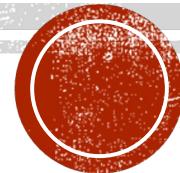
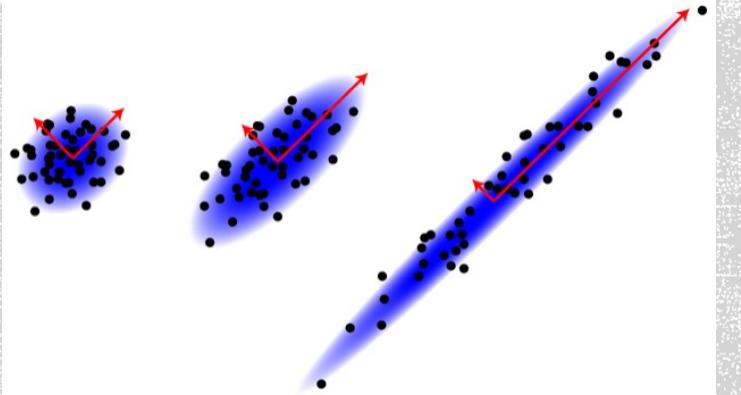


COMPONENTES PRINCIPALES

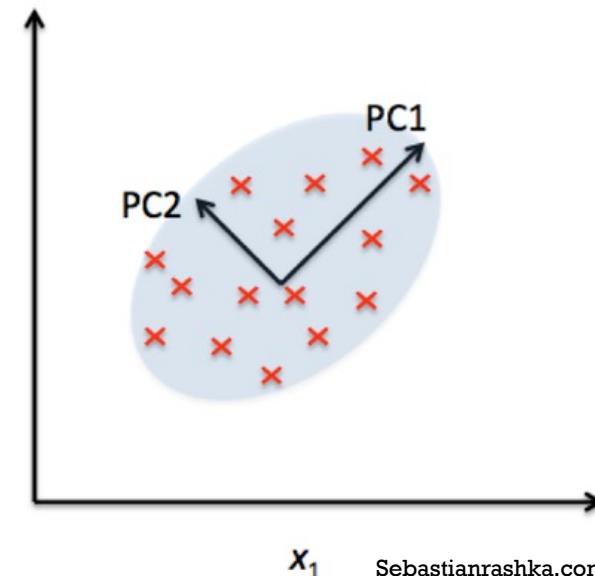


COMPONENTES PRINCIPALES

PCA: Principal Component Analysis

Objetivo: Simplificar el dataset, encontrando una representación de **baja dimensionalidad** que conserva la mayor parte de la información

- **Combinación lineal** de las dimensiones (atributos) originales del dataset que maximiza la varianza
- **Rotación** de los ejes originales
- Permite una **visualización** los datos en problemas de aprendizaje supervisado y no supervisado
- Se limitan las dimensiones que estén altamente **correlacionadas** entre ellas
- PCA permite encontrar la superficie lineal de menos dimensiones más cercana a los puntos en el espacio original (en distancia Euclidiana)



Sebastianrashka.com



COMPONENTES PRINCIPALES

- Hay tantos componentes principales (PCs) como dimensiones, ortogonales entre ellos
- Cada PC es una combinación lineal normalizada de los atributos del dataset (X_1, X_2, \dots, X_N), se buscan los vectores que maximicen la varianza

$$PC_i = \Phi_{1i}X_1 + \Phi_{2i}X_2 + \dots + \Phi_{Ni}X_N, \text{ sujeto a } \sum_{j=1}^N \Phi_{ji}^2 = 1$$

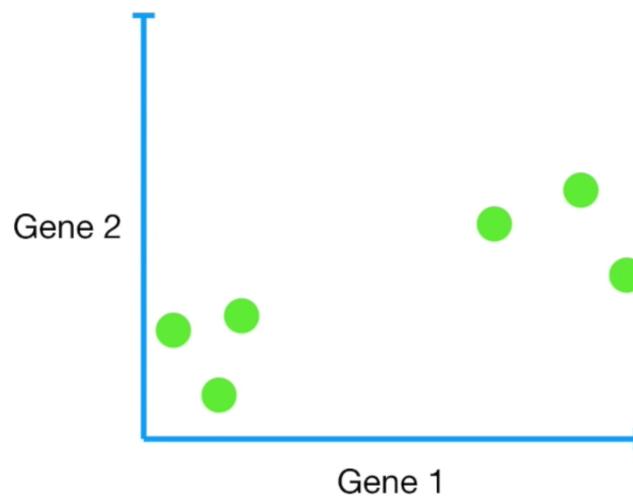
Además $\forall j < i, PC_i \perp PC_j$

- Cada PC tiene asociada una carga o **loading** de cada una de las dimensiones originales (los Φ_{ji}). El vector de loadings de una variable original indica su dirección en el espacio de los PC
- Solo existe una solución posible de espacio de componentes principales, conservando siempre las direcciones aunque puede que el sentido sea el contrario.
- A cada PC se le puede establecer la cantidad de información original especificada (Proporción de varianza explicada). Esta va decreciendo con cada PC considerado, por lo que los primeros p PCs van a representar mucha más información que las primeras p dimensiones originales
- Las instancias originales se proyectan en el espacio dado por los primeros p PCs



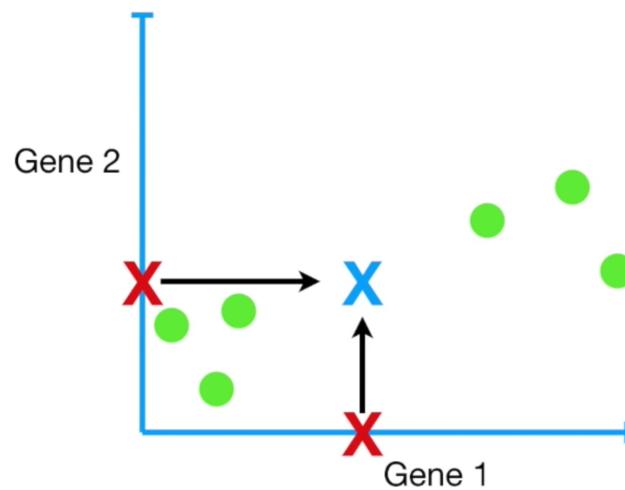
COMPONENTES PRINCIPALES

	Mouse					
	1	2	3	4	5	6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

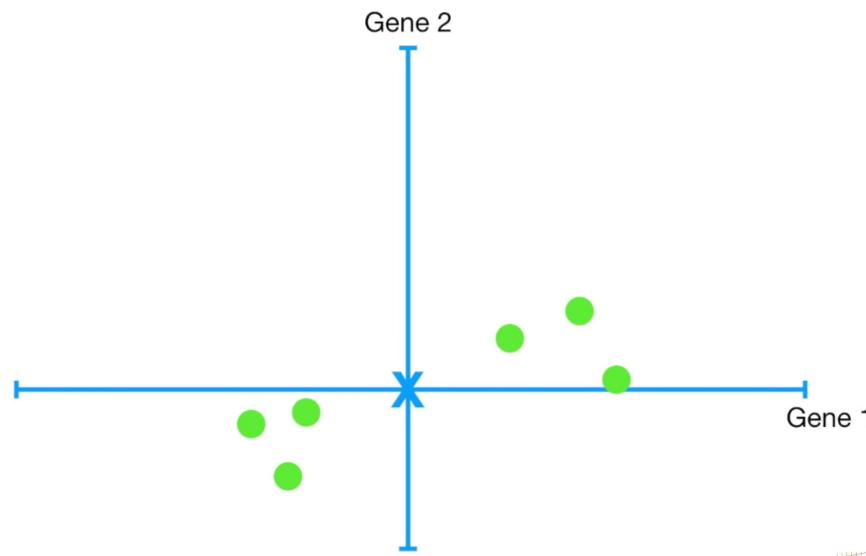


COMPONENTES PRINCIPALES

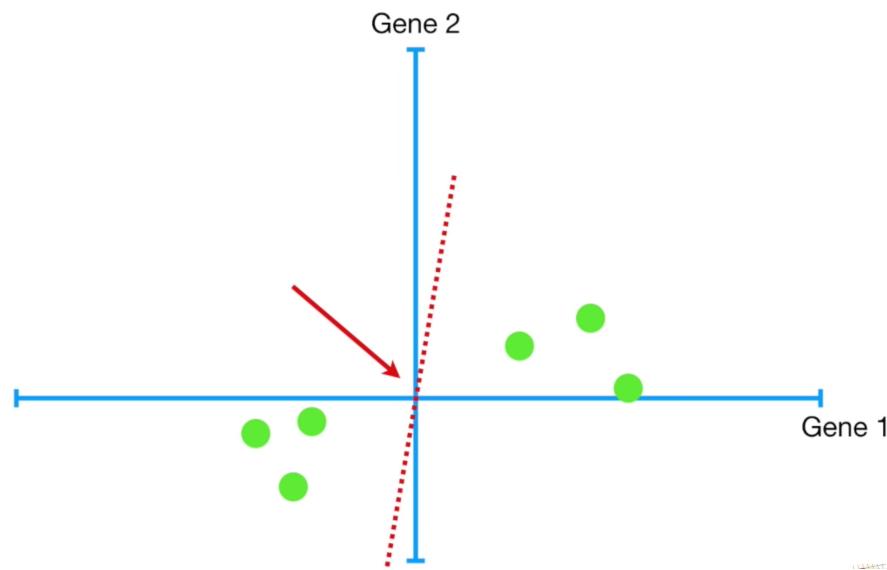
	Mouse					
	1	2	3	4	5	6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



