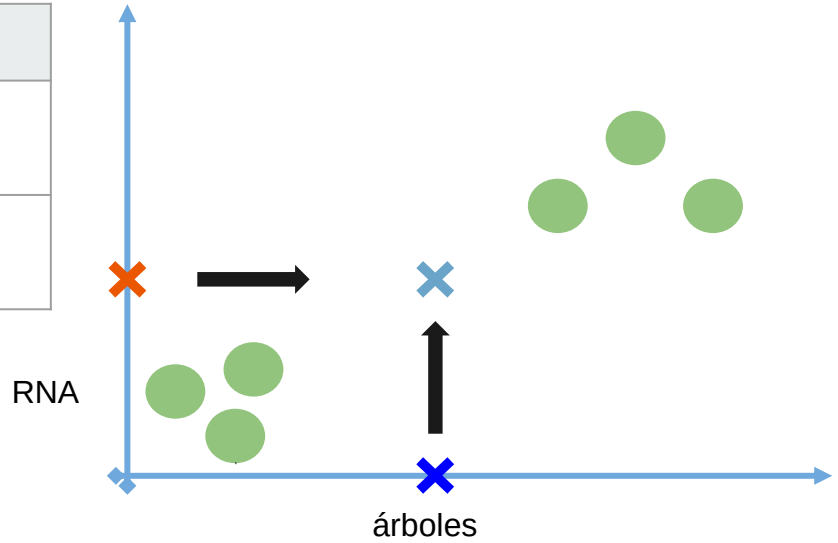
Calculamos el valor promedio para RNA (variable 2)

PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	[illegible]	3	2	1
RNA	6	4	5	3	2.8	1

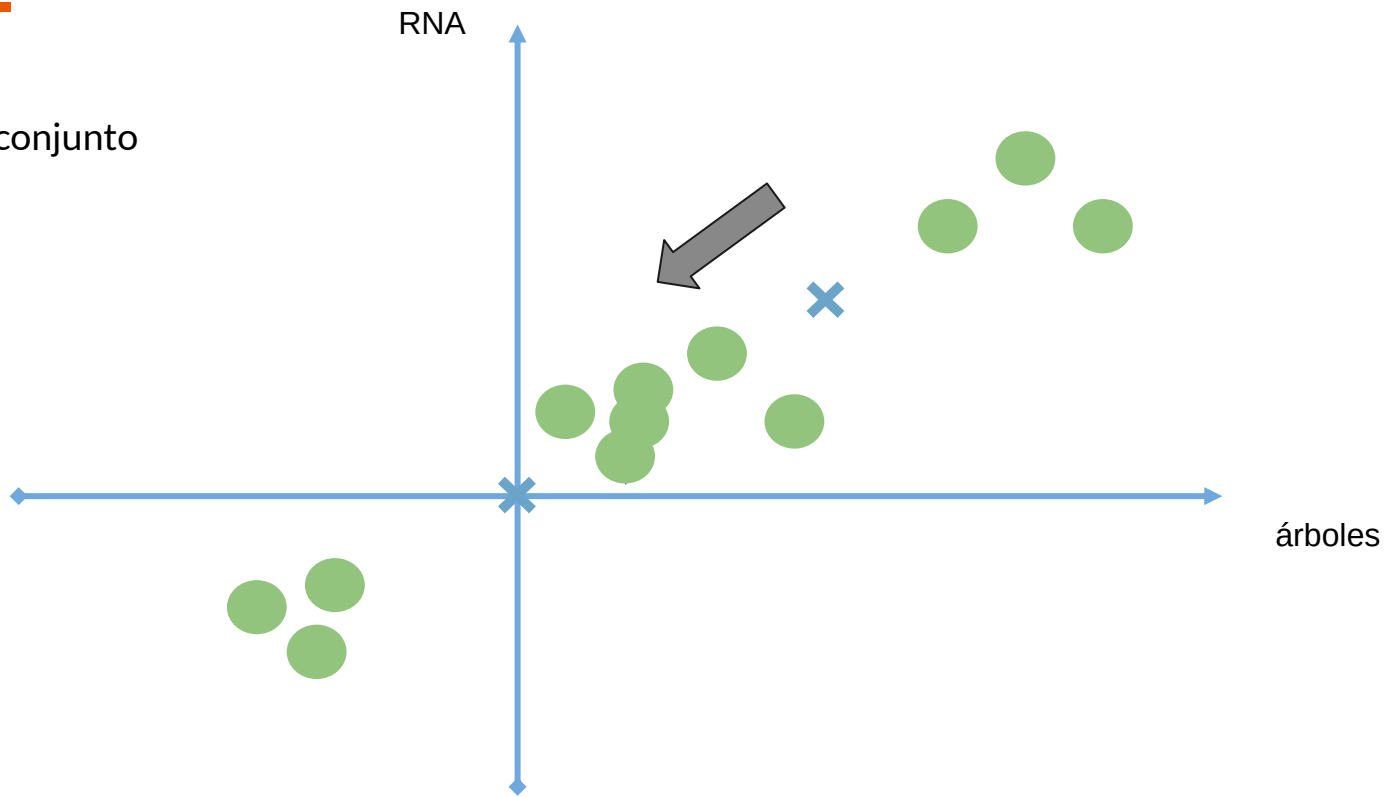
No necesitamos mirar los datos para lo que sigue



Calculamos el centro del conjunto de datos

PCA: Análisis de componentes principales

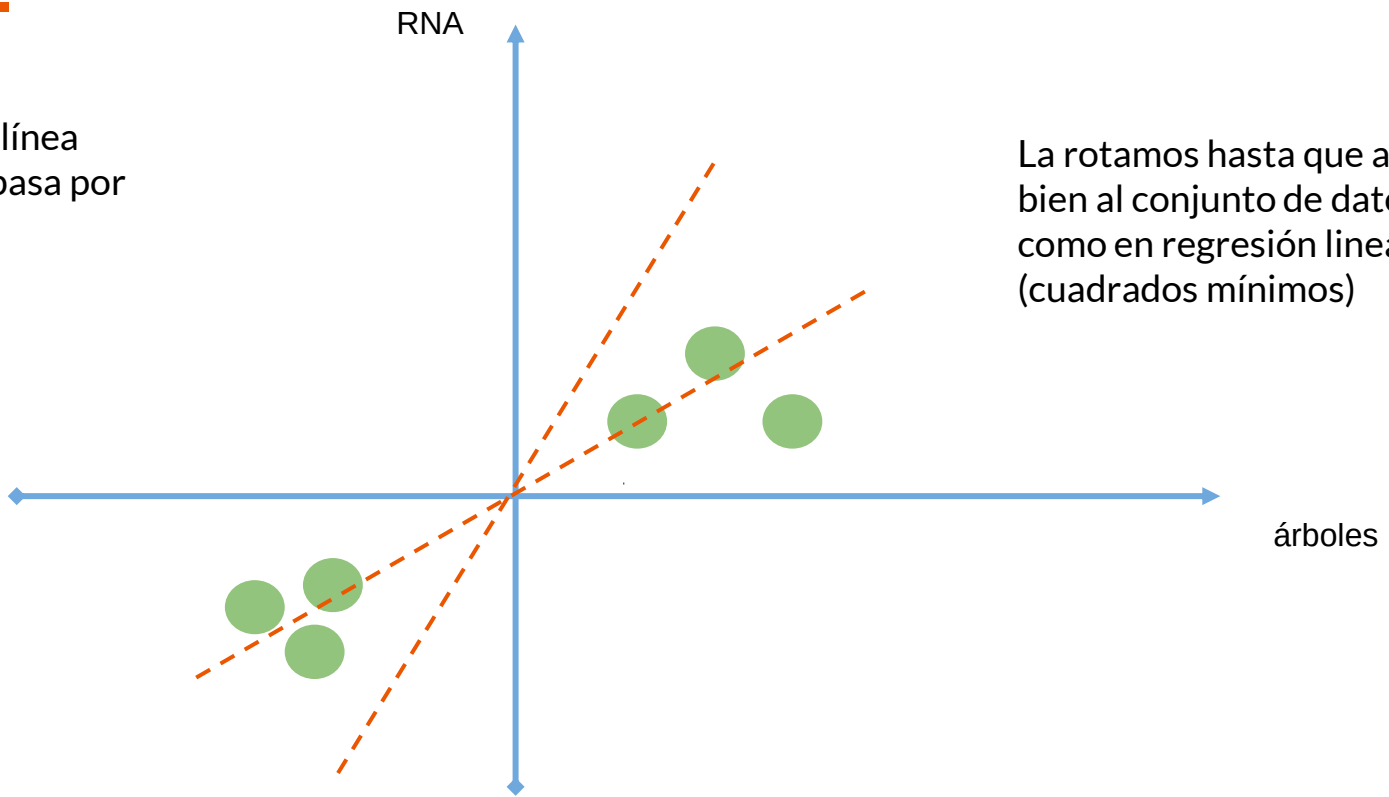
Centramos el conjunto
De datos



PCA: Análisis de componentes principales

Trazamos una línea aleatoria que pasa por el origen de coordenadas

La rotamos hasta que ajuste bien al conjunto de datos como en regresión lineal (cuadrados mínimos)

