EJEMPLO USO DE PICKLE

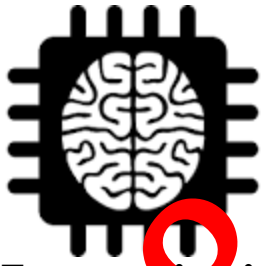
- Vamos a implementar un modelo de árbol de decisión con sklearn para el conjunto de datos de cáncer de seno.
- Guardaremos el objeto del modelo ya implementado y entrenado en un archivo mediante el modulo pickle de python que nos permitirá serializarlo.
- Cargaremos el archivo y de serializaremos el archivo para obtener de nuevo nuestro objeto del modelo entrenado.

TALLER: K-MEANS – CLIENTES SUPERMERCADOS

Desarrollar el taller de clustering de clientes de supermercados 09-SUPERMERCADOS-K-Means-STUD.html (hasta antes de la determinación del k).

UNIVERSIDAD ICESI

AGENDA

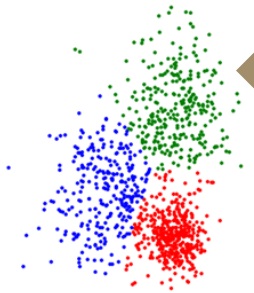


Aprendizaje automático

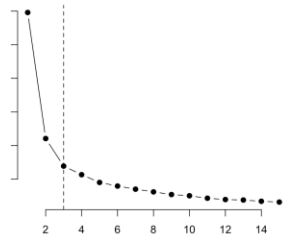
REPASO

Aprendizaje supervisado

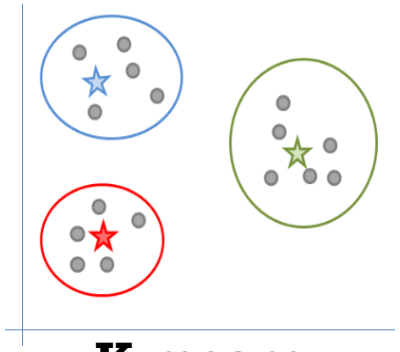
Aprendizaje no supervisado



Clustering



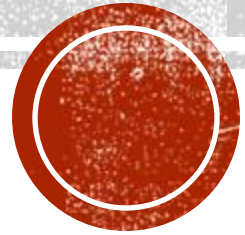
Evaluación del clustering



K-means

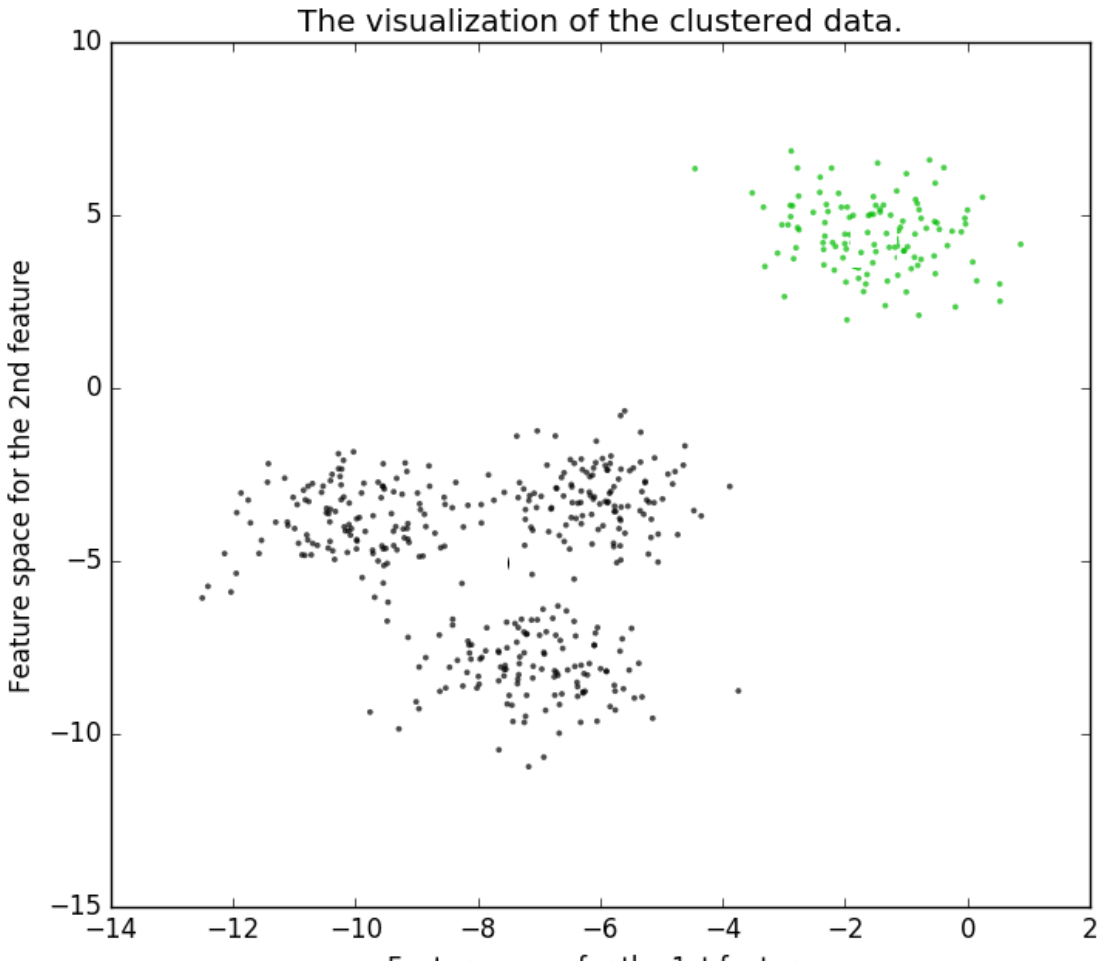
UNIVERSIDAD ICESI

EVALUACIÓN DE CLUSTERING



UNIVERSIDAD ICESI

ESCOGENCIA DEL K



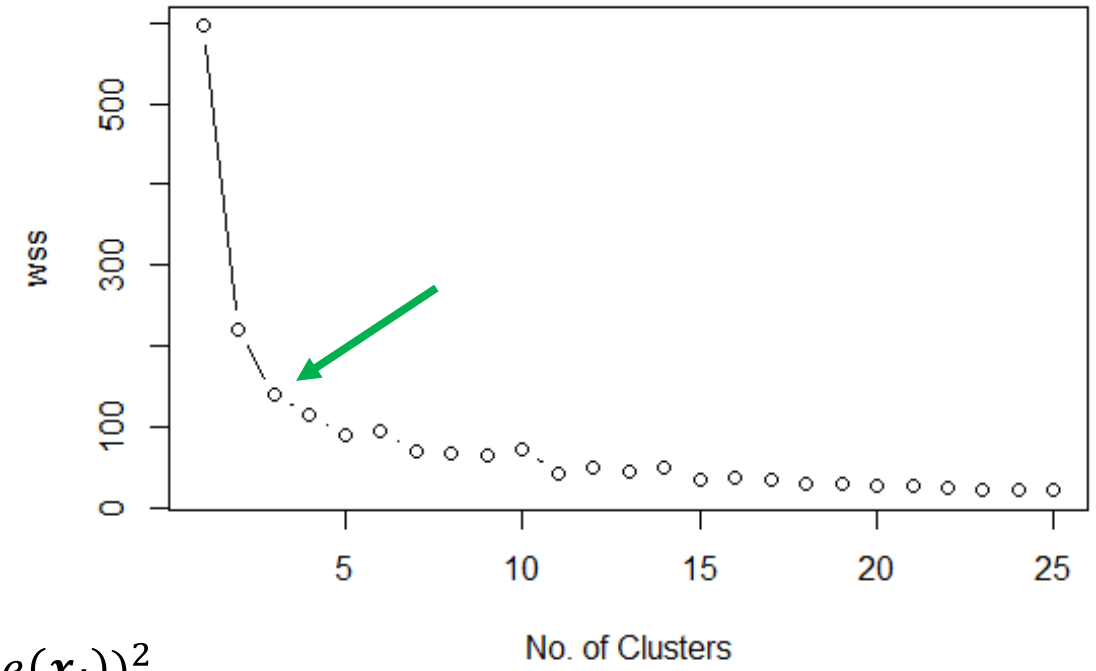
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

UNIVERSIDAD ICESI

¿Cuántos clusters ven uds aquí?

ESCOGENCIA DEL K – CODO

- Heurísticos:
  - Dependen del juicio del analista, se requiere conocimiento del negocio
- Método "del codo":
  - Dibujar WSS para cada valor de K
  - Escoger el valor de K que implica una reducción "considerable" del WSS del clustering resultante, cuando la curva se vuelve asintótica

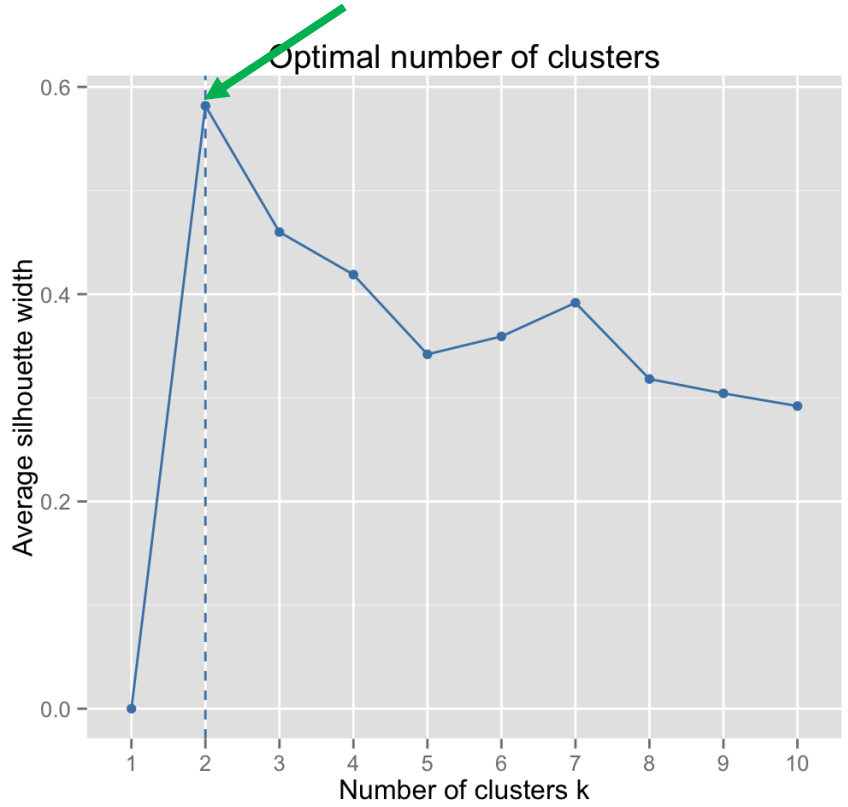


$$WSS = \sum_{i=1}^{\#instancias} distancia(\boldsymbol{x_i} - centroide(\boldsymbol{x_i}))^2$$

UNIVERSIDAD ICESI

ESCOGENCIA DEL K – SILHOUETTE

- Método Silhouette
  - Busca el K que maximice la **separación** entre clusters, con clusters lo más **compactos** posibles
  - Analiza el ajuste de cada instancia al cluster al que fue asignado
  - Qué tan cerca está cada observación de las demás de su propio cluster?
    - 0,7-1,0: el cluster es fuertemente robusto
    - 0,5-0,7: el cluster es razonablemente robusto
    - 0,25-0,5: el cluster puede ser artificial y puede no denotar una noción de estructura necesariamente
    - Inferior a 0,25: el cluster debería descartarse, no indica estructura
  - Se busca la maximización del valor Silhouette: promedio de los clusters



UNIVERSIDAD ICESI

ESCOGENCIA DEL K – SILUETA

- Método Silueta (Silhouette)
  - Calcular el valor de silueta de cada punto:
    - Cohesión del punto con su cluster C_i (promedio de distancias con puntos de su mismo cluster):