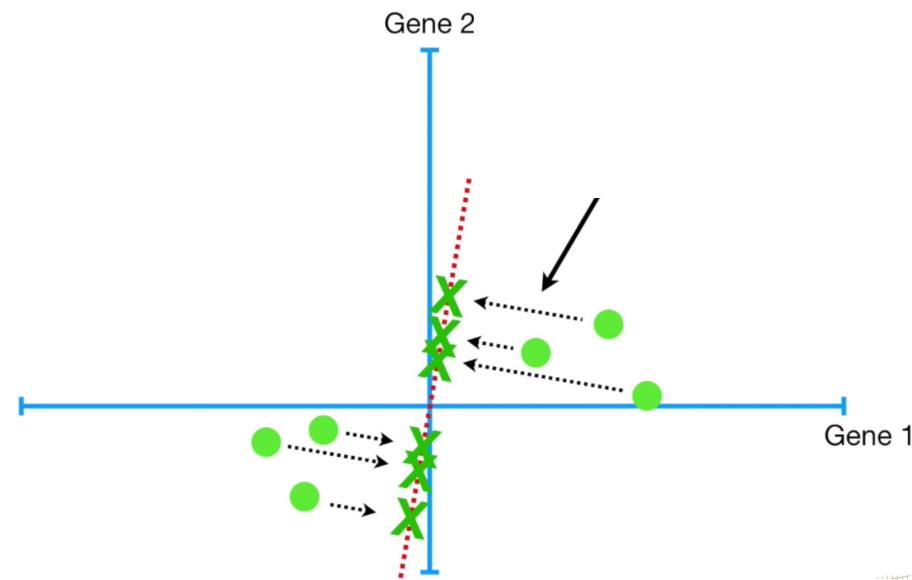
COMPONENTES PRINCIPALES



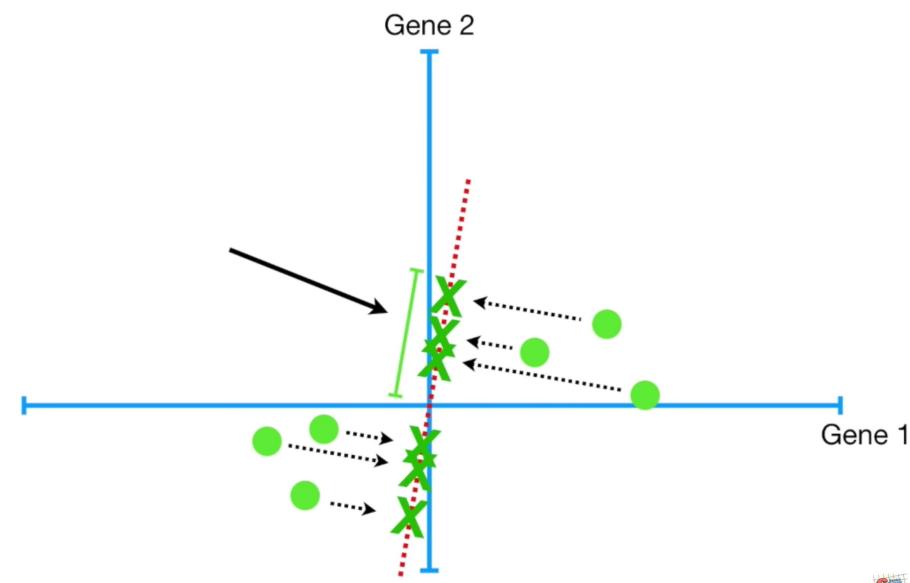
COMPONENTES PRINCIPALES



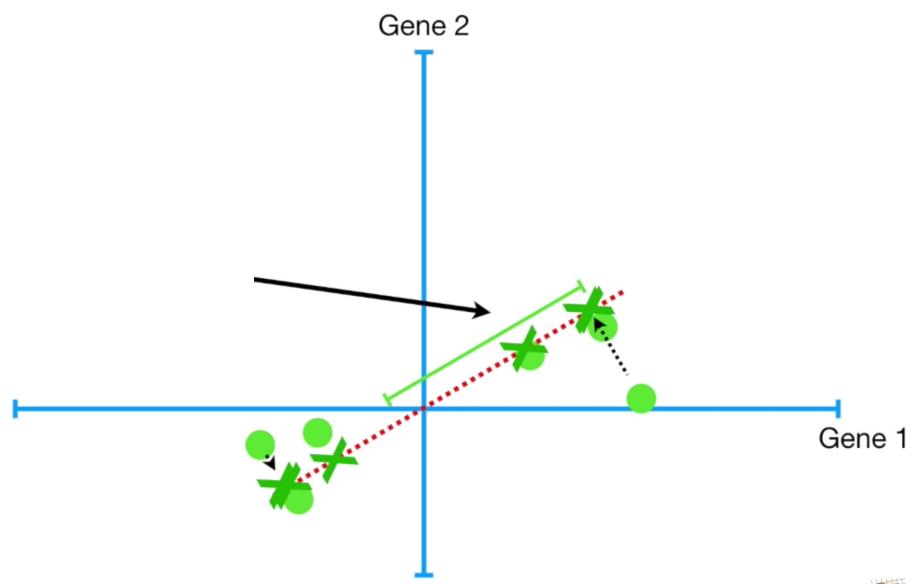
COMPONENTES PRINCIPALES



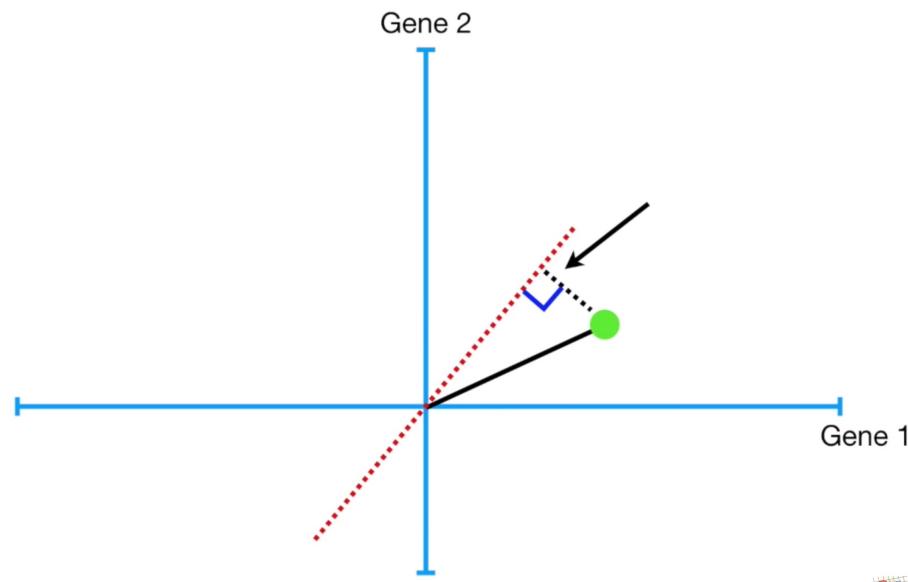
COMPONENTES PRINCIPALES



COMPONENTES PRINCIPALES



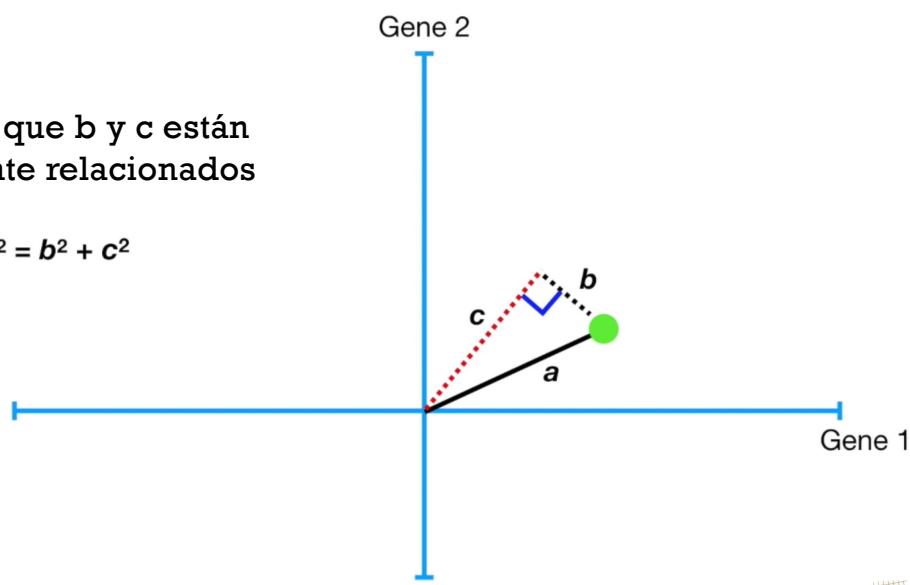
COMPONENTES PRINCIPALES