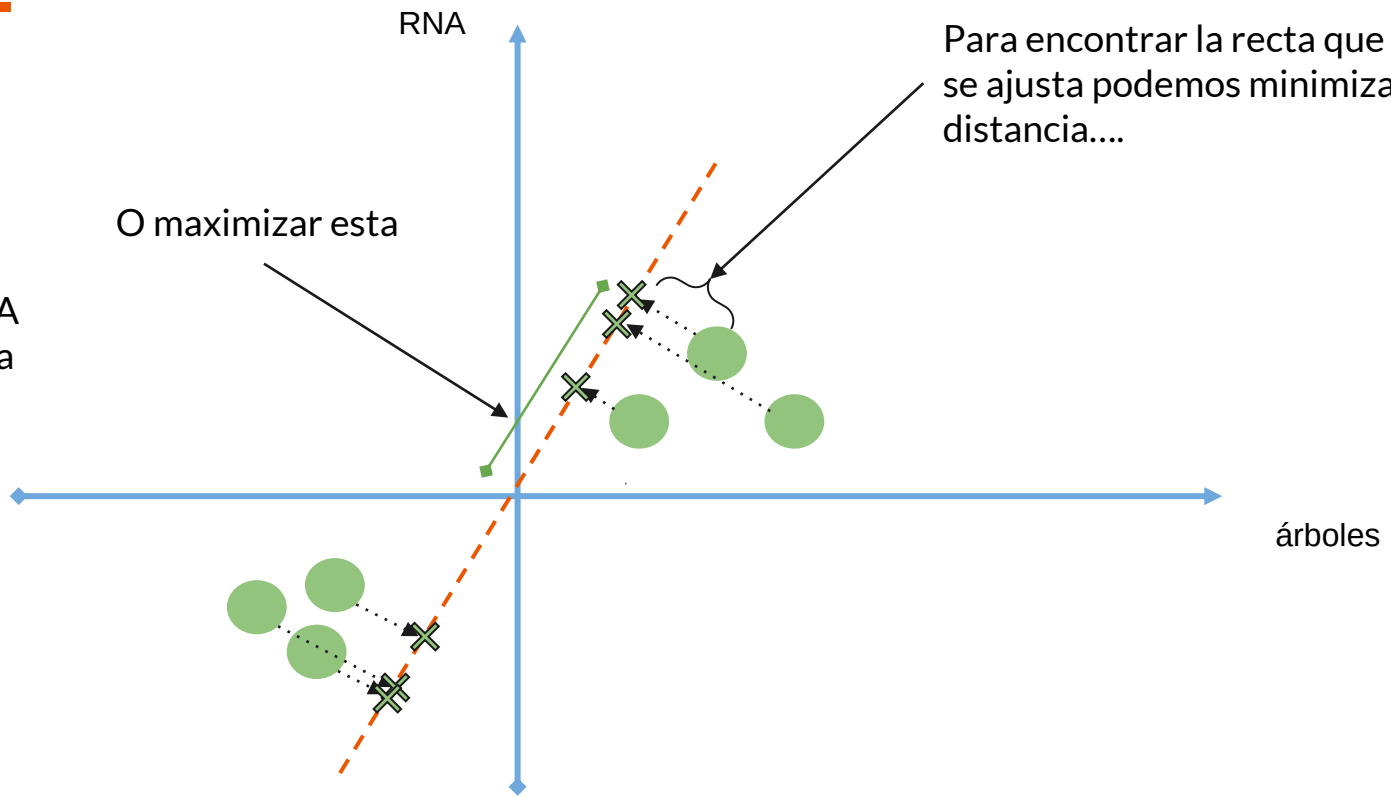


PCA por dentro

PCA: Análisis de componentes principales

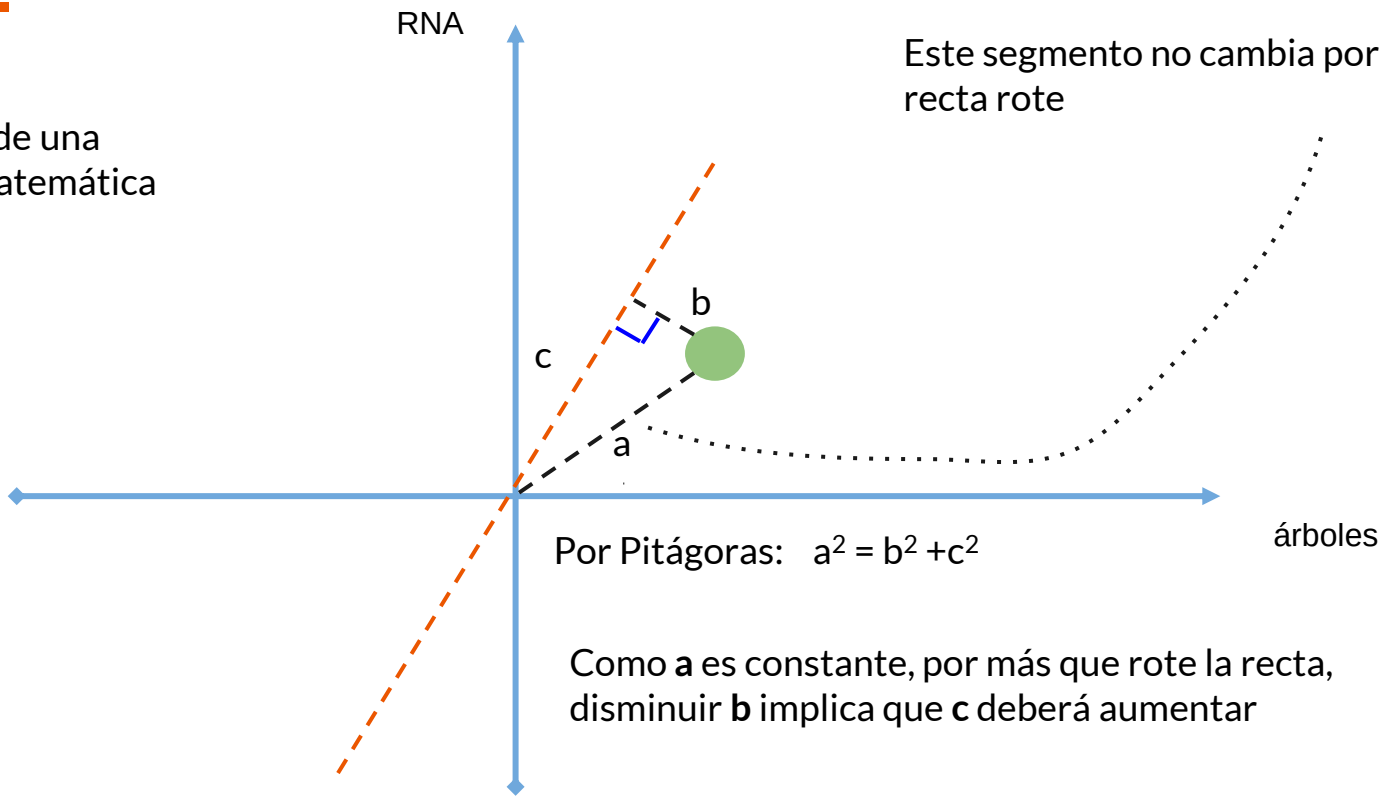
Los pasos son similares

Volviendo a la línea inicial PCA proyecta en ella cada punto….



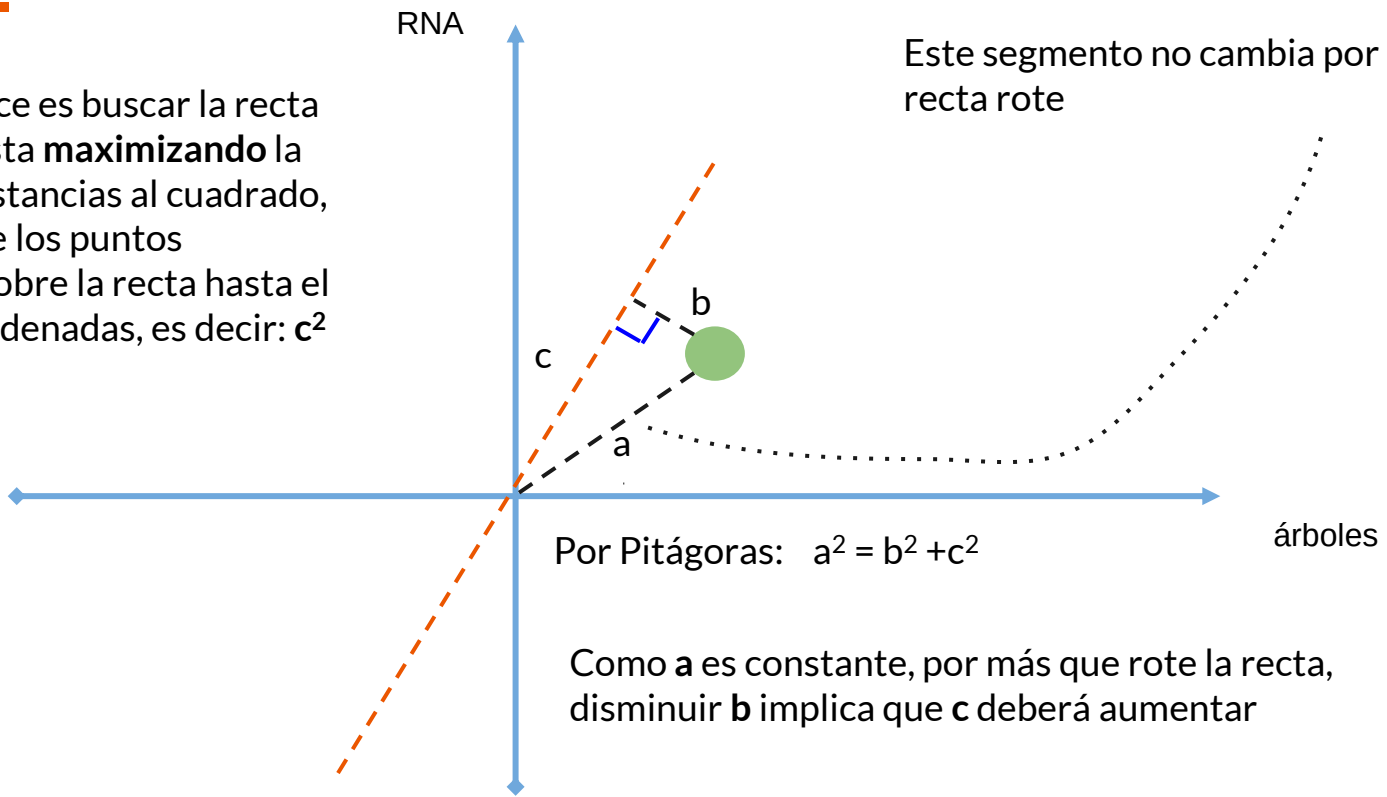
PCA: Análisis de componentes principales

Veámoslo desde una perspectiva matemática



PCA: Análisis de componentes principales

PCA lo que hace es buscar la recta que mejor ajusta **maximizando** la suma de las distancias al cuadrado, medidas desde los puntos proyectados sobre la recta hasta el origen de coordenadas, es decir: $\mathbf{c^2}$