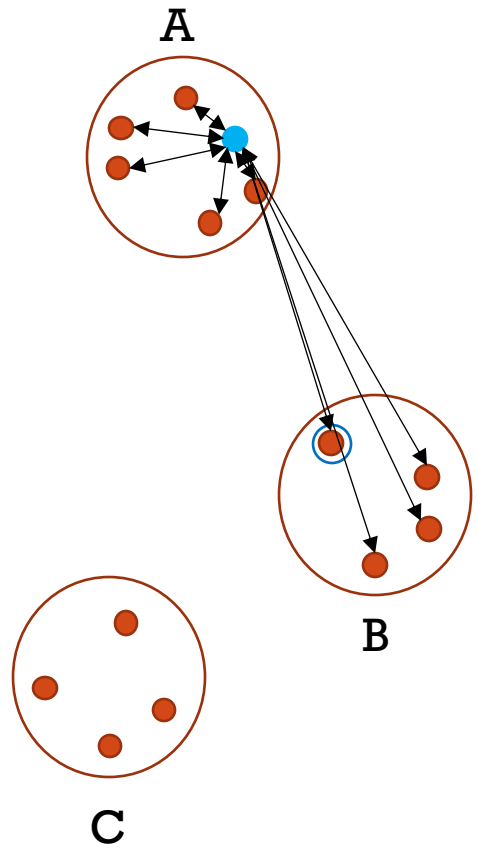
$$cohesión(p) = a(p) = \frac{\sum_{p' \in C_i,\ p' \neq p} distancia(p, p')}{|C_i| - 1}$$

    - Separación de los puntos de otros clusters (distancia promedio con los puntos del cluster más cercano):

$$separación(p) = b(p) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left(\frac{\sum_{p' \in C_j} distancia(p, p')}{|C_j|} \right)$$

    - El valor de silueta del punto es entonces:

$$silueta(p) = s(p) = \frac{b(p) - a(p)}{\max\left(b(p), a(p)\right)}$$



ESCOGENCIA DEL K – SILUETA

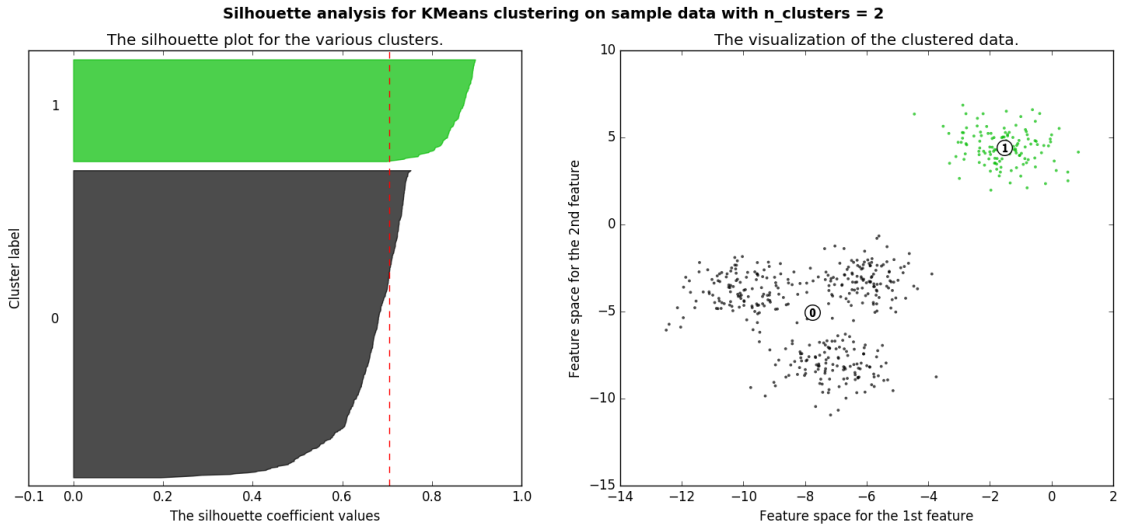
- Método Silueta (Silhouette)
  - Calcular el valor de silueta de cada cluster (promedio de las siluetas de sus puntos).

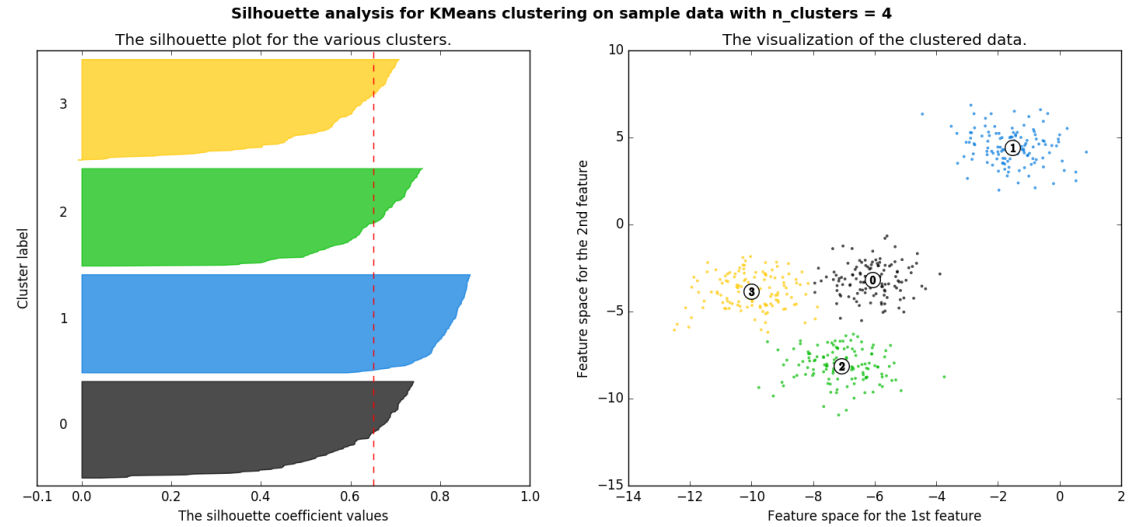
$$silueta(C_i) = \frac{1}{|C_i|} \sum_{p \in C_i} s(p)$$

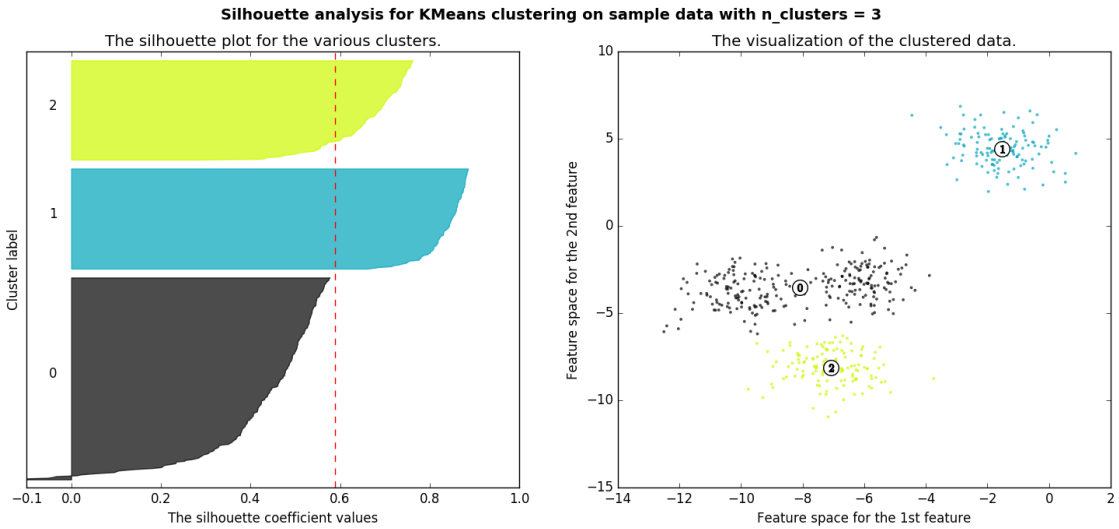
  - Analizar los puntos y clusters, buscando posibles problemas de asignación dados por el valor del K:
    - El rango de la silueta está entre -1 y 1
    - Una silueta de 0 implica que la asignación de un punto a su cluster es indiferente
    - Se espera que los puntos del mismo cluster estén más cercanos al punto en cuestión: para que la silueta sea positiva tenemos que $a(p) < b(p)$

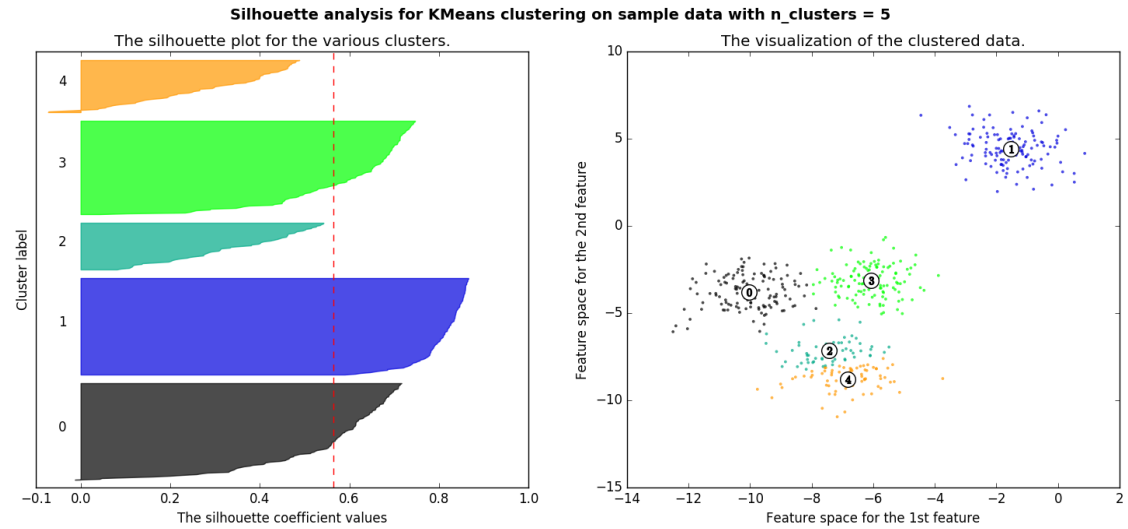
ROUSSEEUW (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.

ESCOGENCIA DEL K – SILHOUETTE





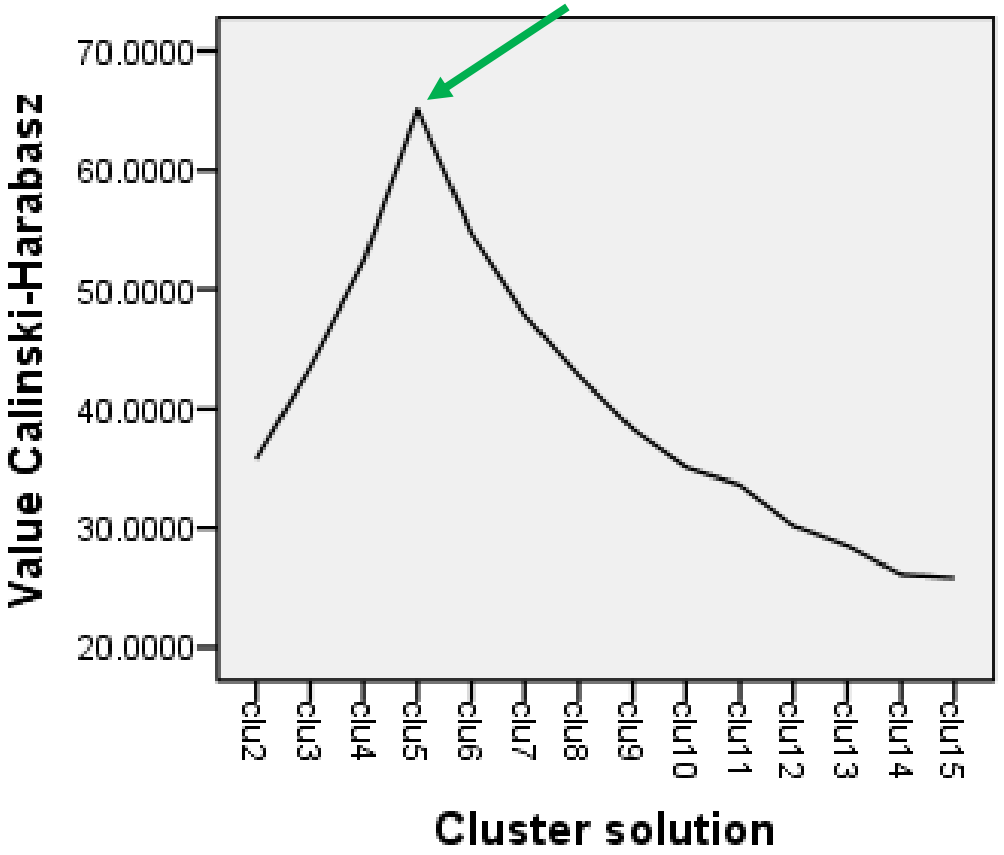




ESCOGENCIA DEL K – CALINSKI-HARABASZ

- Método de Calinski-Harabasz:
  - Se busca el K que maximice la **separación** entre clusters, con clusters lo más **compactos** posibles
  - TSS = variación total (entre todos los datos y el centro global)
  - WSS = variación intra-cluster (entre los puntos de cada cluster y sus centroides
  - BSS = variación inter-cluster (entre los centroides de los clusters y el centro global). BSS = TSS - WSS
  - CH = ratio entre la variación entre clusters (BSS) y el promedio de la variación interna de los clusters (WSS). Se busca maximizar CH:

$$CH = \frac{BSS}{WSS} * \frac{N-k}{k-1}$$

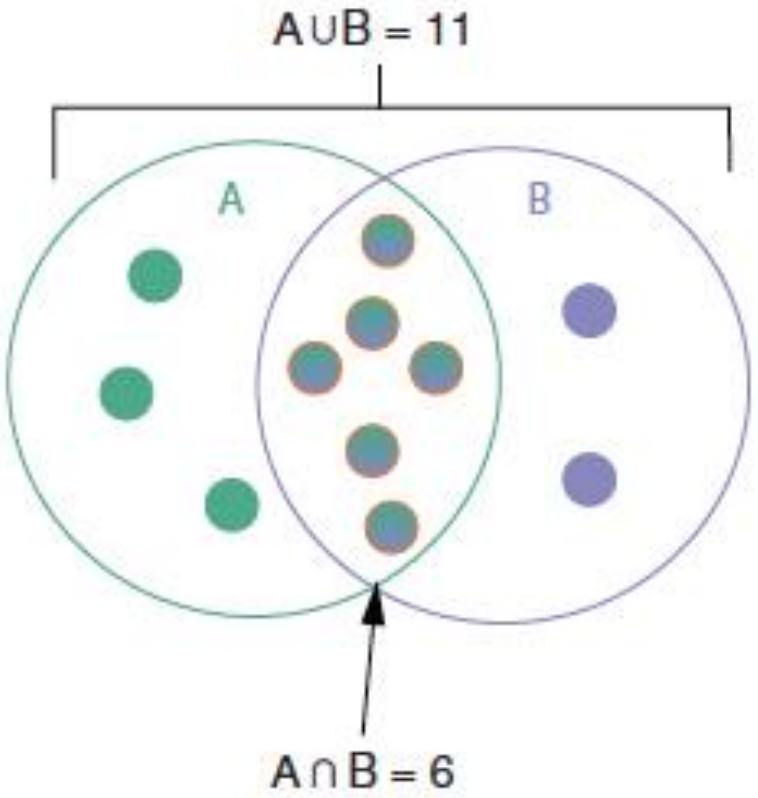


UNIVERSIDAD ICESI

BOOTSTRAP DE LOS CLUSTERS CREADOS

- La robustez del clustering depende de la escogencia adecuada del **k**
  - Algunos clusters encontrados representan una estructura bien determinada
  - Otros terminan siendo repositorios de las instancias que no encajan en ninguno de los clusters estructurales
- Clusters estructurales: estables y resistentes a cambios de los datos
  - Repetición del clustering a partir de técnicas de resampleo (e.g. bootstrapping)
  - Algoritmo de clustering bootstrap
    1. Crear un clustering del dataset original
    2. Crear una nuevo dataset del mismo tamaño a partir del original, con repeticiones, y realizar su clustering
    3. Para cada cluster original encontrar el cluster resampleado más similar, utilizando la medida de Jaccard: $Jaccard(A,B) = \frac{A \cap B}{A \cup B}$. Si esta medida es inferior a 0,5, el cluster correspondiente se considera inestable y se disuelve. Esto es una indicación de que no es buen cluster.
    4. Se repiten los pasos 2 y 3 varias veces

MEDIDA DE JACCARD



Jaccard Similarity

$$\frac{(A \cap B)}{(A \cup B)} = 6/11 \backsimeq 0.55$$

Zumel, 2020

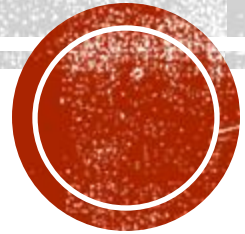
UNIVERSIDAD ICESI

TALLER: EVALUACIÓN DE CLUSTERING

Continuar con el taller de clustering de clientes de supermercado 09-SUPERMERCADOS- K-Means-STUD.html con la parte dedicada a la determinación y evaluación del número de clusters.

UNIVERSIDAD ICESI

APRENDIZAJE NO SUPERVISADO



UNIVERSIDAD ICESI

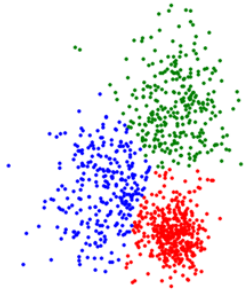
AGENDA

Aprendizaje automático

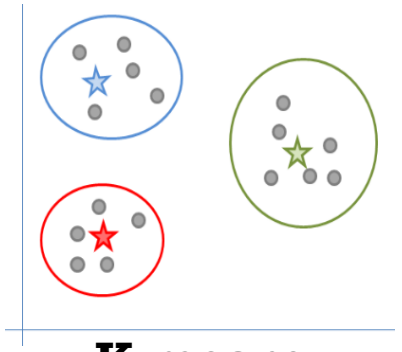
REPASO

Aprendizaje supervisado

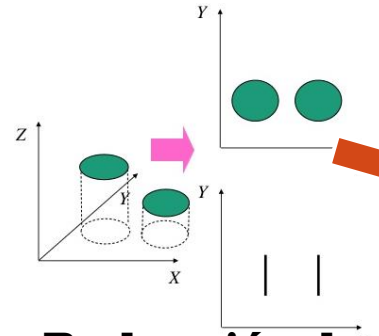
Aprendizaje no supervisado



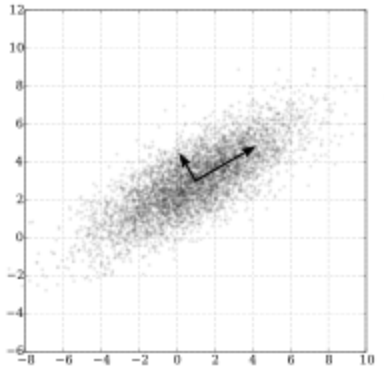
Clustering



K-means

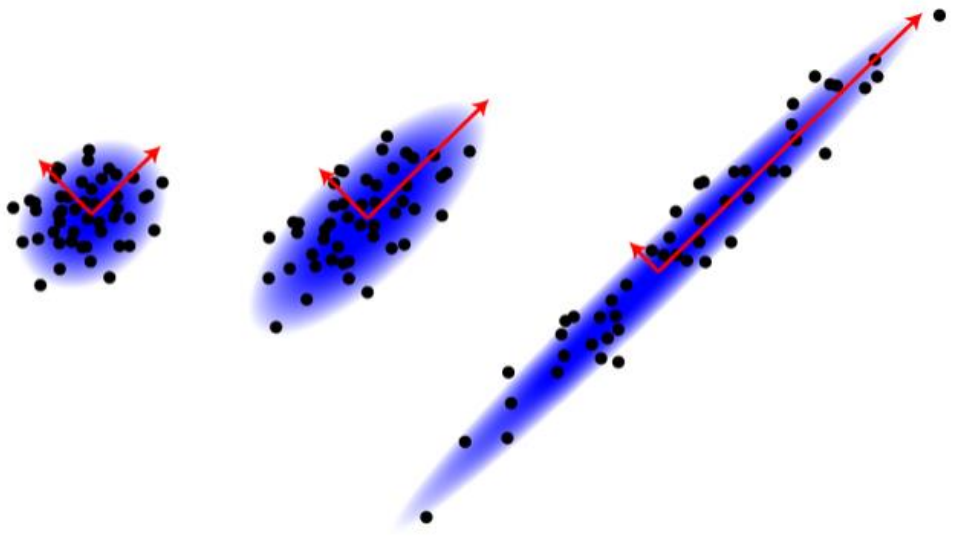


Reducción de dimensionalidad



Componentes principales

COMPONENTES PRINCIPALES



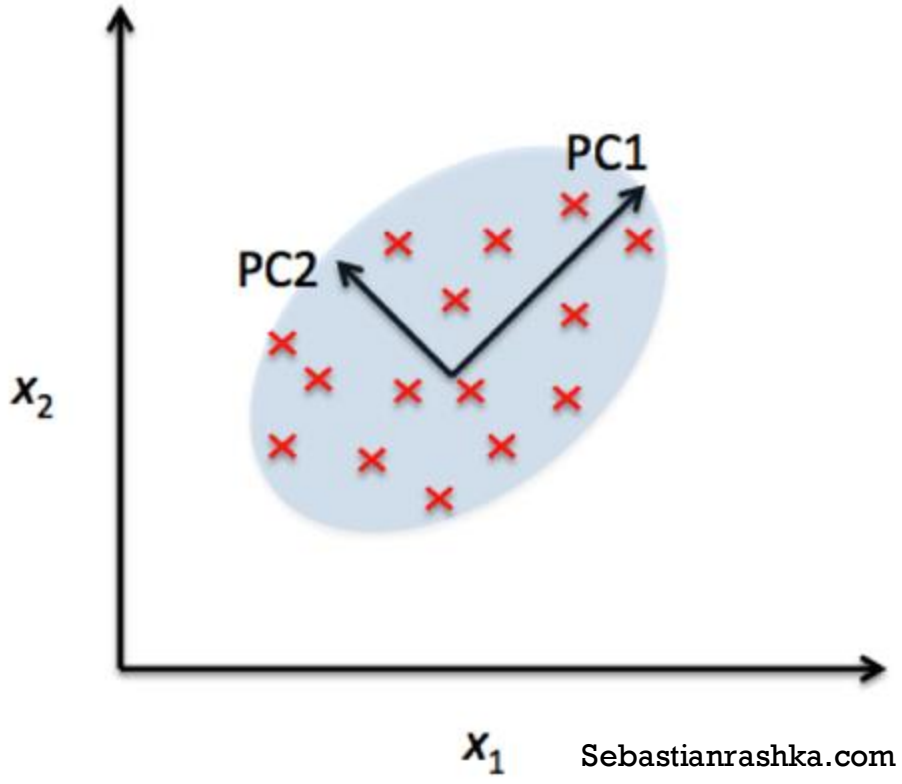
UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

PCA: Principal Component Analysis

Objetivo: Simplificar el dataset, encontrando una representación de **baja dimensionalidad** que conserve 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 los atributos altamente **correlacionados**
- PCA permite encontrar la superficie lineal de menos dimensiones más cercana a los puntos en el espacio original (en distancia Euclidiana)