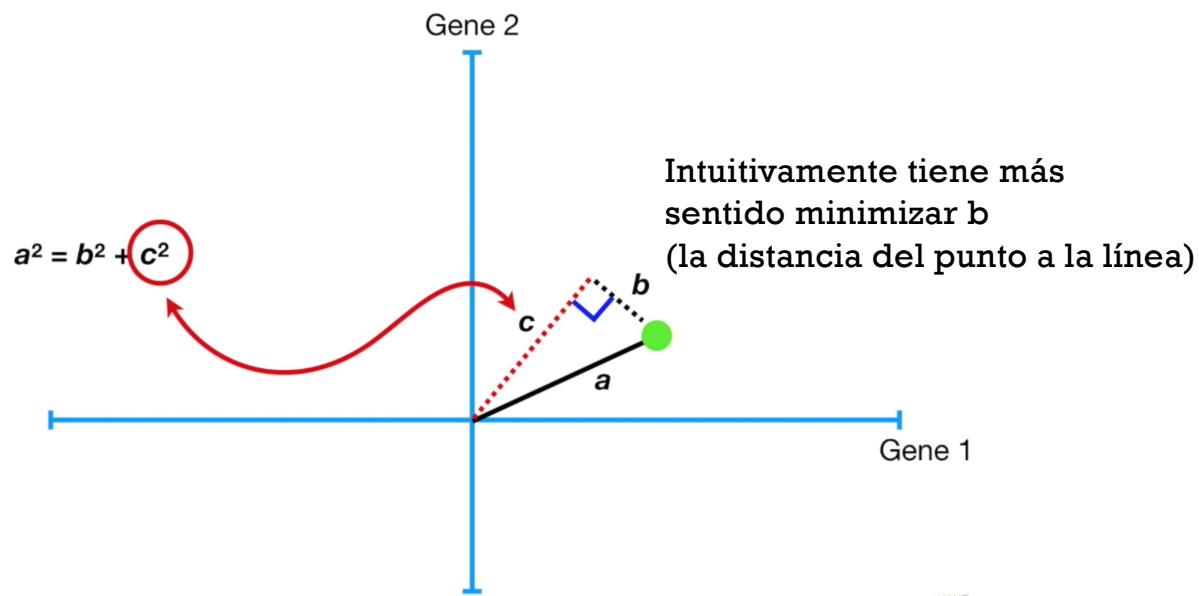
COMPONENTES PRINCIPALES

Se muestra que b y c están inversamente relacionados

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

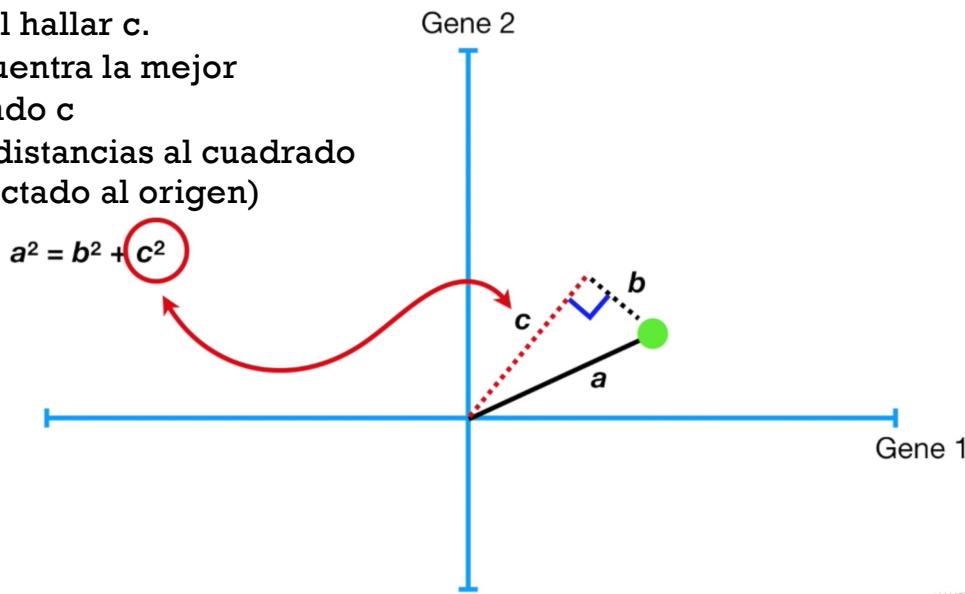


COMPONENTES PRINCIPALES

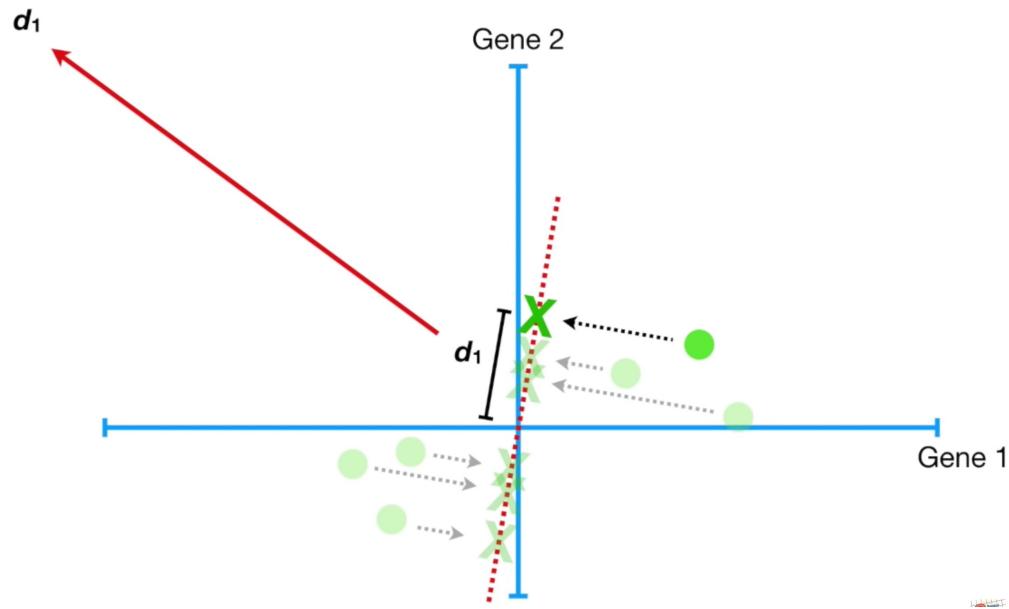


COMPONENTES PRINCIPALES

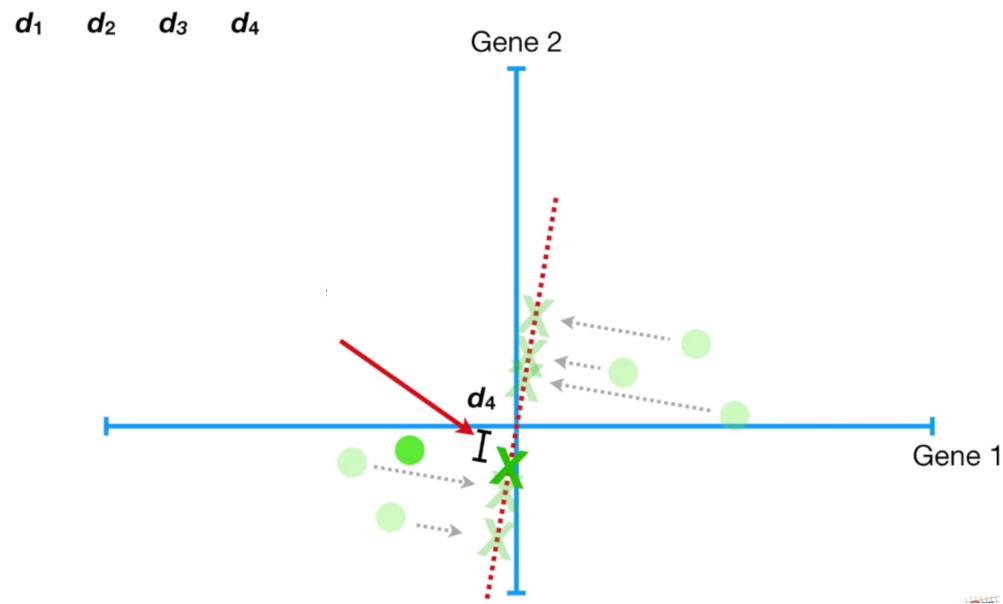
Pero es más fácil hallar c .
Luego PCA encuentra la mejor