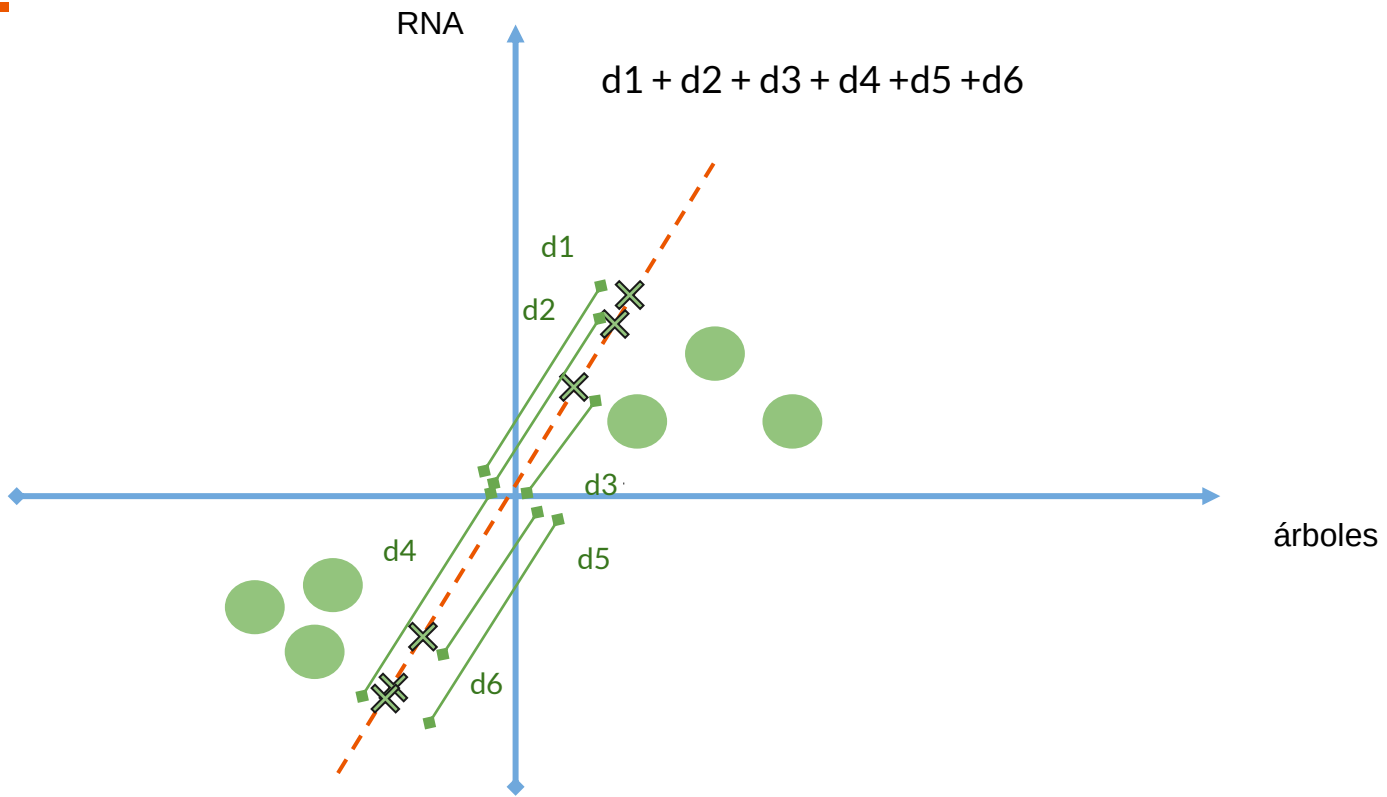


Este segmento no cambia por más que la recta rote

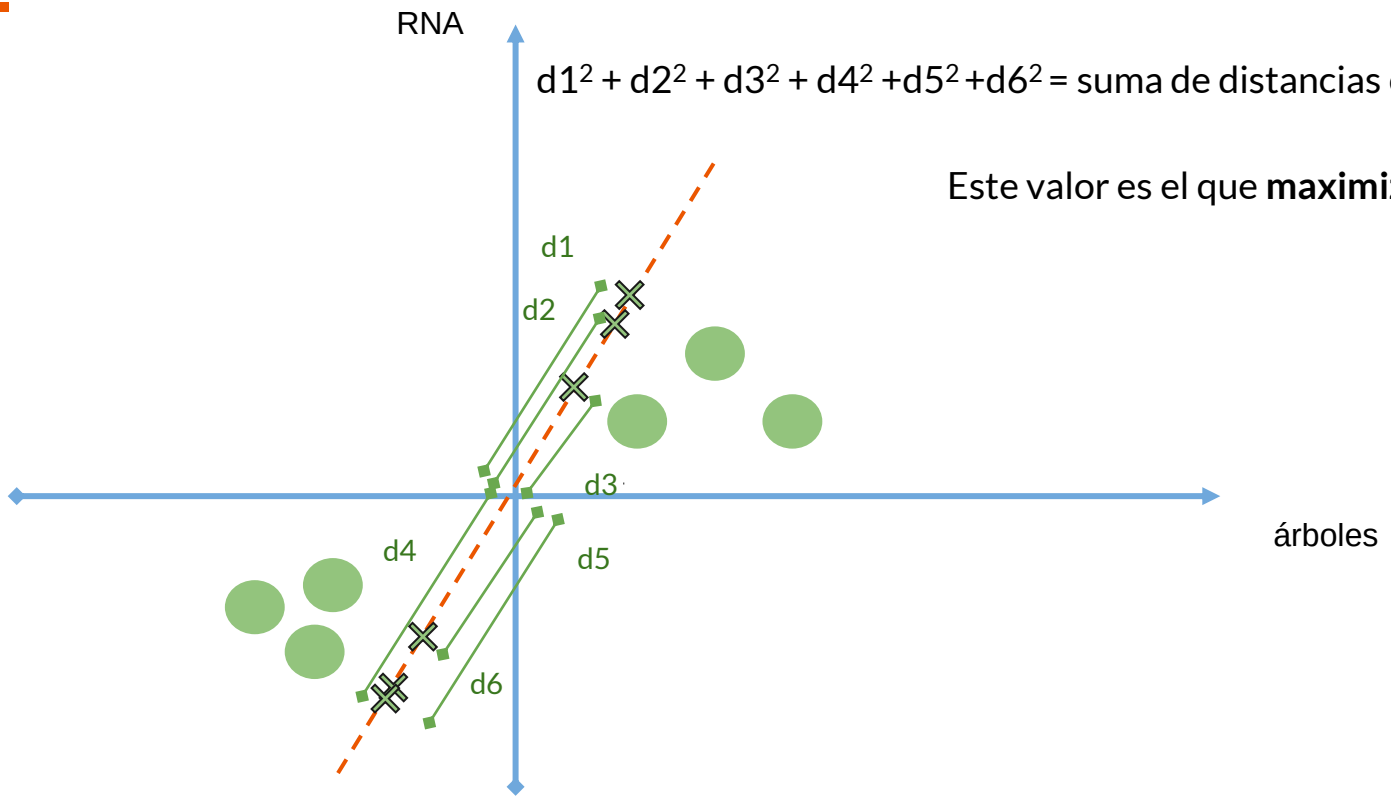
Por Pitágoras: $a^2 = b^2 + c^2$

Como **a** es constante, por más que rote la recta, disminuir **b** implica que **c** deberá aumentar

PCA: Análisis de componentes principales



PCA: Análisis de componentes principales

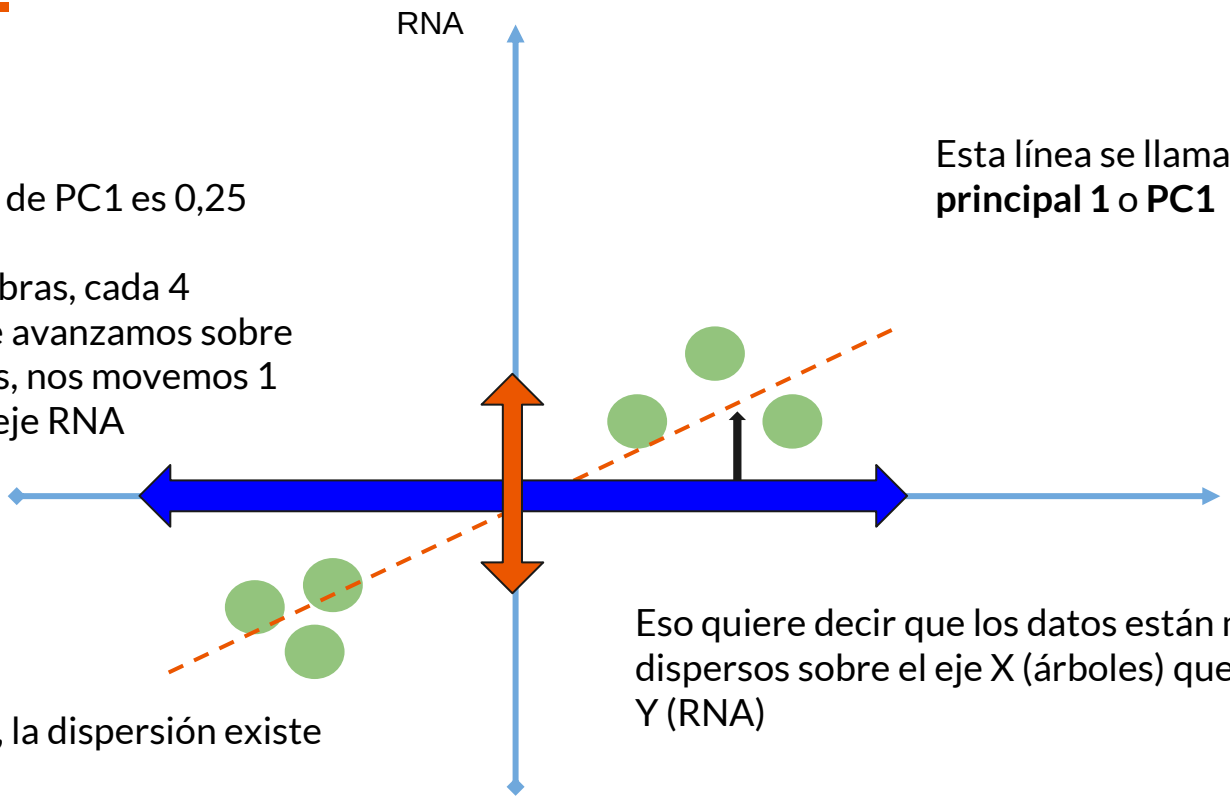


$d1^2 + d2^2 + d3^2 + d4^2 + d5^2 + d6^2$ = suma de distancias cuadradas

Este valor es el que **maximiza PCA**

Volviendo a la recta ya ajustada

PCA: Análisis de componentes principales



La pendiente de PC1 es 0,25

En otras palabras, cada 4 unidades que avanzamos sobre el eje: árboles, nos movemos 1 unidad en el eje RNA