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

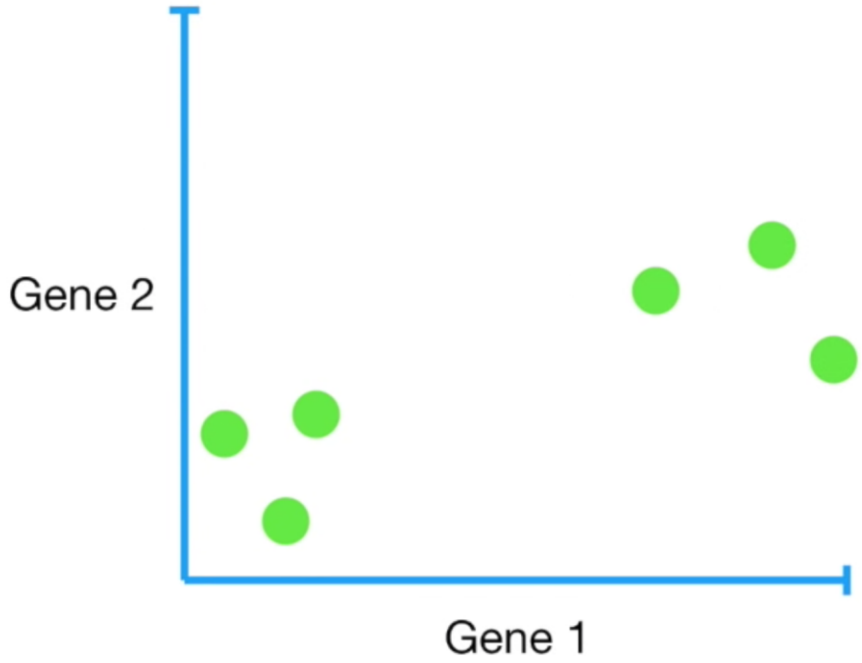
$$PC_i = \Phi_{1i}X_1 + \Phi_{2i}X_2 + \cdots + \Phi_{Ni}X_N, \text{ sujeto a } \sum_{j=1}^{N} {\Phi_{ji}}^2 = 1$$

$$\text{Además } \forall j < i, PC_i \perp PC_j$$

- 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

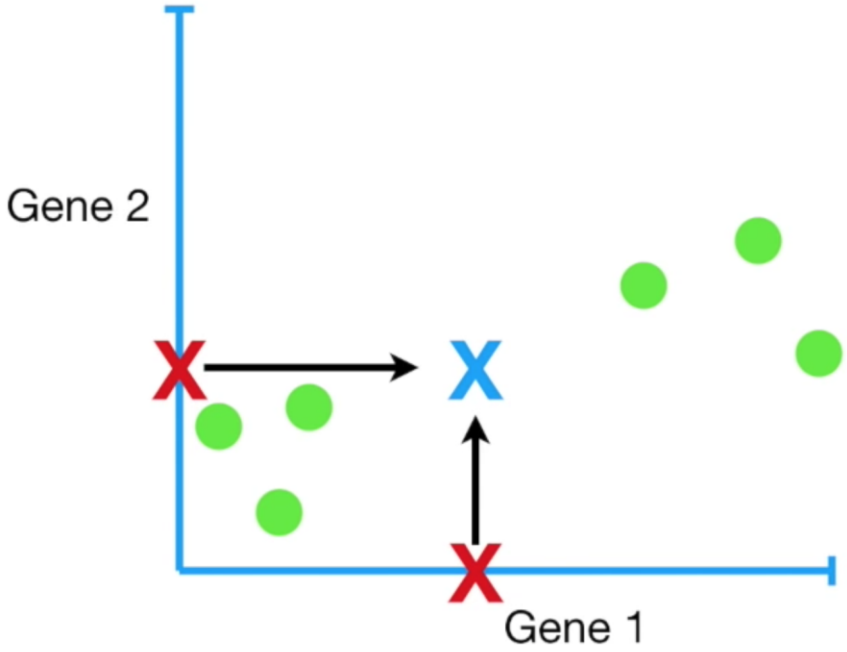
COMPONENTES PRINCIPALES

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

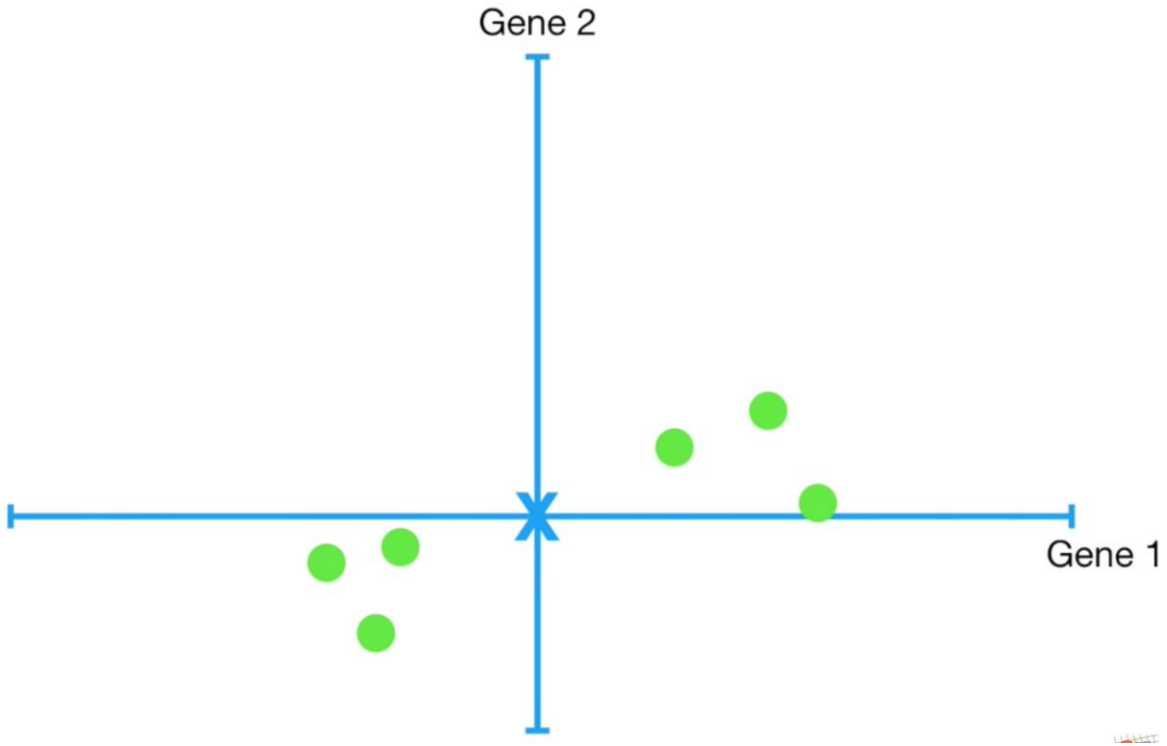


COMPONENTES PRINCIPALES

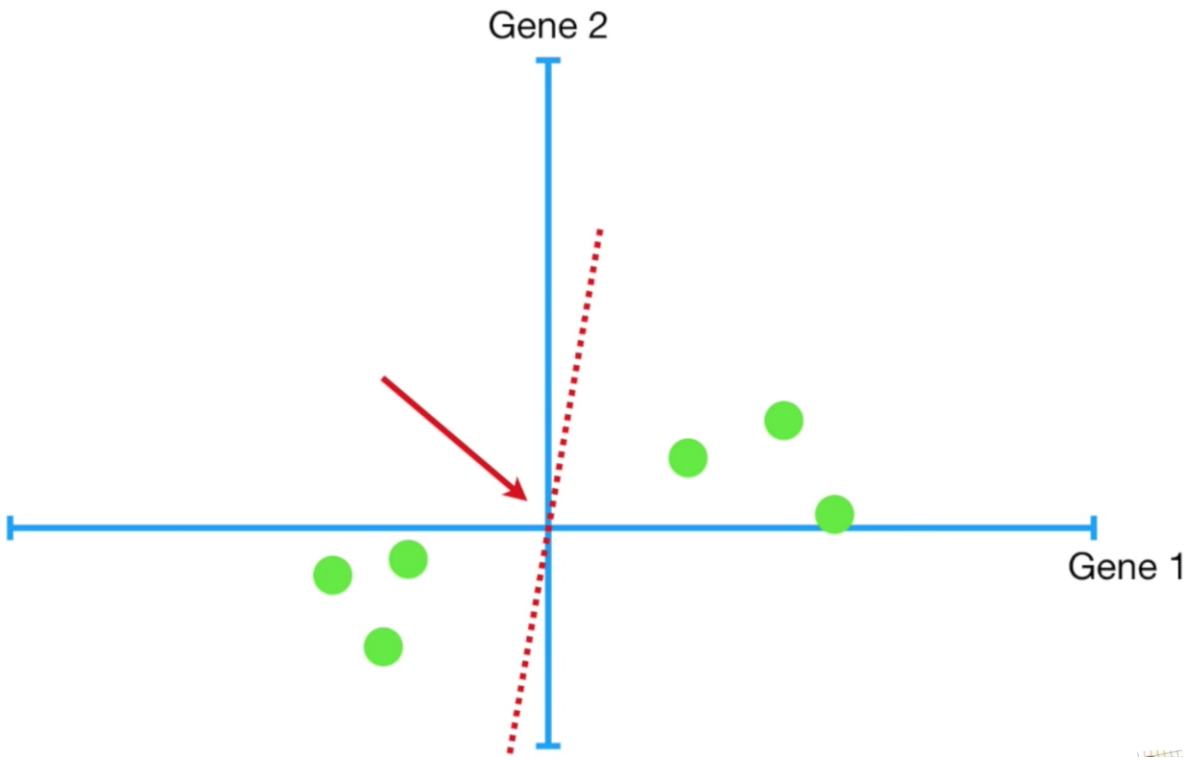
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