línea maximizando c
(la suma de las distancias al cuadrado
Del punto proyectado al origen)



COMPONENTES PRINCIPALES

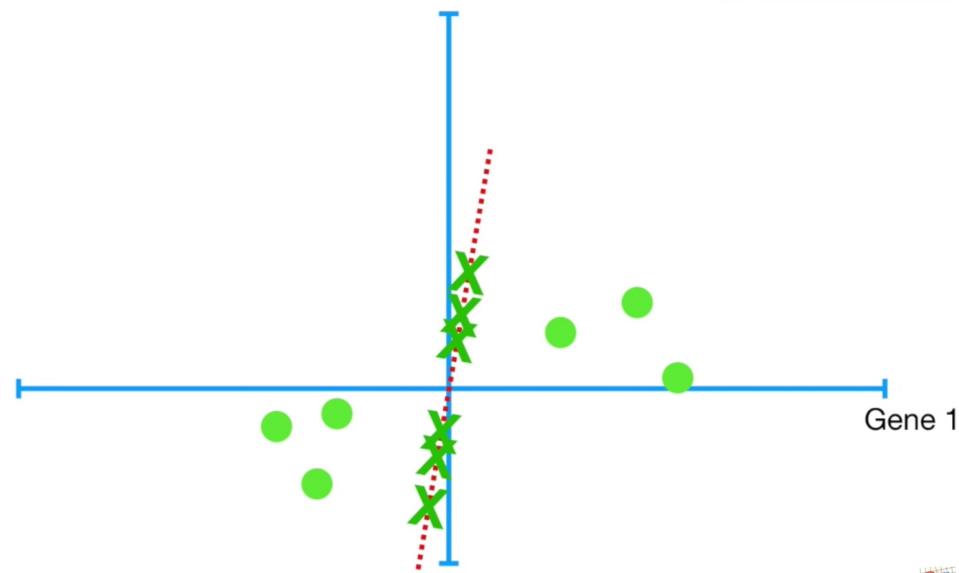


COMPONENTES PRINCIPALES

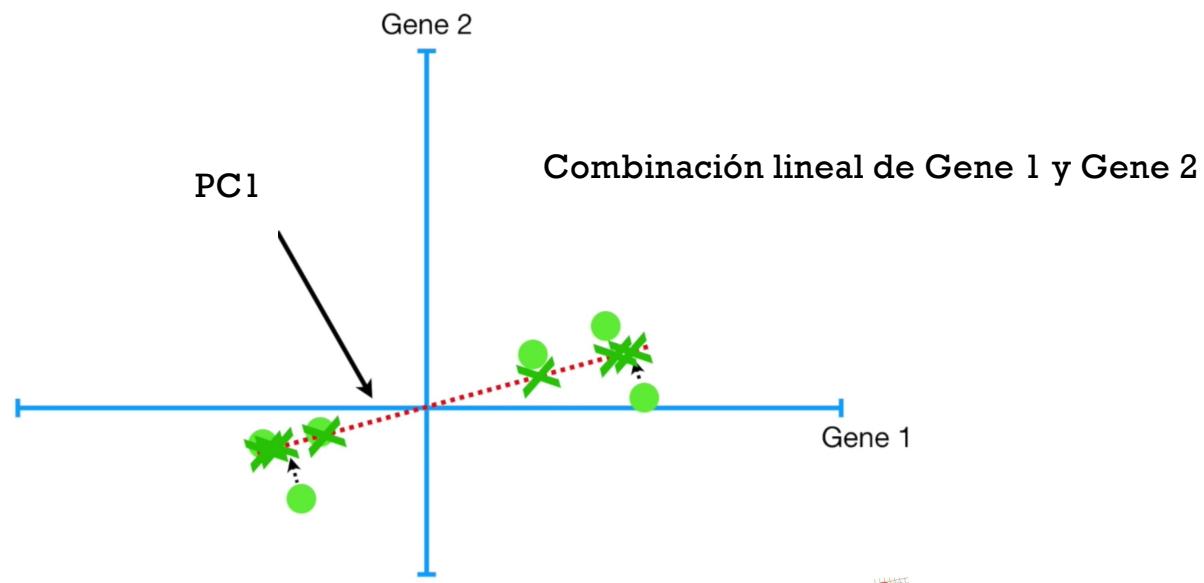


COMPONENTES PRINCIPALES

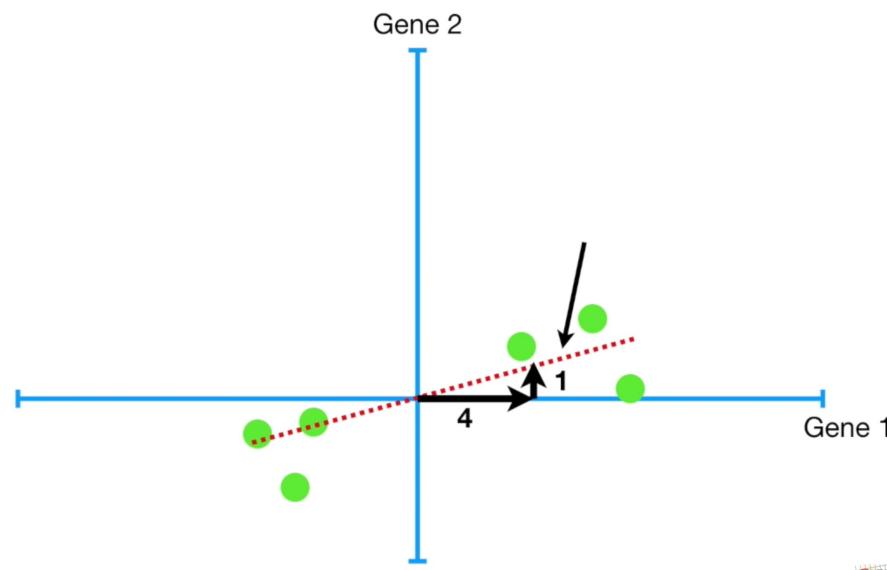
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS(distances)}$$