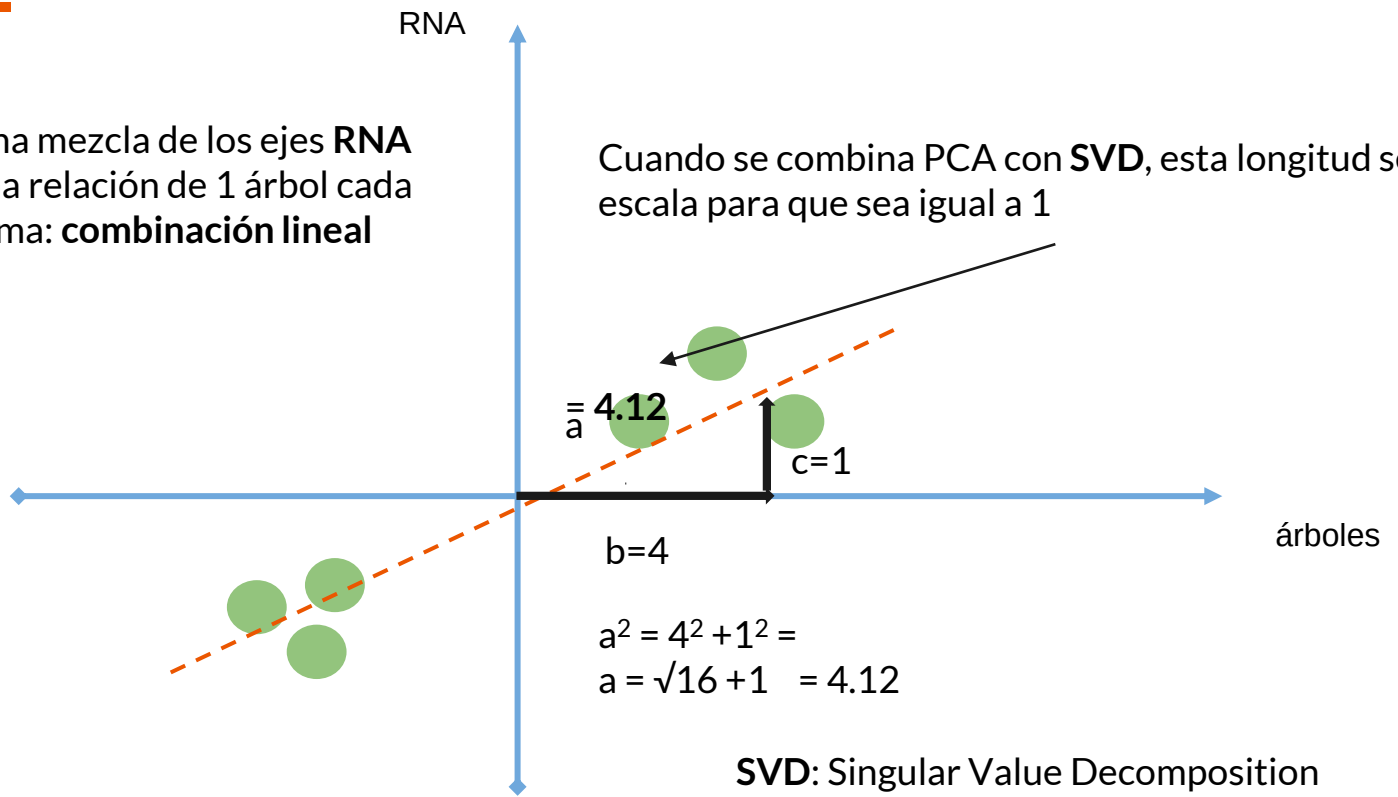
Esta línea se llama **componente principal 1** o **PC1**

Eso quiere decir que los datos están mucho más dispersos sobre el eje X (árboles) que sobre el eje Y (RNA)

Sobre el eje RNA, la dispersión existe pero es menor

PCA: Análisis de componentes principales

Cómo **PC1** es una mezcla de los ejes **RNA** y **árboles** (en una relación de 1 árbol cada 4 RNA), se la llama: **combinación lineal de variables**



Cuando se combina PCA con **SVD**, esta longitud se escala para que sea igual a 1

$a^2 = 4^2 + 1^2 =$

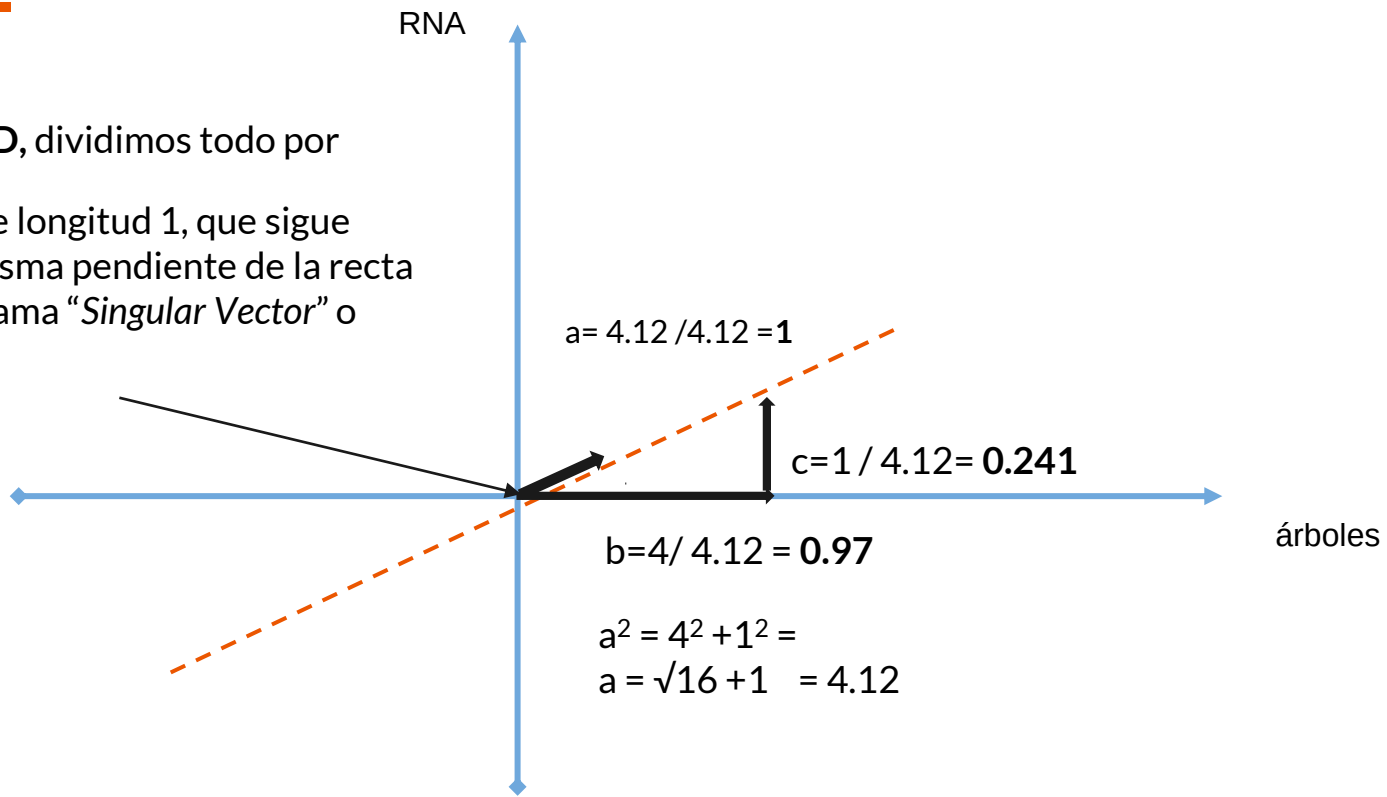
$a = \sqrt{16 + 1} = 4.12$

SVD: Singular Value Decomposition

PCA: Análisis de componentes principales

Para aplicar **SVD,** dividimos todo por 4.12

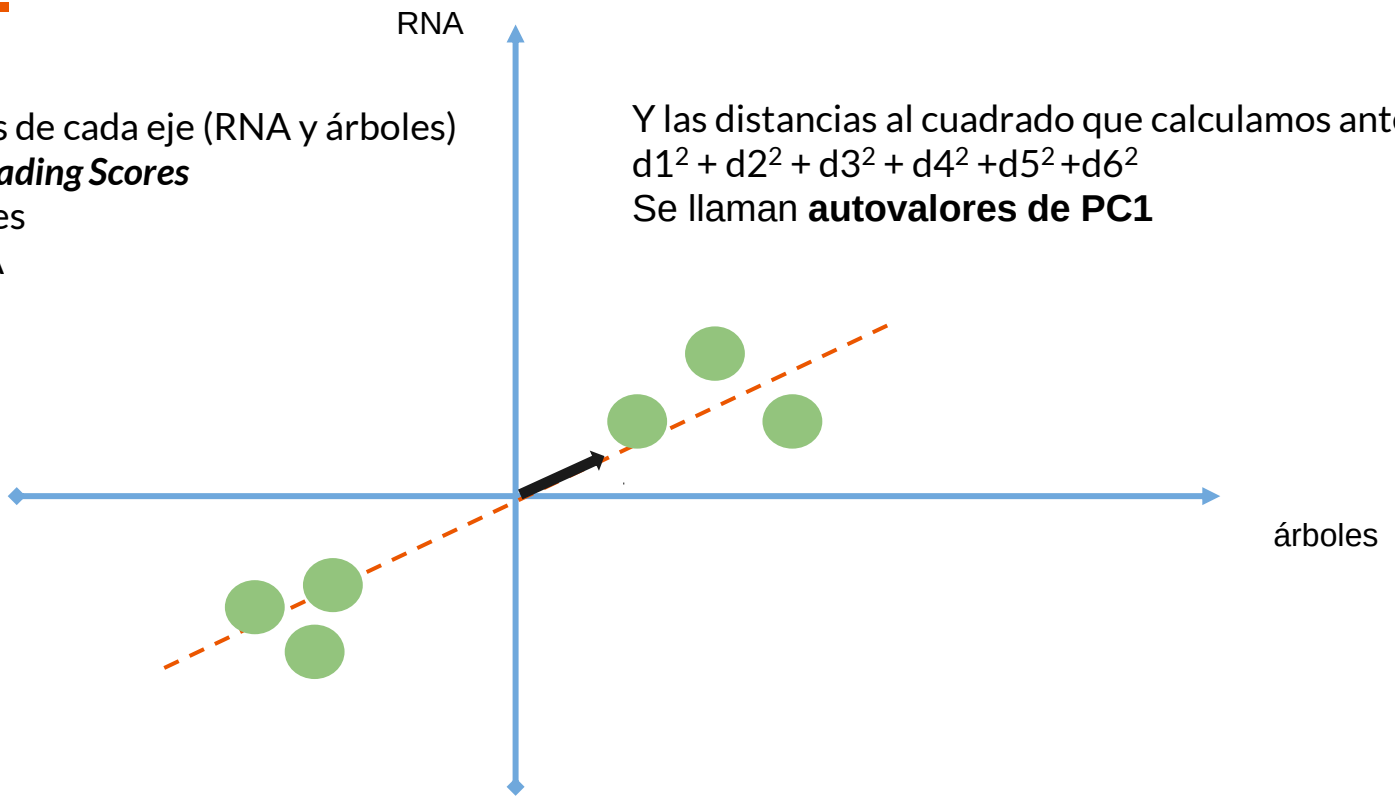
Este vector, de longitud 1, que sigue teniendo la misma pendiente de la recta calculada se llama "*Singular Vector*" o **autovector**