COMPONENTES PRINCIPALES

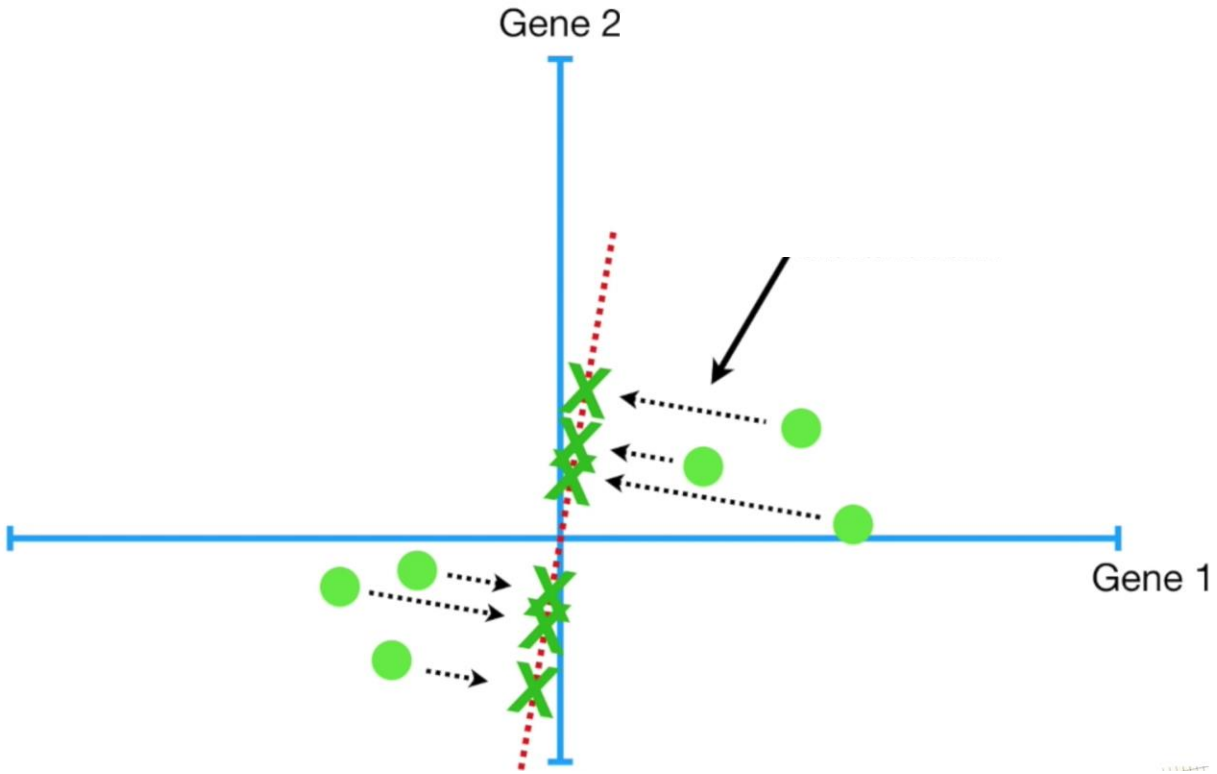


COMPONENTES PRINCIPALES



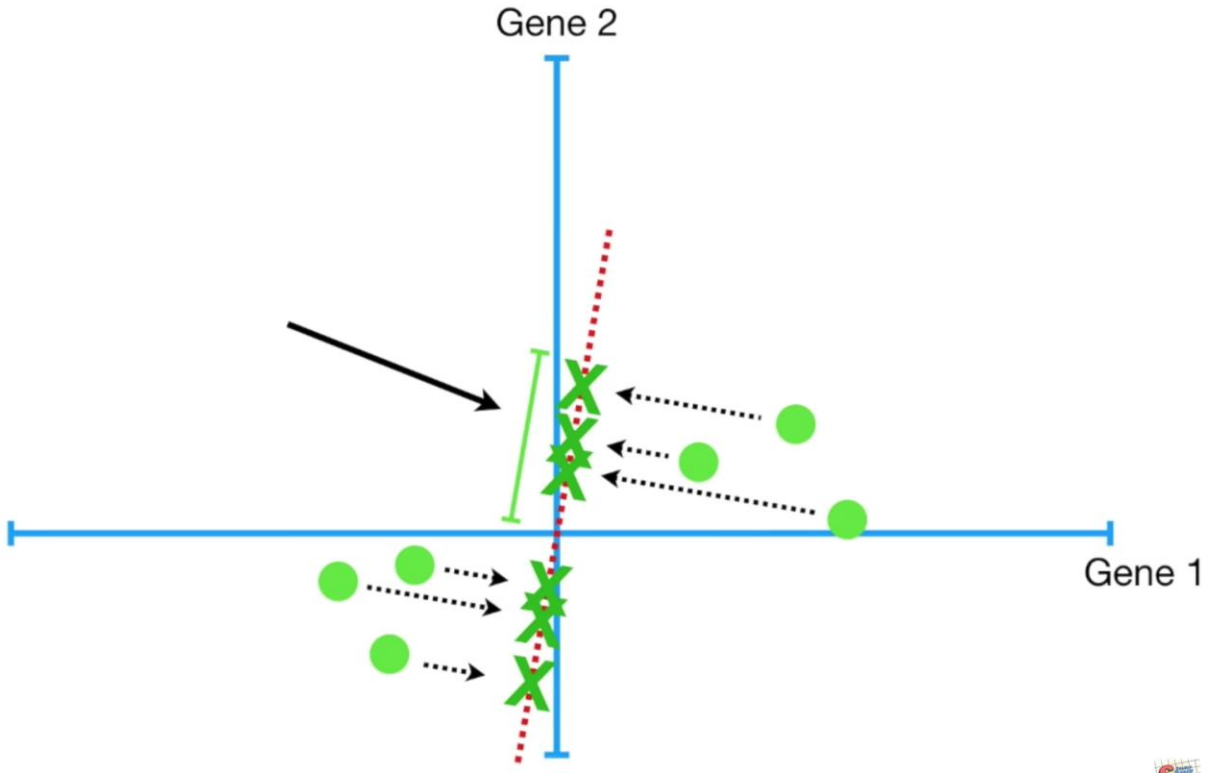
UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES



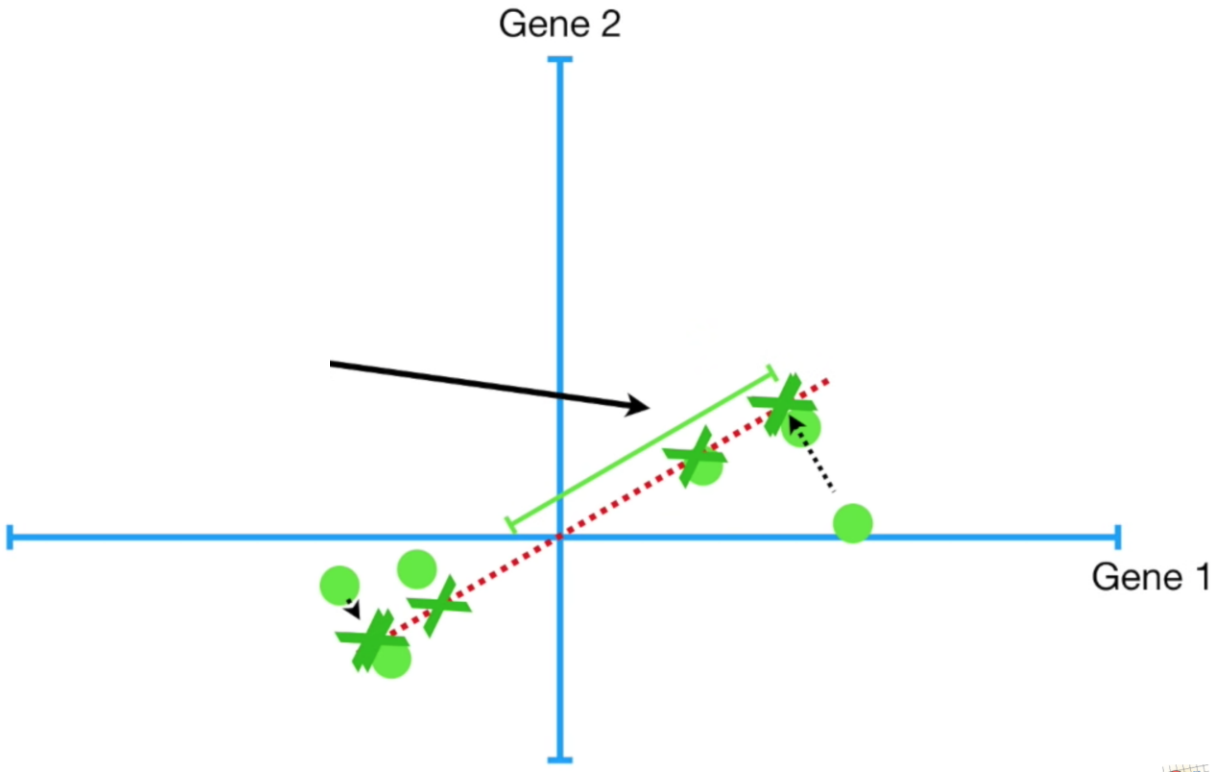
UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES



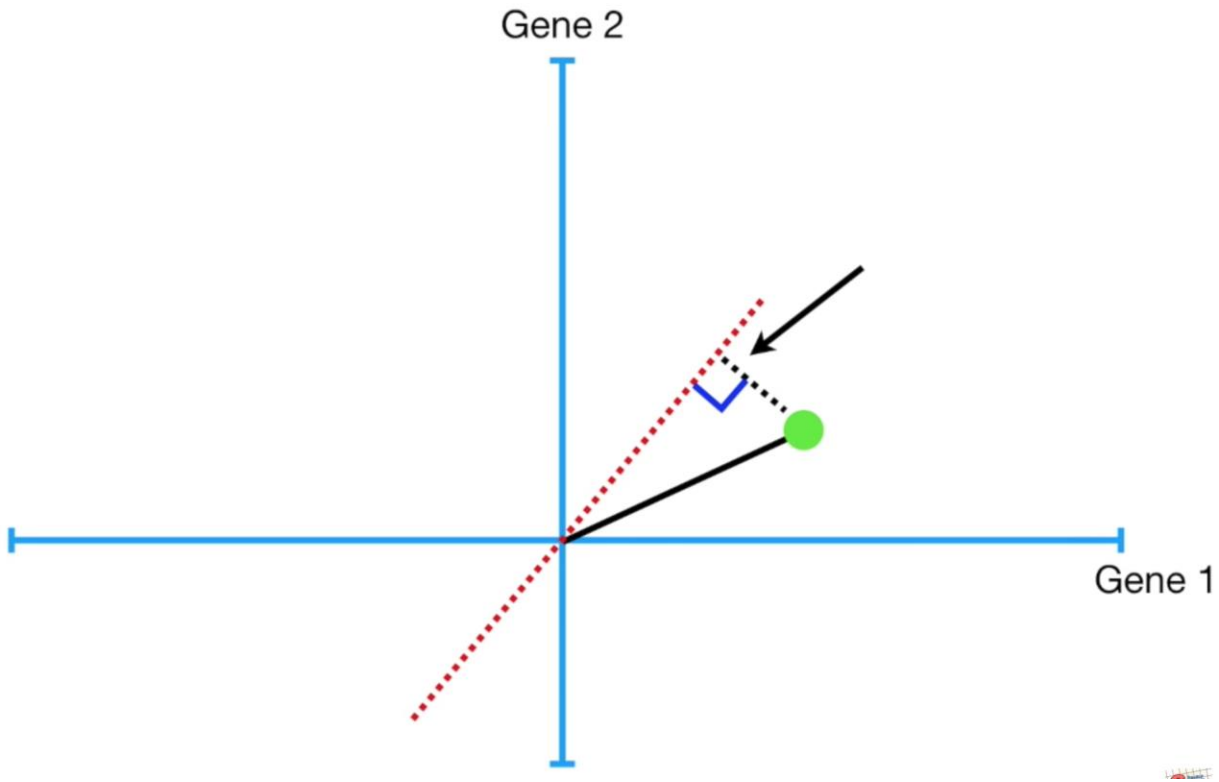
UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES



UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

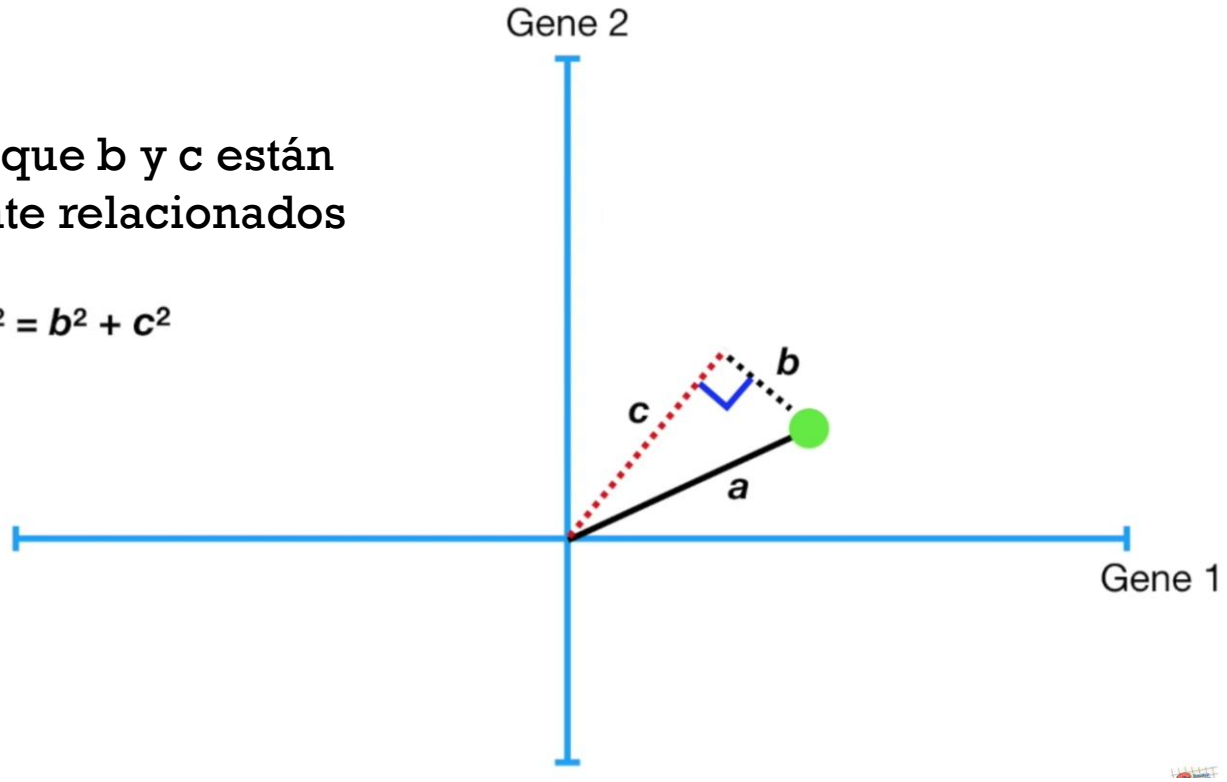


UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

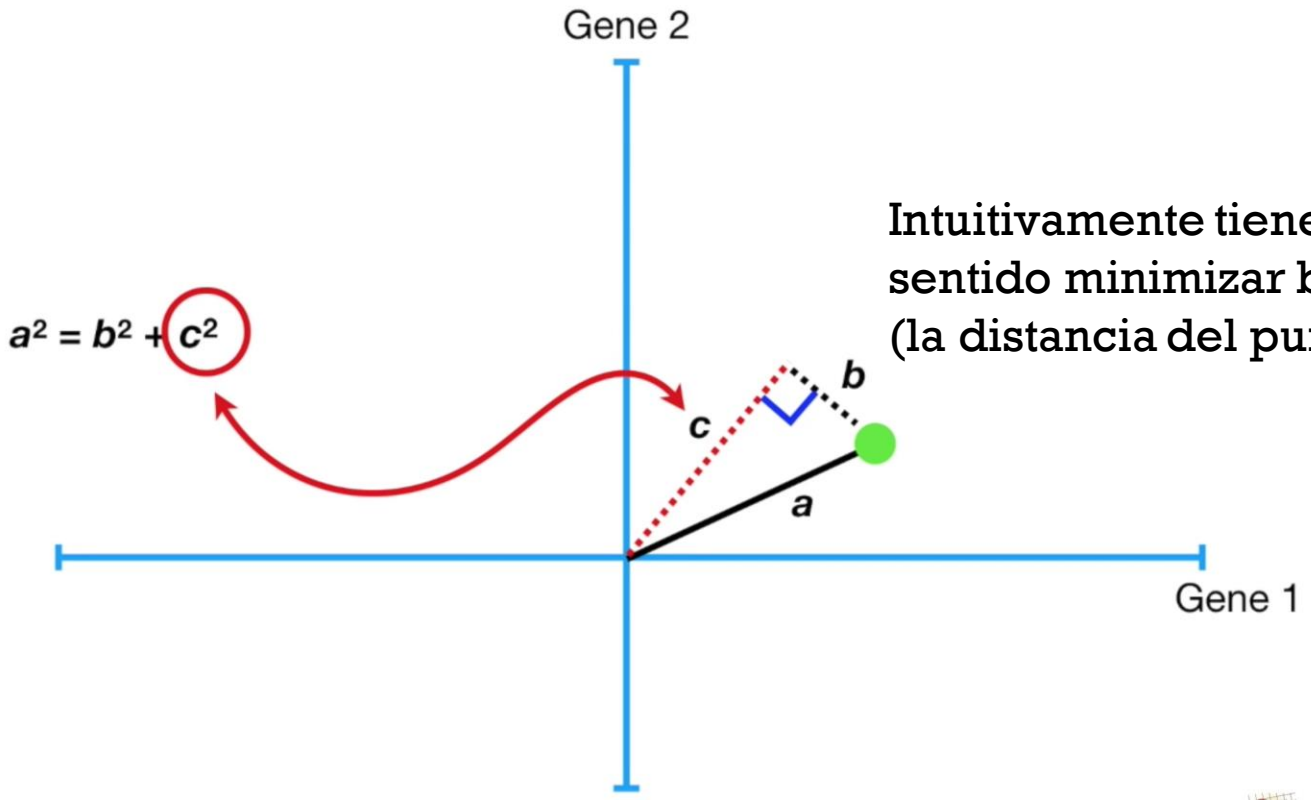
Se muestra que b y c están inversamente relacionados

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



UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

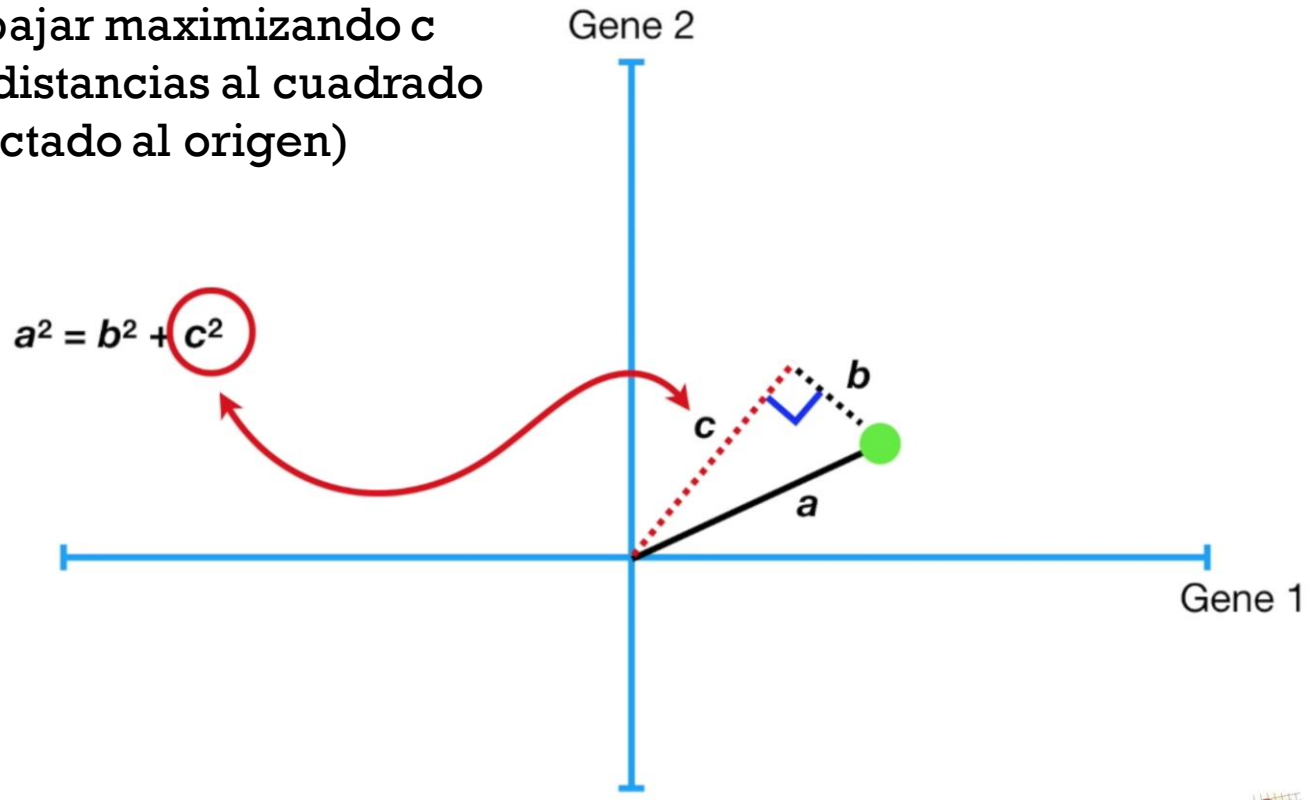


Intuitivamente tiene más sentido minimizar b
(la distancia del punto a la línea)

UNIVERSIDAD ICESI

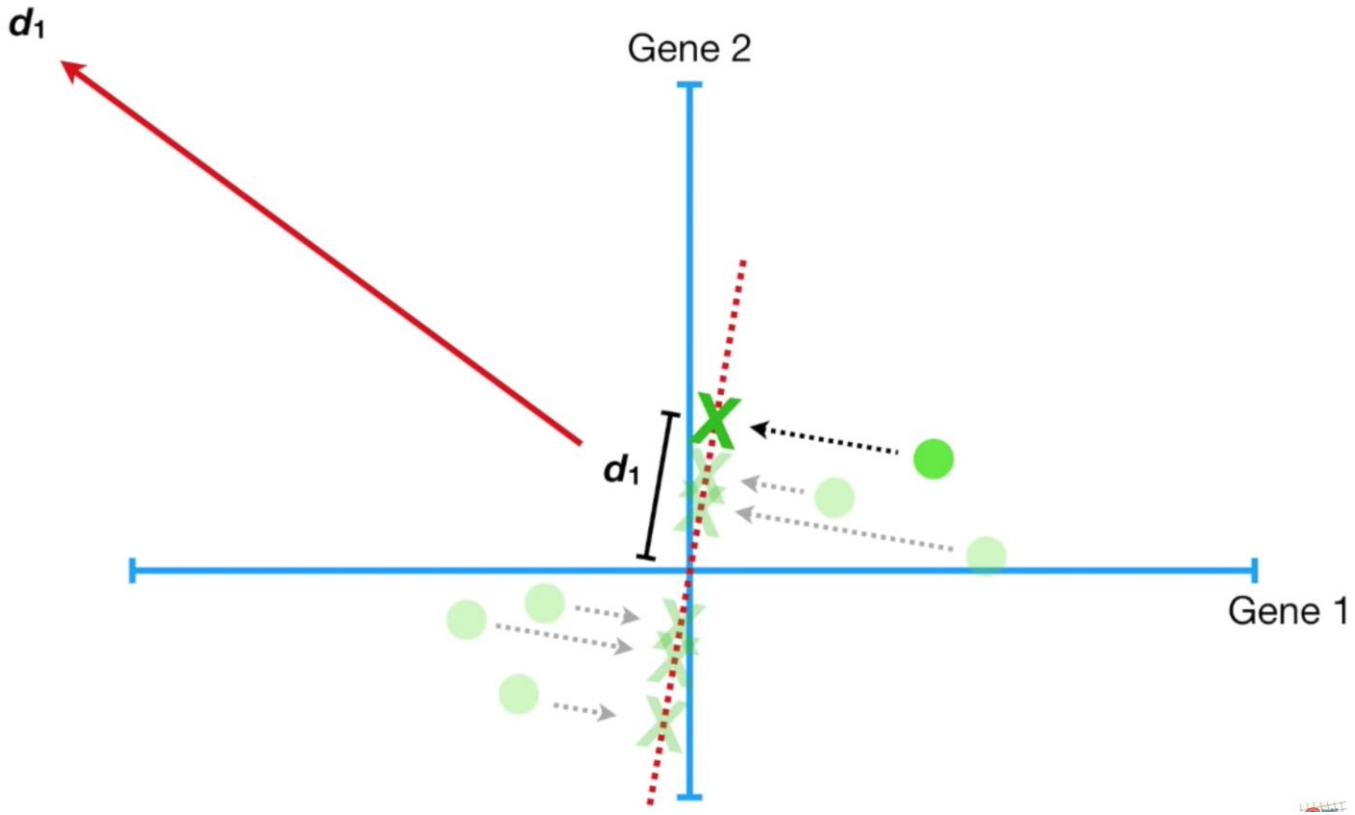
COMPONENTES PRINCIPALES

Es más fácil trabajar maximizando c
(la suma de las distancias al cuadrado
del punto proyectado al origen)



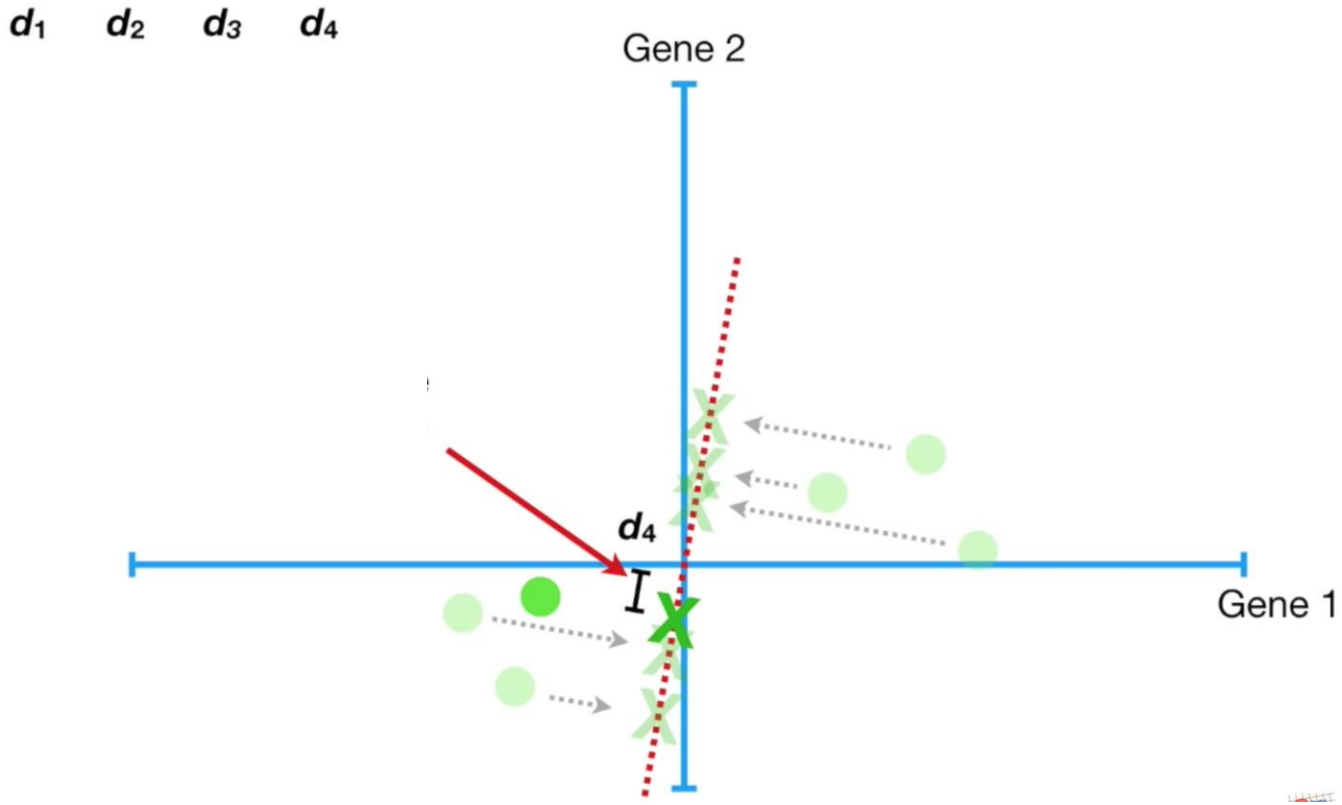
UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES