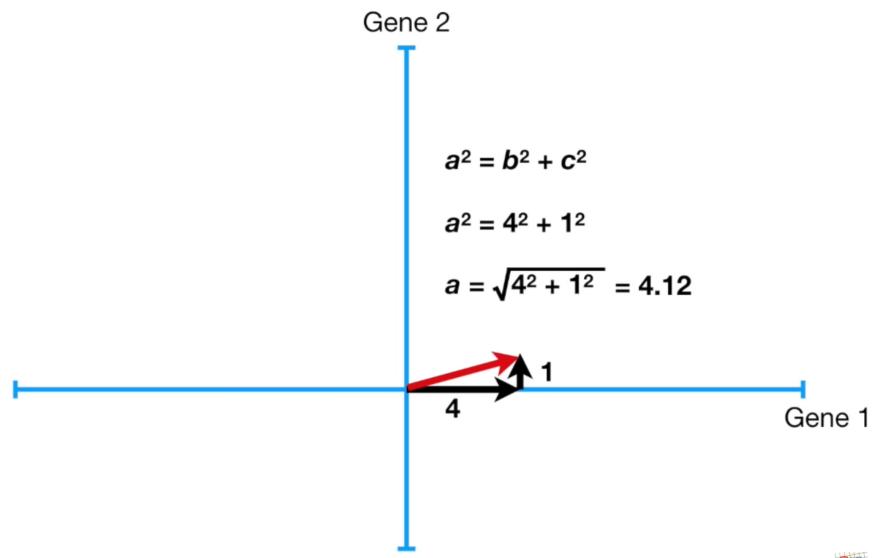
COMPONENTES PRINCIPALES



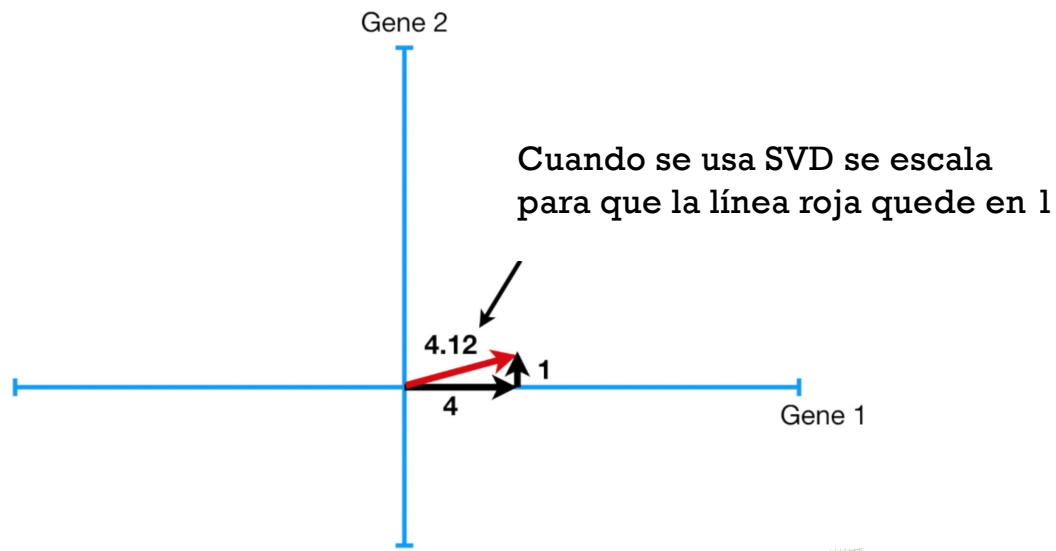
COMPONENTES PRINCIPALES



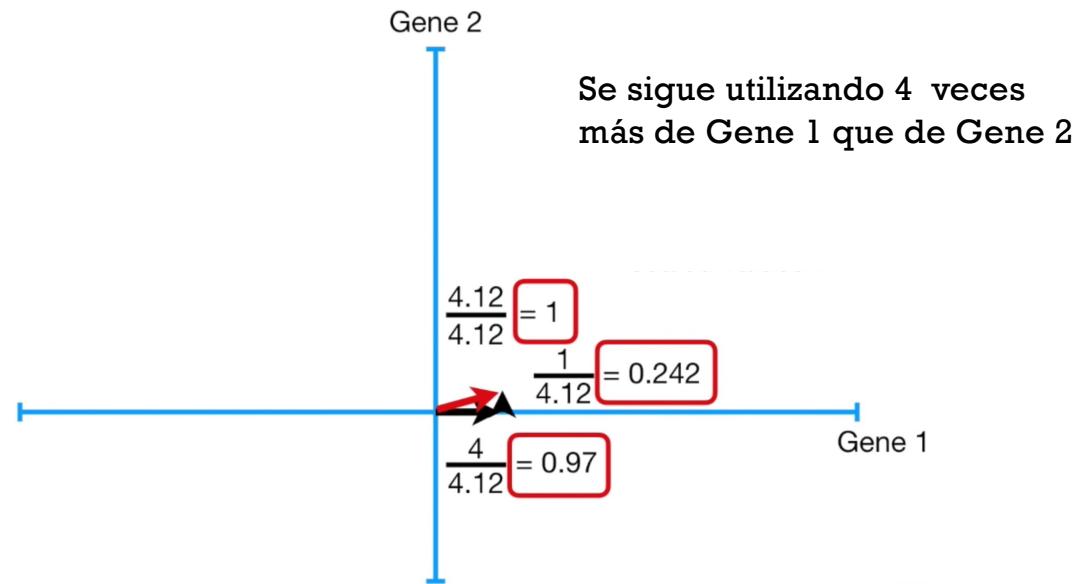
COMPONENTES PRINCIPALES



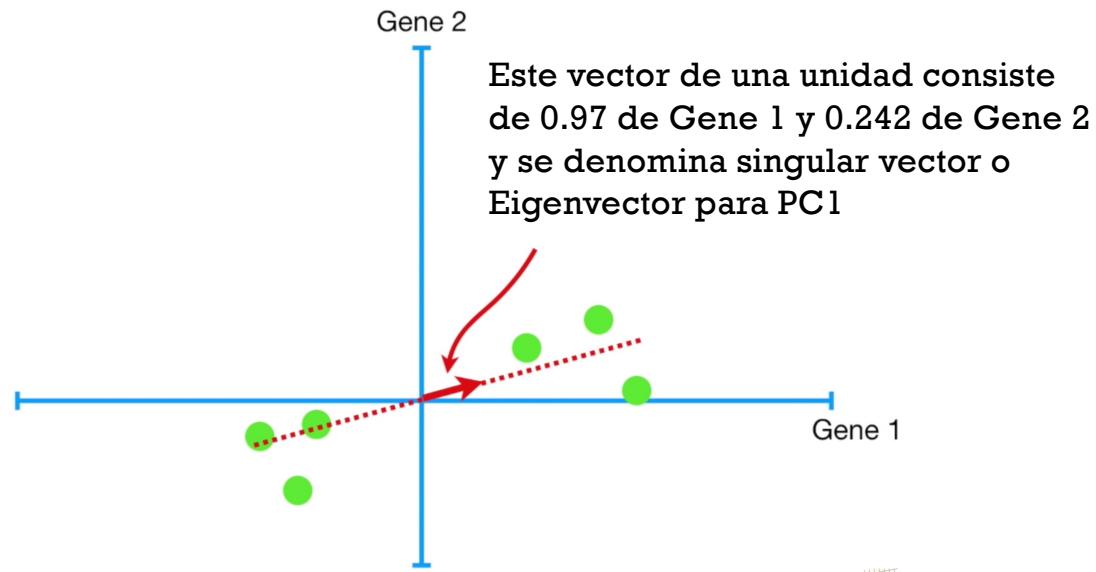
COMPONENTES PRINCIPALES



COMPONENTES PRINCIPALES