PCA: Análisis de componentes principales

Las proporciones de cada eje (RNA y árboles)
Son llamadas: ***Loading Scores***

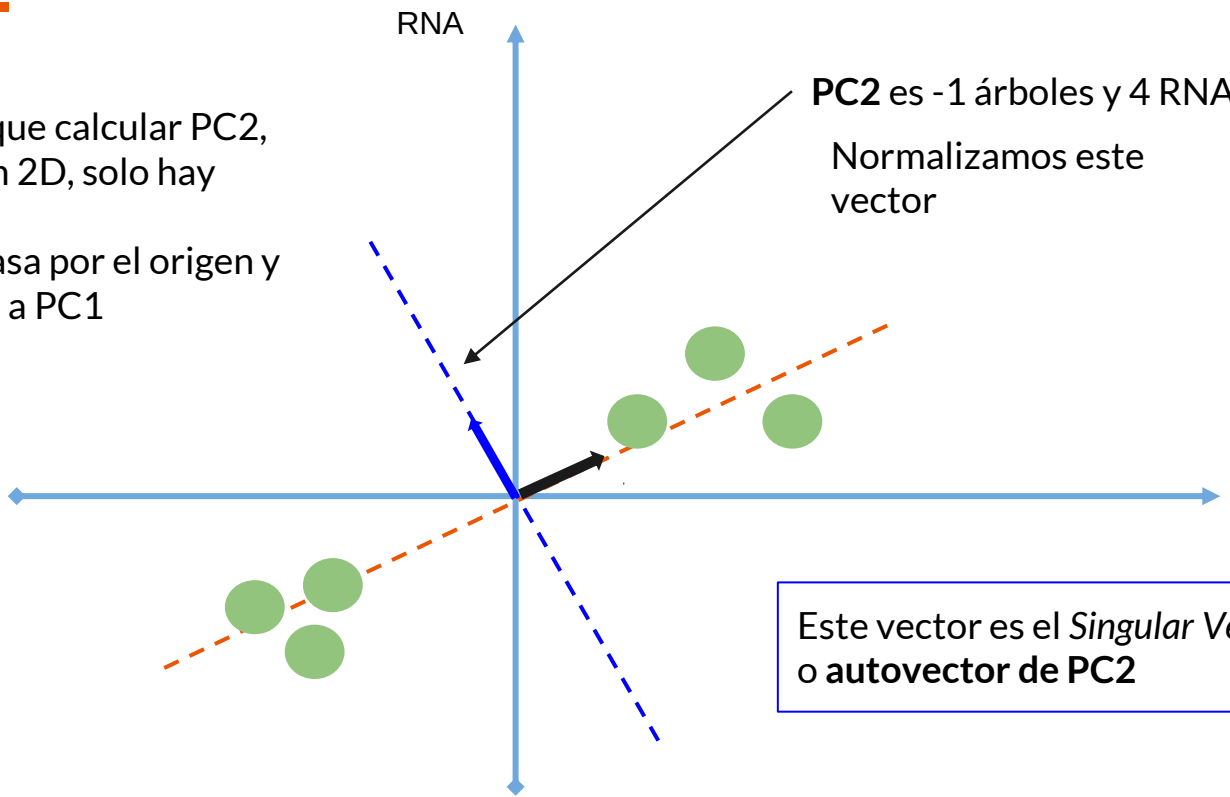
- 0,97 árboles
- 0.242 RNA



Y las distancias al cuadrado que calculamos antes:
$d1^2 + d2^2 + d3^2 + d4^2 + d5^2 + d6^2$
Se llaman **autovalores de PC1**

PCA: Análisis de componentes principales

Ahora tenemos que calcular PC2, como estamos en 2D, solo hay una posibilidad:
Solo una recta pasa por el origen y es perpendicular a PC1



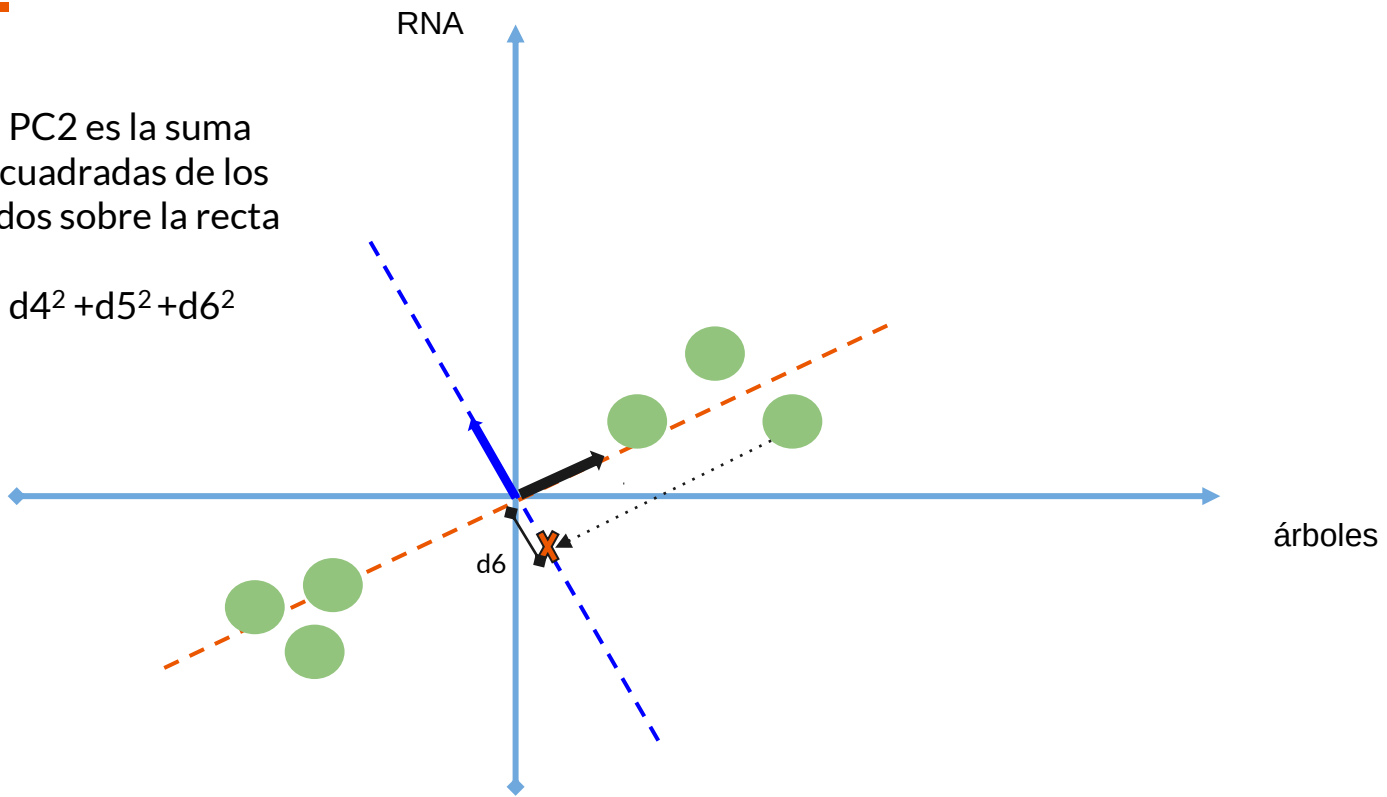
PC2 es -1 árboles y 4 RNA

Normalizamos este vector

Este vector es el *Singular Vector* de PC2 o **autovector de PC2**

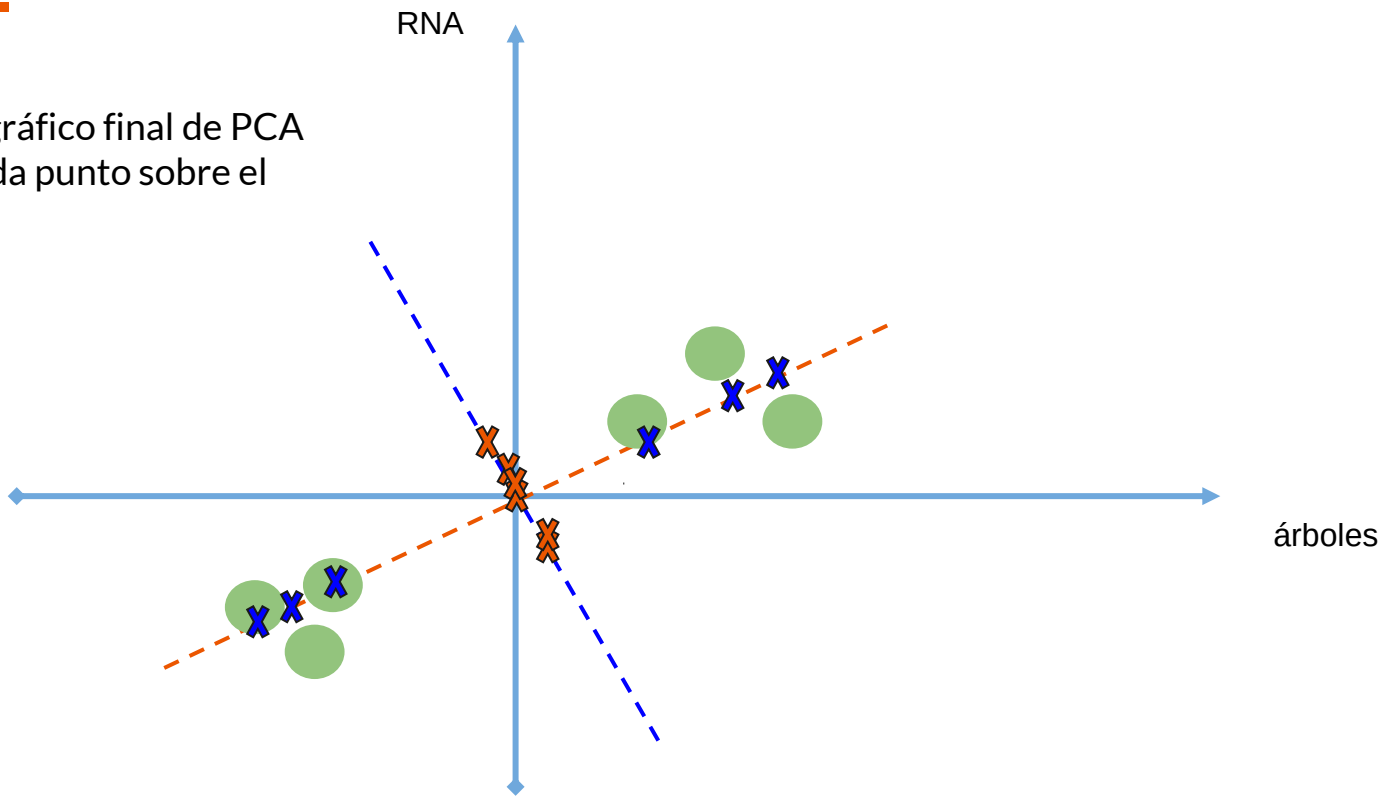
PCA: Análisis de componentes principales

El **autovector** de PC2 es la suma de las distancias cuadradas de los puntos proyectados sobre la recta PC2:
$d1^2 + d2^2 + d3^2 + d4^2 + d5^2 + d6^2$



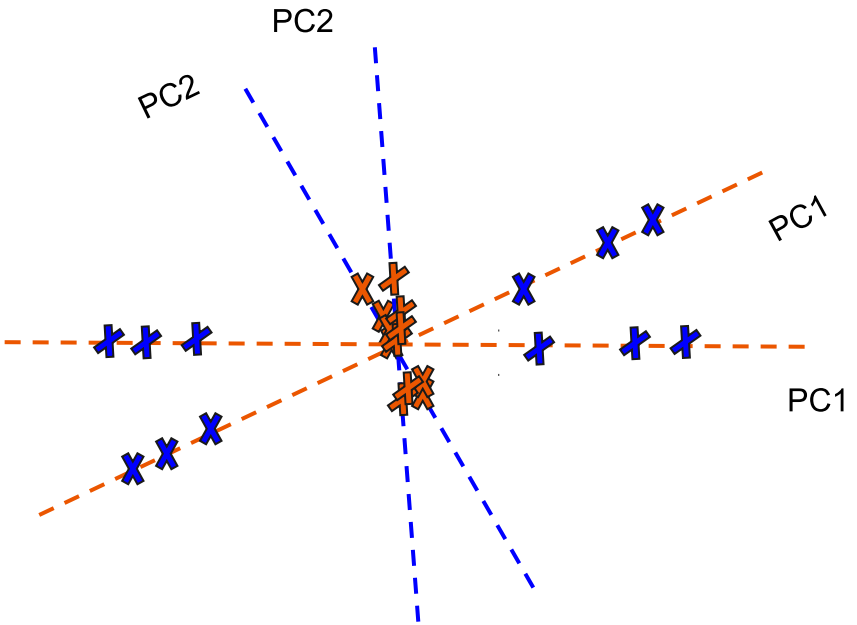
PCA: Análisis de componentes principales

Para graficar el gráfico final de PCA
Proyectamos cada punto sobre el
eje PC2 y PC 1



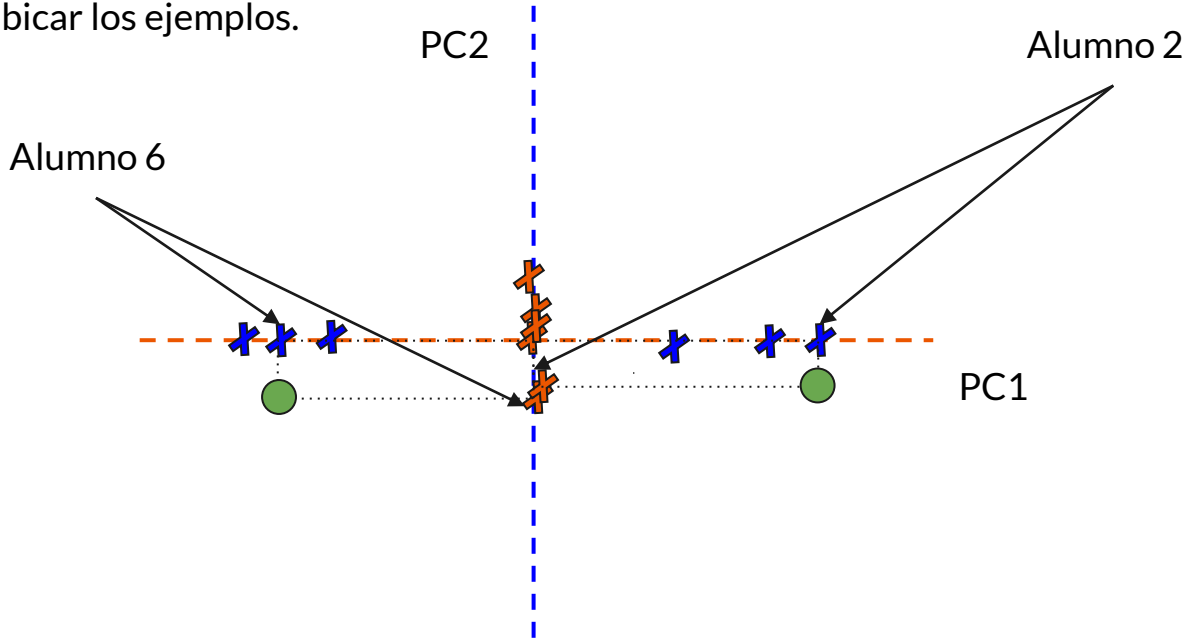
PCA: Análisis de componentes principales

Rotamos los ejes, dejando a PC1, horizontal



PCA: Análisis de componentes principales

Usamos los puntos proyectados en los ejes para reubicar los ejemplos.



PCA: Análisis de componentes principales

Suma distancias cuadradas de PC1 = autovalor de PC1
Suma distancias cuadradas de PC2 = autovalor de PC2

Podemos calcular la variación centrada en el origen:

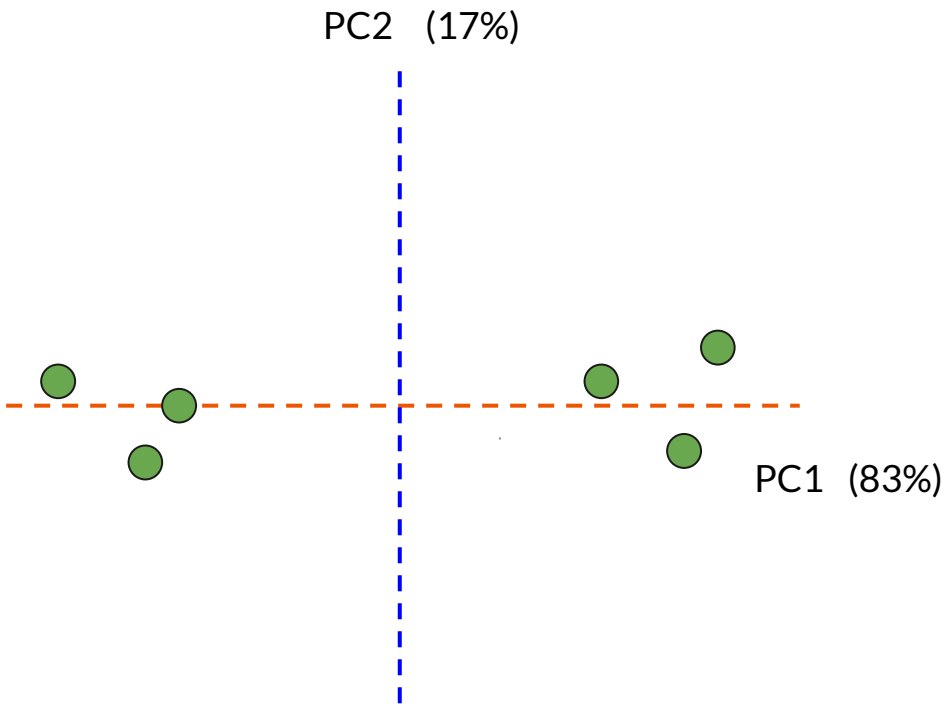
$$\frac{\text{Suma distancias cuadradas de PC1}}{n-1} = \text{variación de PC1}$$

$$\frac{\text{Suma distancias cuadradas de PC2}}{n-1} = \text{variación de PC2}$$

Imaginemos que la variación de PC1 es **15**, y la de PC2 es **3**.
La variación total es 15+3 = 18

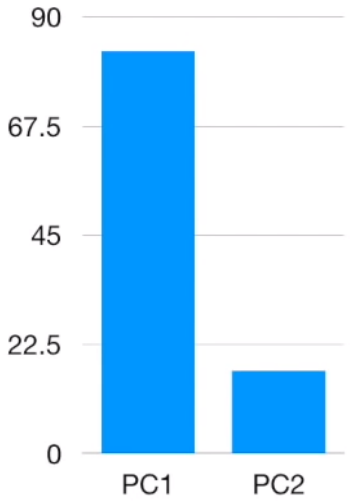
Eso quiere decir que PC1 acumula el **83 %** de la variación total. (15/18 = 0.83)

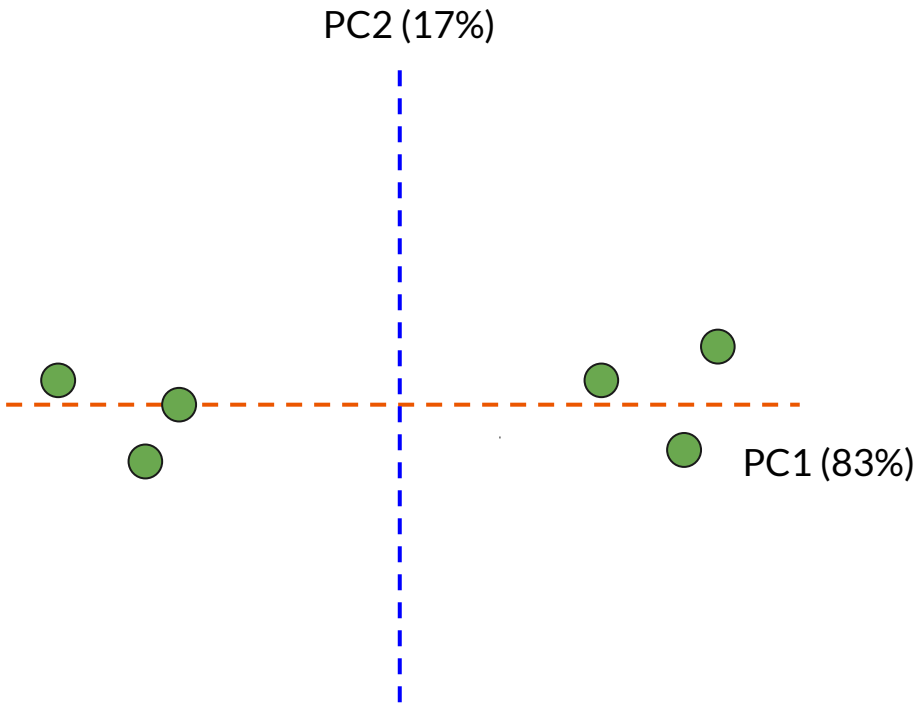
Y PC2, el **17%** (3/18 = 0.17)



PCA: Análisis de componentes principales

Scree plot es un gráfico donde se representa el porcentaje de variación de cada PC (componente principal)