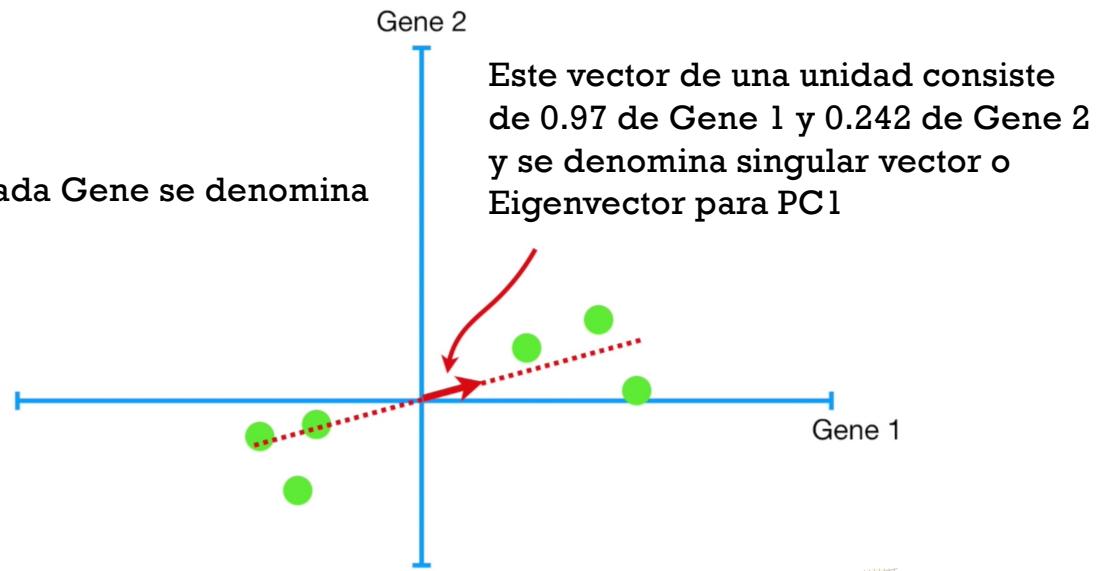


COMPONENTES PRINCIPALES



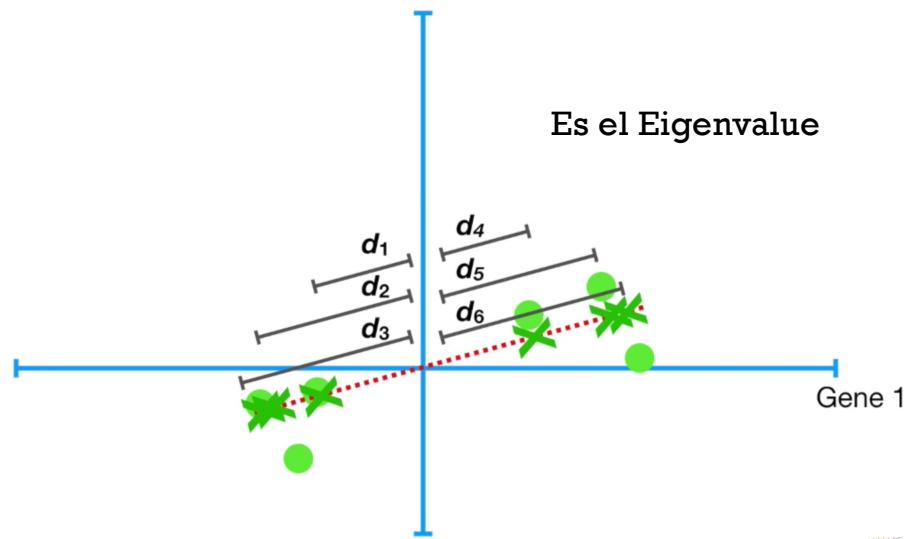
COMPONENTES PRINCIPALES

La proporción de cada Gene se denomina loading o carga



COMPONENTES PRINCIPALES

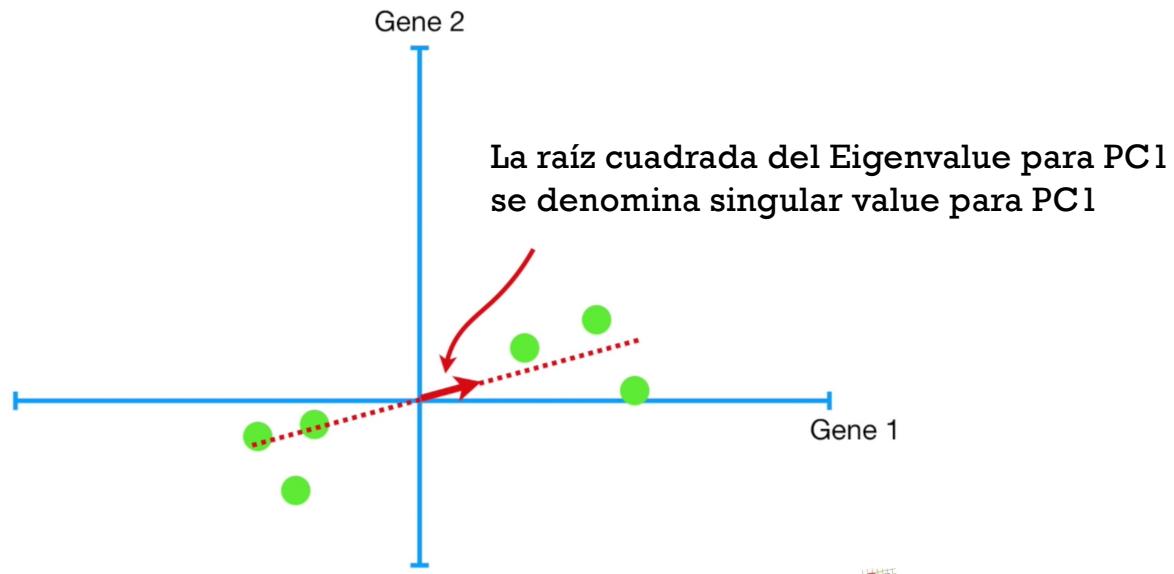
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS(distances)}$$