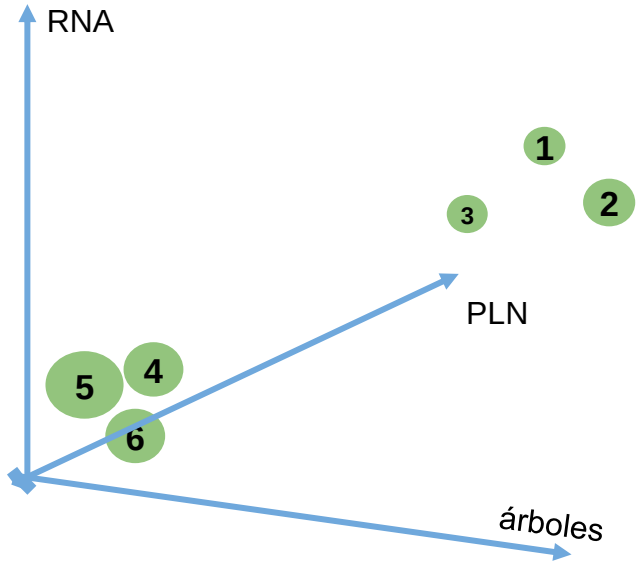


3 Variables = 3 dimensiones

PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	8	3	2	1
RNA	6	4	5	3	2.8	1
PLN	12	9	10	2.5	1.3	2



PCA: Análisis de componentes principales

Los pasos son los mismos

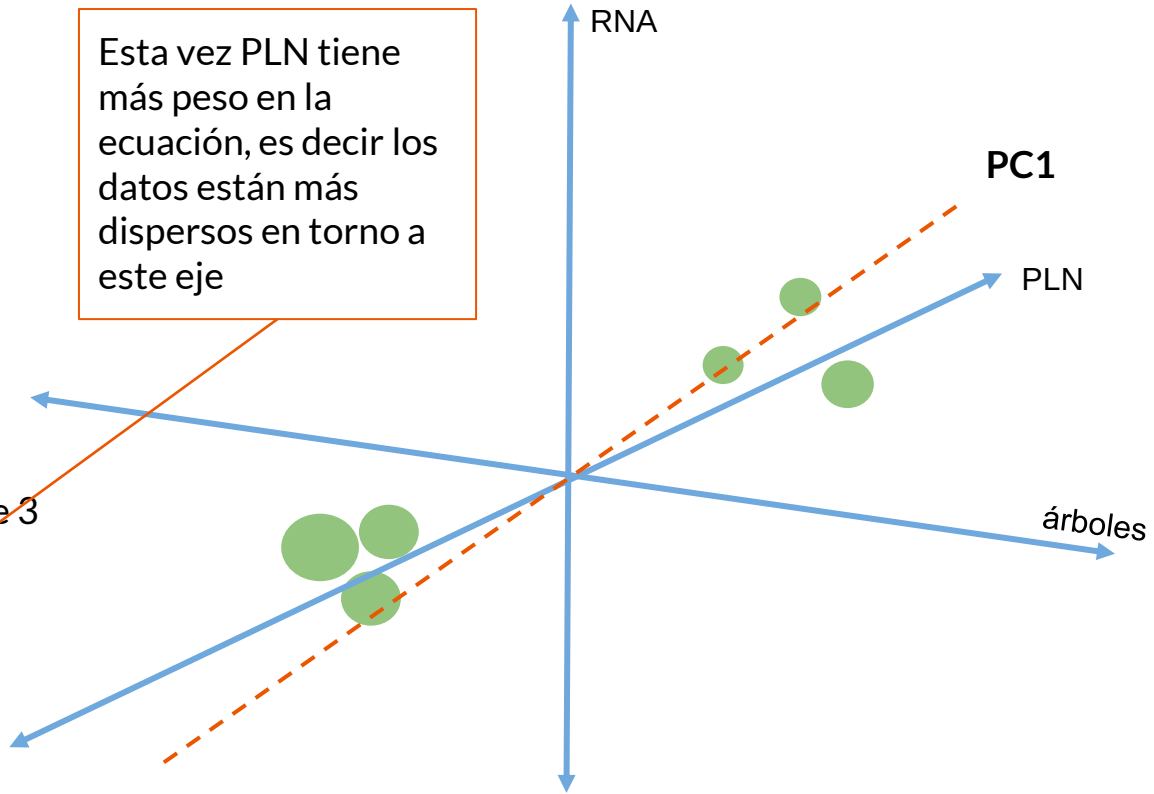
1. Centramos los datos en el eje de coordenadas
2. Buscamos la recta que mejor ajuste que pase por el eje de coordenadas

La recta será PC1

Solo que en este caso, la recta tiene 3 coordenadas:
Árboles, RNA y PLN

- **0.62 árboles**
- **0.15 RNA**
- **0.77 PLN**

Esta vez PLN tiene más peso en la ecuación, es decir los datos están más dispersos en torno a este eje



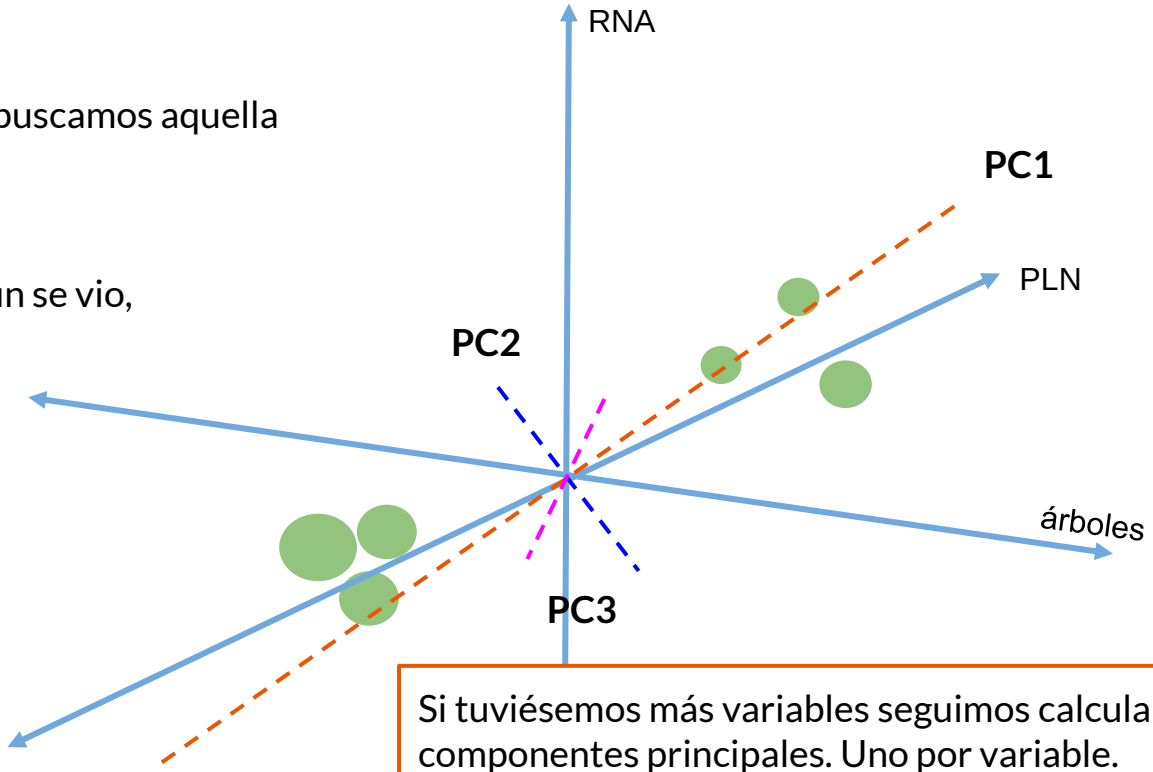
PCA: Análisis de componentes principales

Luego buscamos PC2
Esta vez hay varias posibilidades, buscamos aquella que:

1. Pase por el origen
2. Sea perpendicular a PC1
3. Y tenga el mejor ajuste, según se vio, maximizando su autovector

Finalmente calculamos PC3:

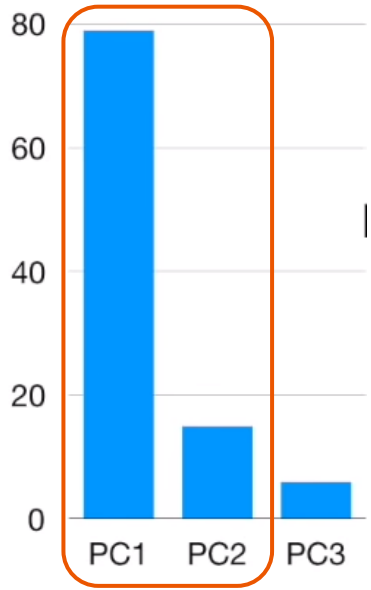
- Perpendicular a las otras dos
- Pase por el origen



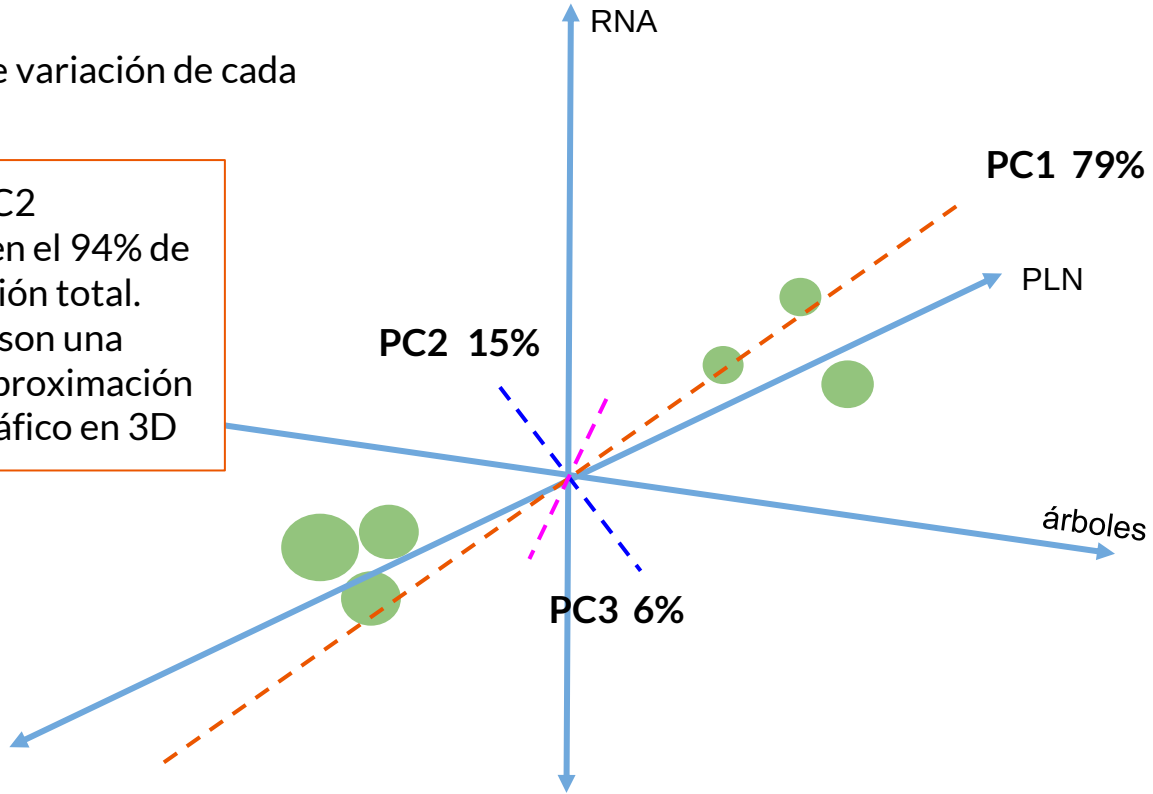
Si tuviésemos más variables seguimos calculando componentes principales. Uno por variable.