<https://www.youtube.com/watch?v=FgakZw6K1QQ>

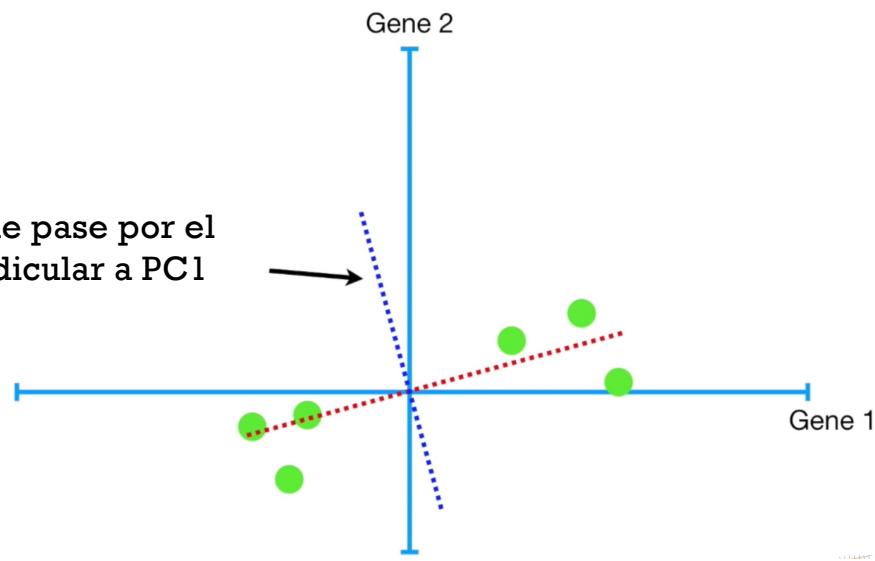


COMPONENTES PRINCIPALES



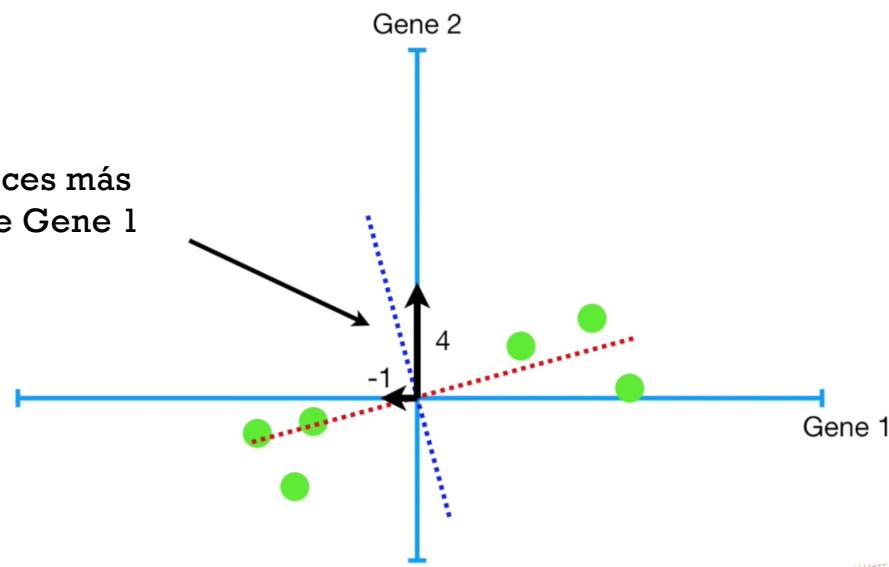
COMPONENTES PRINCIPALES

Se escoge la línea que pase por el origen y sea perpendicular a PC1



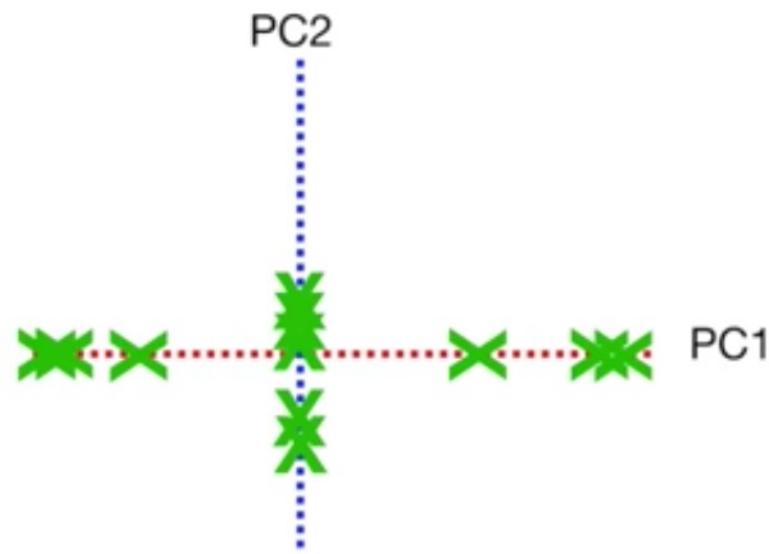
COMPONENTES PRINCIPALES

Gene 2 es 4 veces más importante que Gene 1



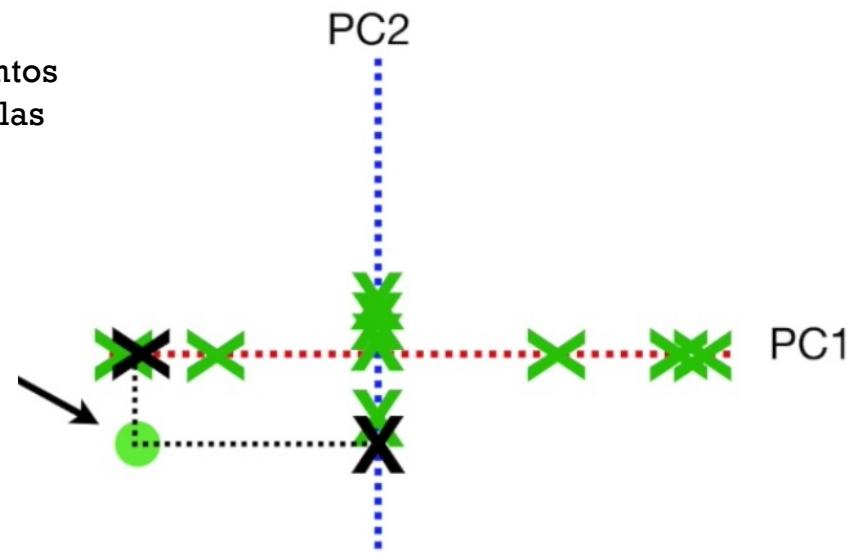
COMPONENTES PRINCIPALES

Rotamos y utilizamos los puntos
Proyectados para encontrar las
muestras



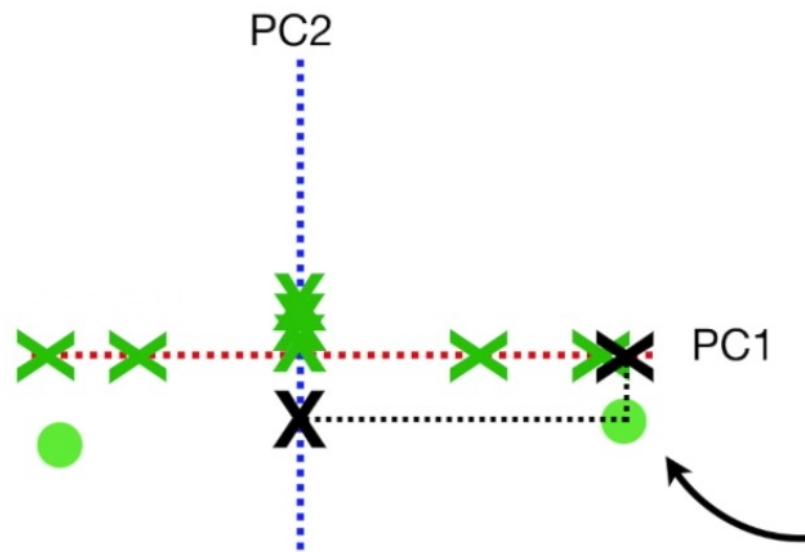
COMPONENTES PRINCIPALES

Rotamos y utilizamos los puntos
Proyectados para encontrar las
muestras



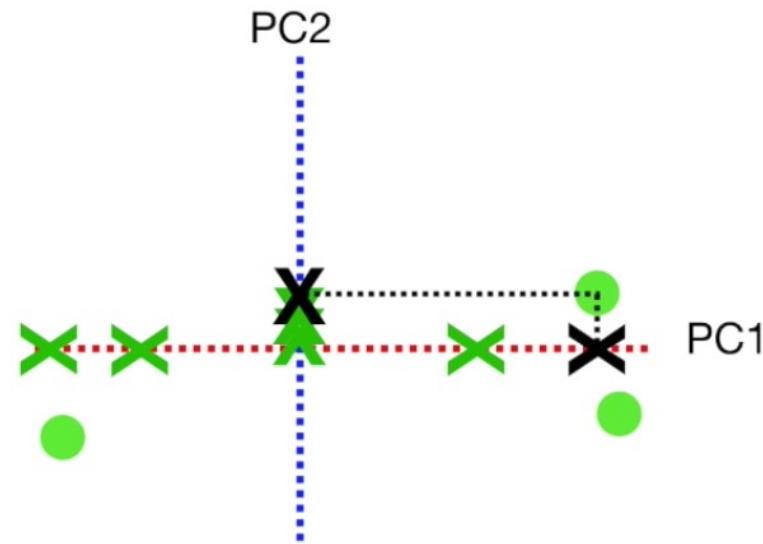
COMPONENTES PRINCIPALES

Rotamos y utilizamos los puntos
Proyectados para encontrar las
muestras



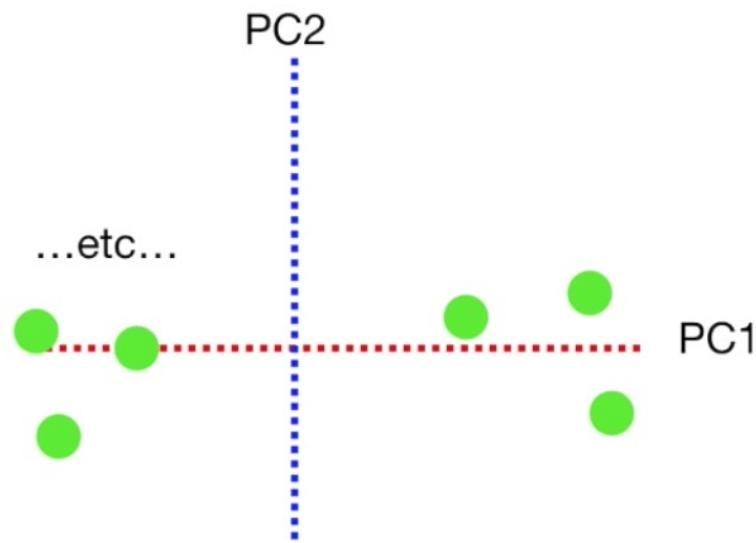
COMPONENTES PRINCIPALES

Rotamos y utilizamos los puntos
Proyectados para encontrar las
muestras



COMPONENTES PRINCIPALES

Rotamos y utilizamos los puntos
Proyectados para encontrar las
muestras



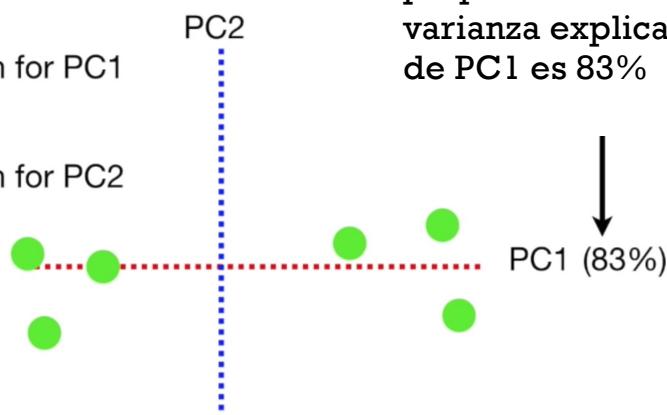
COMPONENTES PRINCIPALES

Si la variación de PC1
fuese 15 y la de PC2 3

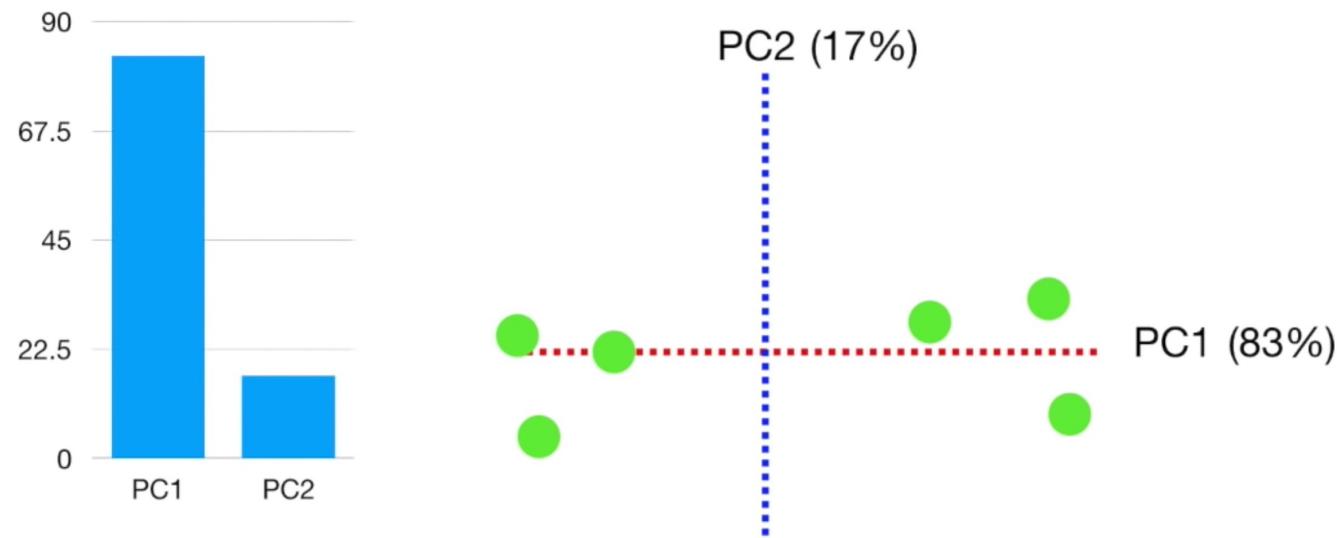