PCA: Análisis de componentes principales

Calculamos luego la proporción de variación de cada componente principal.



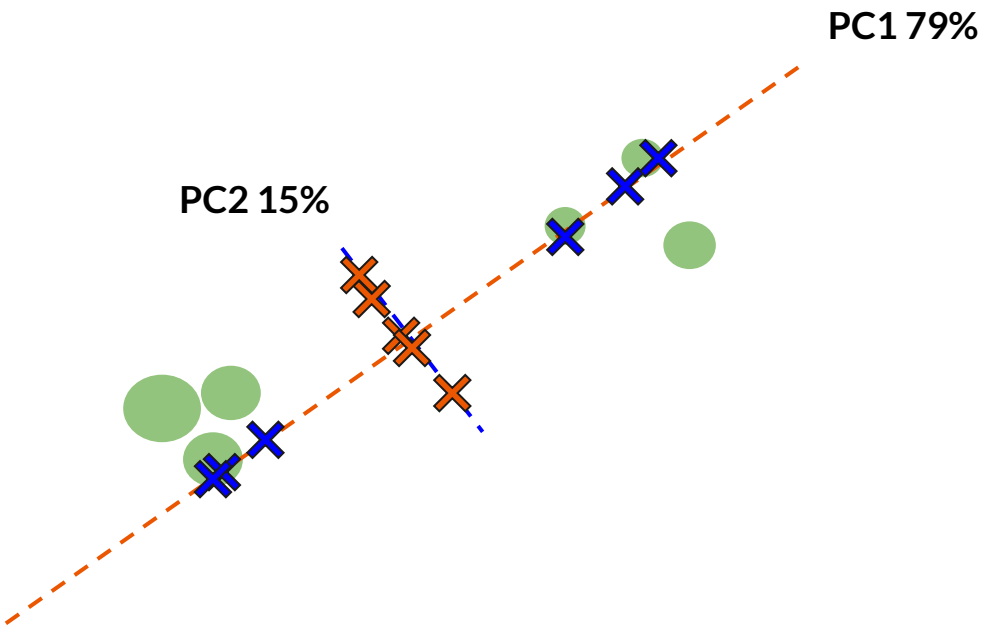
PC1 y PC2 contienen el 94% de la variación total. Es decir son una buena aproximación de un gráfico en 3D



PCA: Análisis de componentes principales

Para convertir el gráfico actual en un gráfico de 2D.
Eliminamos todo, dejamos solo PC1 y PC2

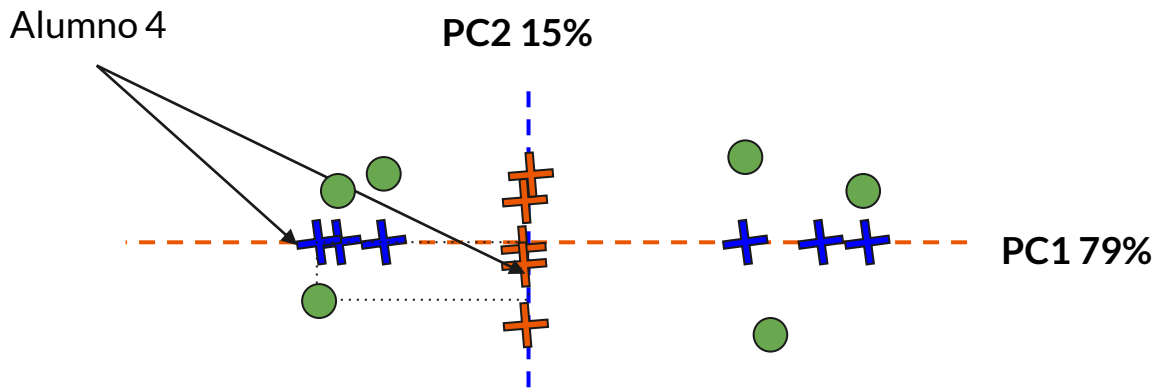
Proyectamos los ejemplos sobre los ejes



PCA: Análisis de componentes principales

Rotamos el gráfico hasta que PC1 quede horizontal.

Graficamos los ejemplos usando las proyecciones



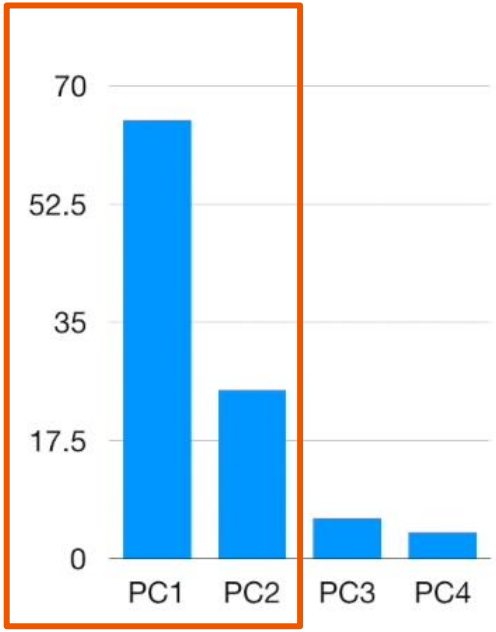
4D

PCA: Análisis de componentes principales

Puntaje por tema en un examen de ciencia de datos

Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	8	3	2	1
RNA	6	4	5	3	2.8	1
PLN						
SVM						

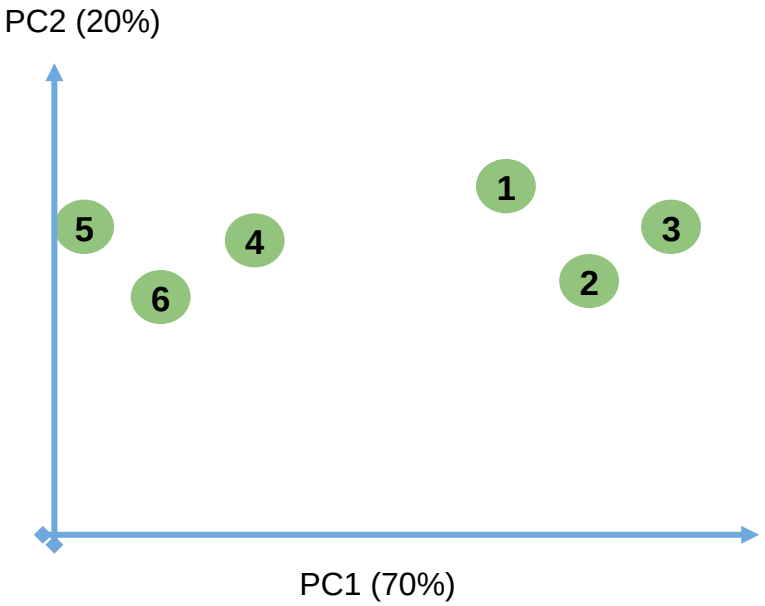
PC1 y PC2 se llevan el 90% de la variación total, eso quiere decir que podemos tener una buena representación 2D, de cuán dispersos están los datos.



PCA: Análisis de componentes principales

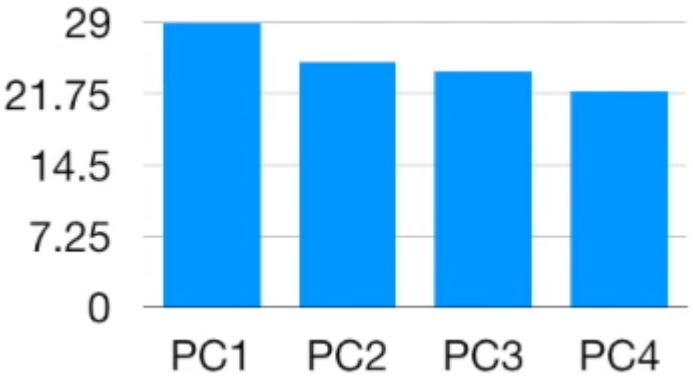
Puntaje por tema en un examen de ciencia de datos

Puntos	A 1	A 2	A 3	A 4	A 5	A 6
árboles	9	10	8	3	2	1
RNA	6	4	5	3	2.8	1
PLN	12	9	10	2.5	1.3	2
SVM	5	7	6	2	4	7

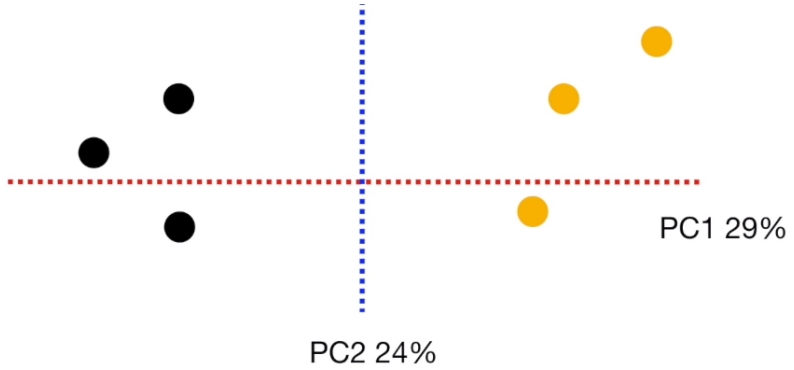


PCA: Análisis de componentes principales

¿Qué pasa si el gráfico de Scree es así?



Usar la representación de PC1 y PC2 no será suficientemente buena para ver la dispersión de los datos.
De todas formas, aún es factible detectar agrupamiento de ejemplos



PCA: Análisis de componentes principales

Resumen:

- El análisis de componentes principales, o PCA, nos sirve para identificar si hay agrupamiento de datos en el espacio de entrada.
- Podemos identificar correlaciones, *clusters* o bien entender cuán dispersos están los datos y sobre todo, sobre qué ejes o variables.
- Es útil especialmente cuando no podemos representar el espacio de entrada sobre un eje cartesiano.