$$\frac{\text{SS}(\text{distances for PC1})}{n - 1} = \text{Variation for PC1}$$

$$\frac{\text{SS}(\text{distances for PC2})}{n - 1} = \text{Variation for PC2}$$

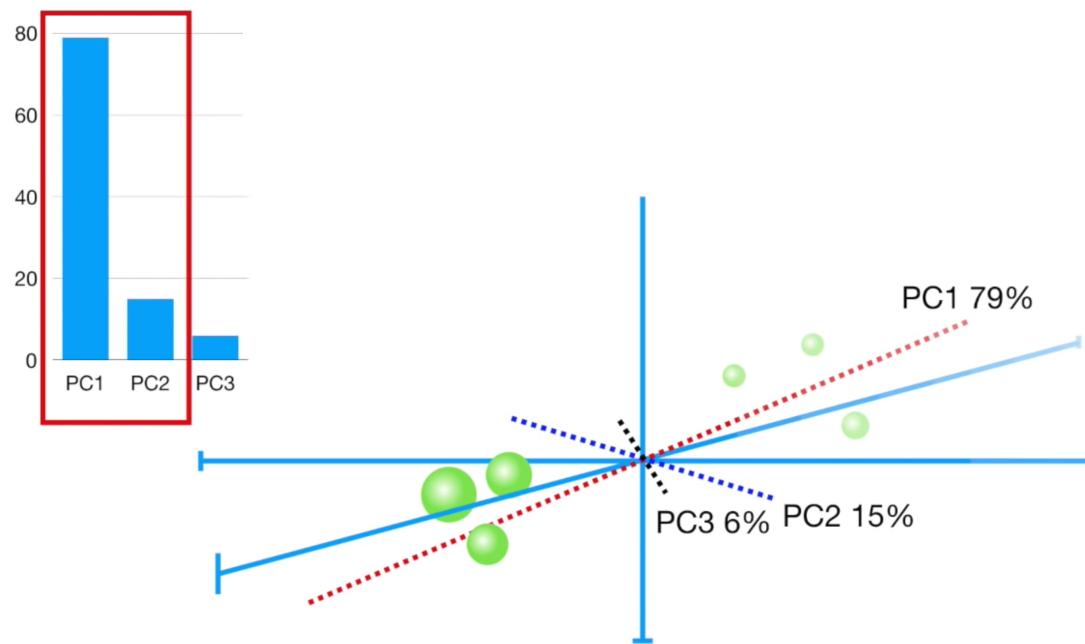
La variación total en
ambos PCs es 18 y la
proporción de
varianza explicada
de PC1 es 83%



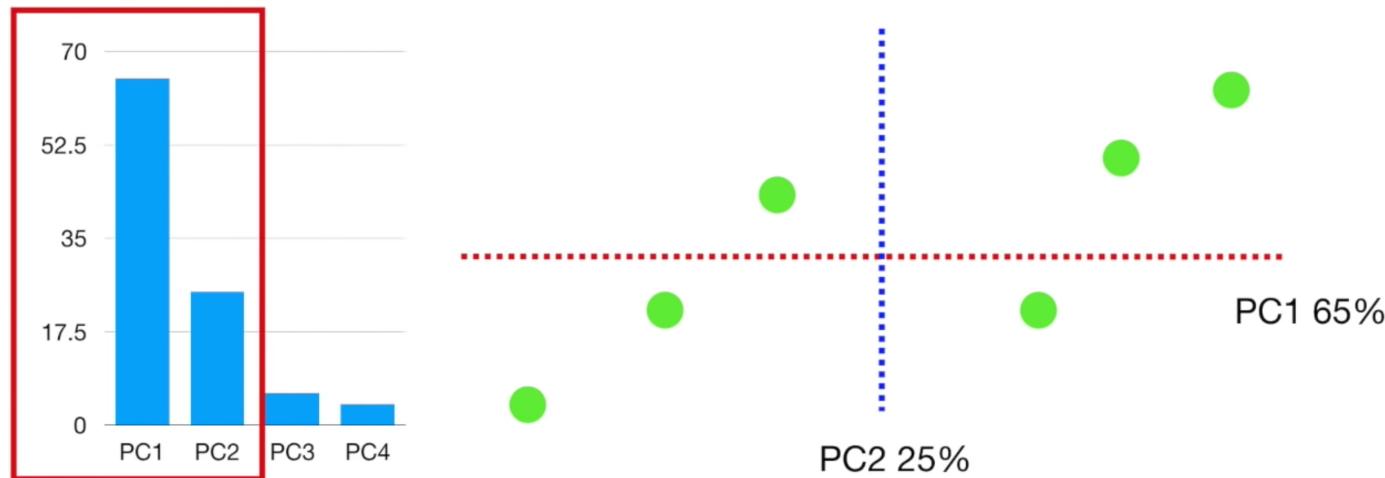
COMPONENTES PRINCIPALES



COMPONENTES PRINCIPALES



COMPONENTES PRINCIPALES



COMPONENTES PRINCIPALES

Consideraciones

- La varianza de cada uno de los atributos (dimensiones) originales depende de su escala, por lo que se debe **normalizar** los datos originales
- El número de dimensiones originales puede ser superior al número de instancias del dataset, pero limitaría el número de PCs al número de instancias -1
- Puede que la varianza esté bien distribuida en los atributos originales, por lo que aplicar PCA no tendría efecto
- El considerar solo los componentes principales más importantes permite ignorar el posible ruido que puedan tener los datos, enfocándose solo en la información más importante. A partir de una transformación inversa del espacio de los PCs hacia el espacio original podemos entonces **filtrar el ruido**.