UNIVERSIDAD ICESI

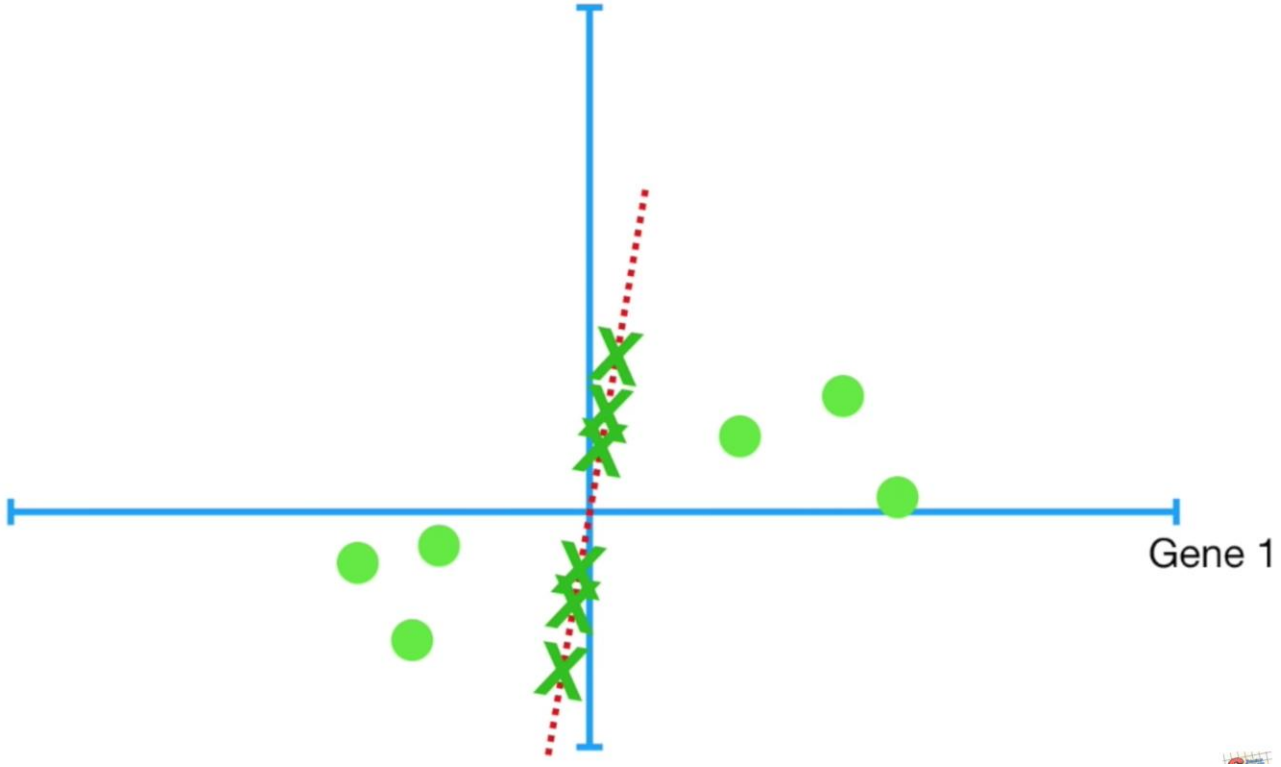
COMPONENTES PRINCIPALES



UNIVERSIDAD ICESI

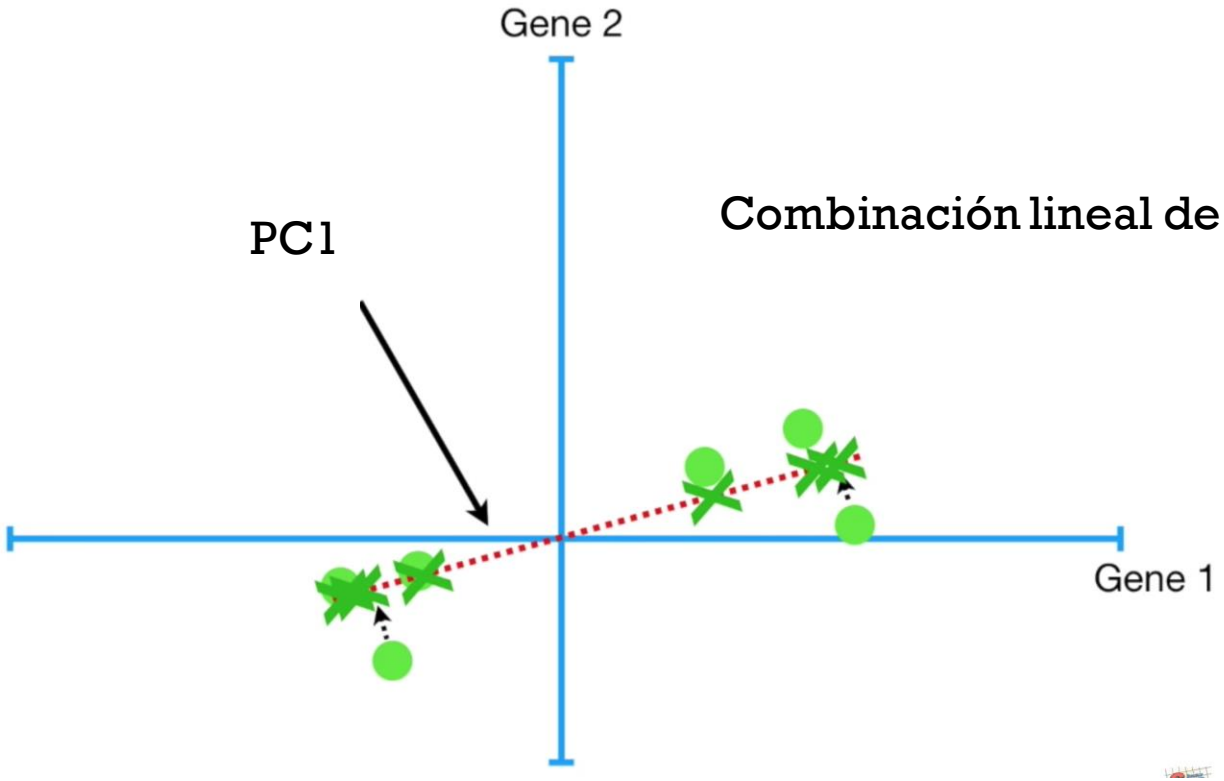
COMPONENTES PRINCIPALES

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



UNIVERSIDAD ICESI

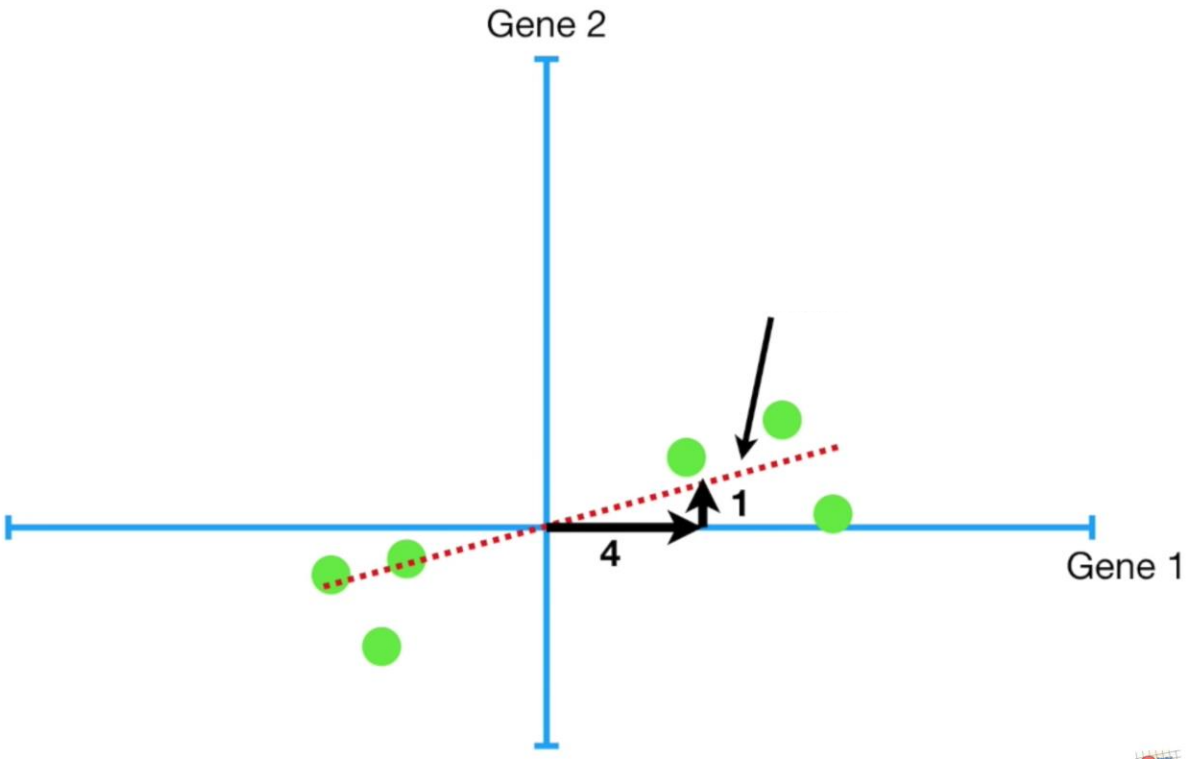
COMPONENTES PRINCIPALES



Combinación lineal de Gene 1 y Gene 2

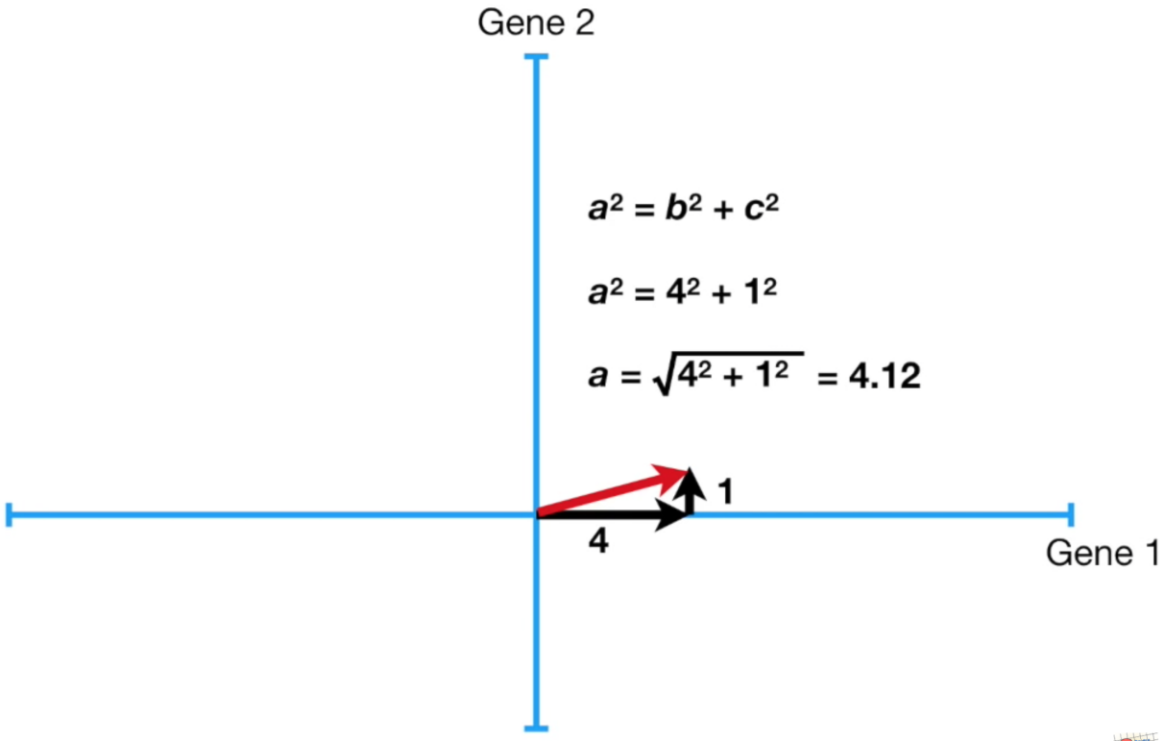
UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES



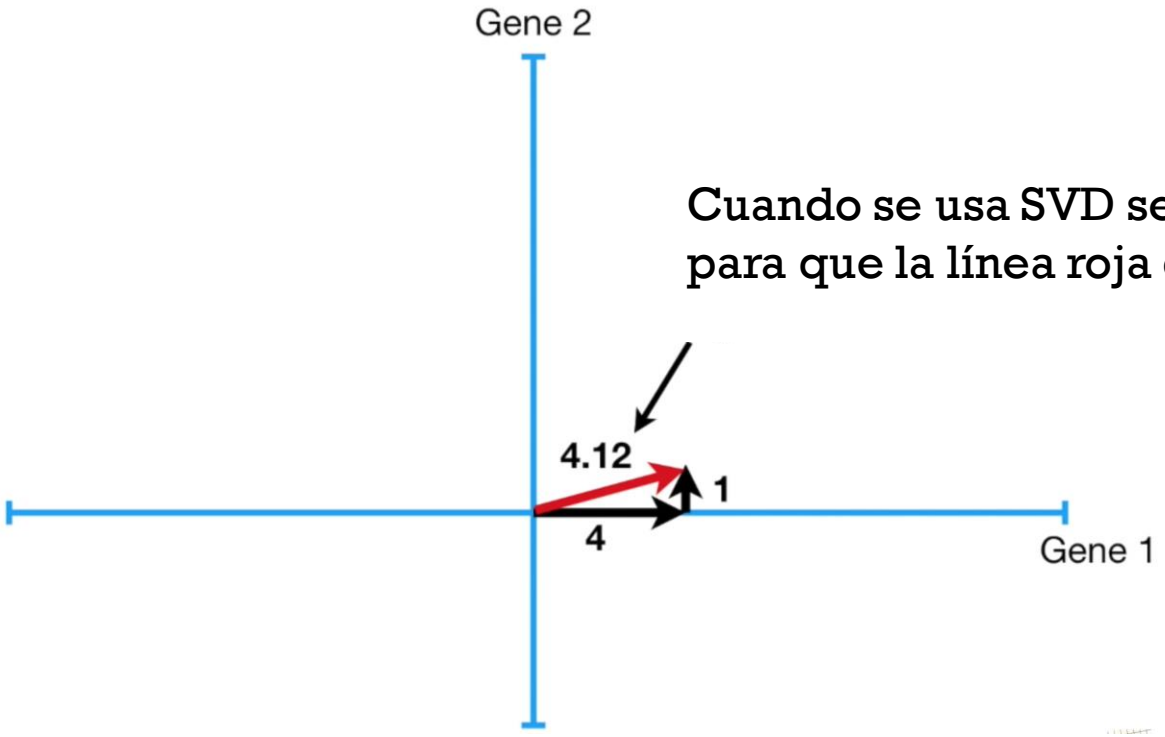
UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES



UNIVERSIDAD ICESI

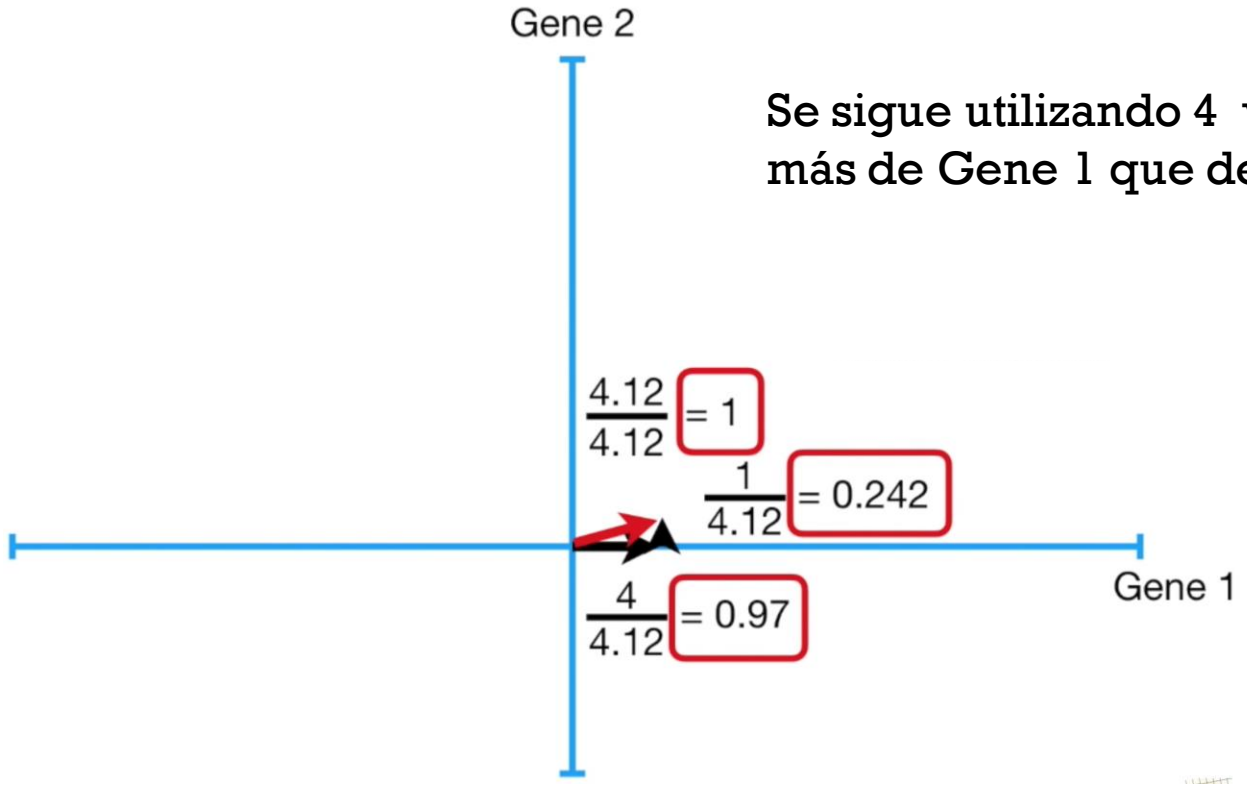
COMPONENTES PRINCIPALES



Cuando se usa SVD se escala para que la línea roja quede en 1

UNIVERSIDAD ICESI

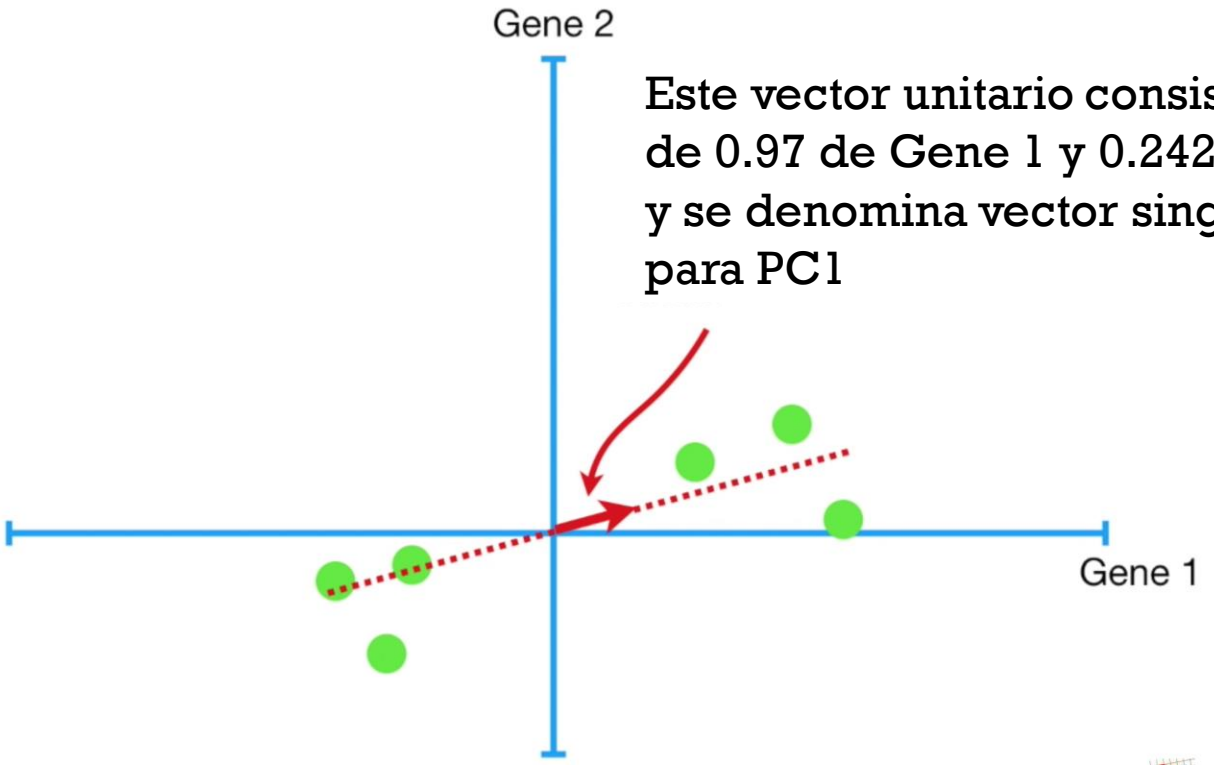
COMPONENTES PRINCIPALES



Se sigue utilizando 4 veces más de Gene 1 que de Gene 2

UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

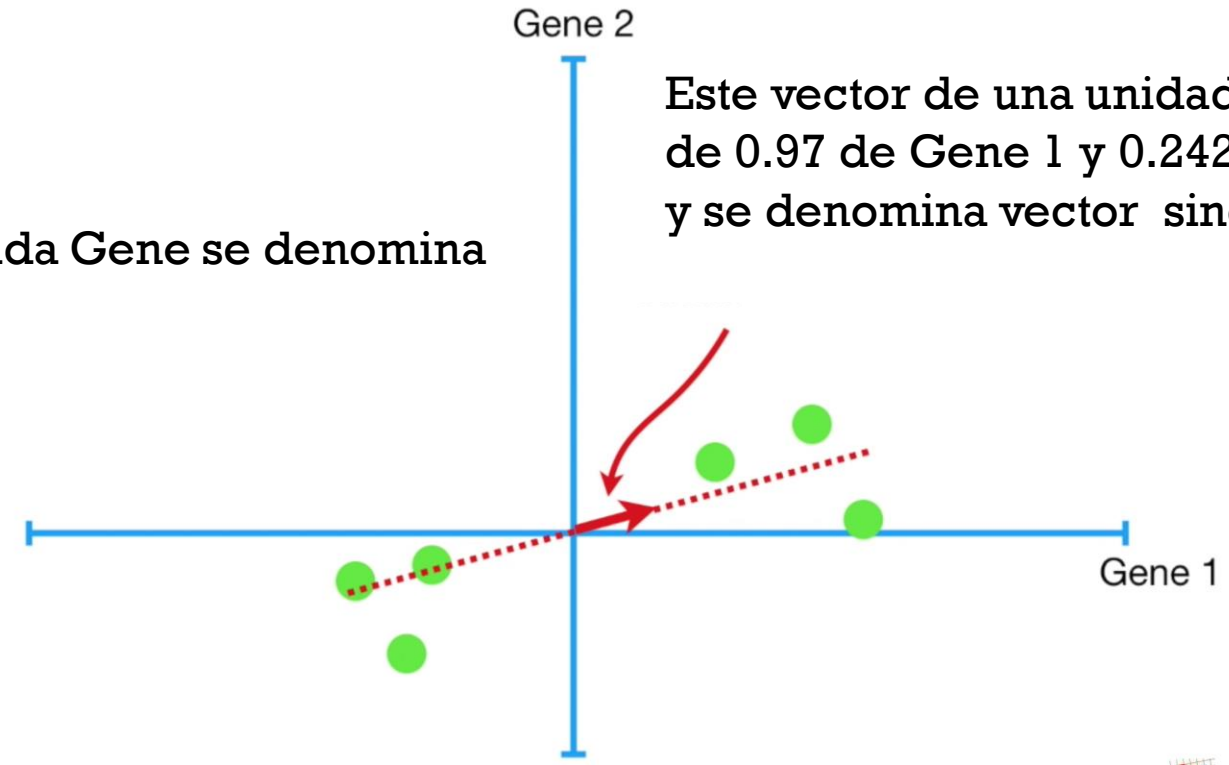


Este vector unitario consiste
de 0.97 de Gene 1 y 0.242 de Gene 2
y se denomina vector singular
para PC1

UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

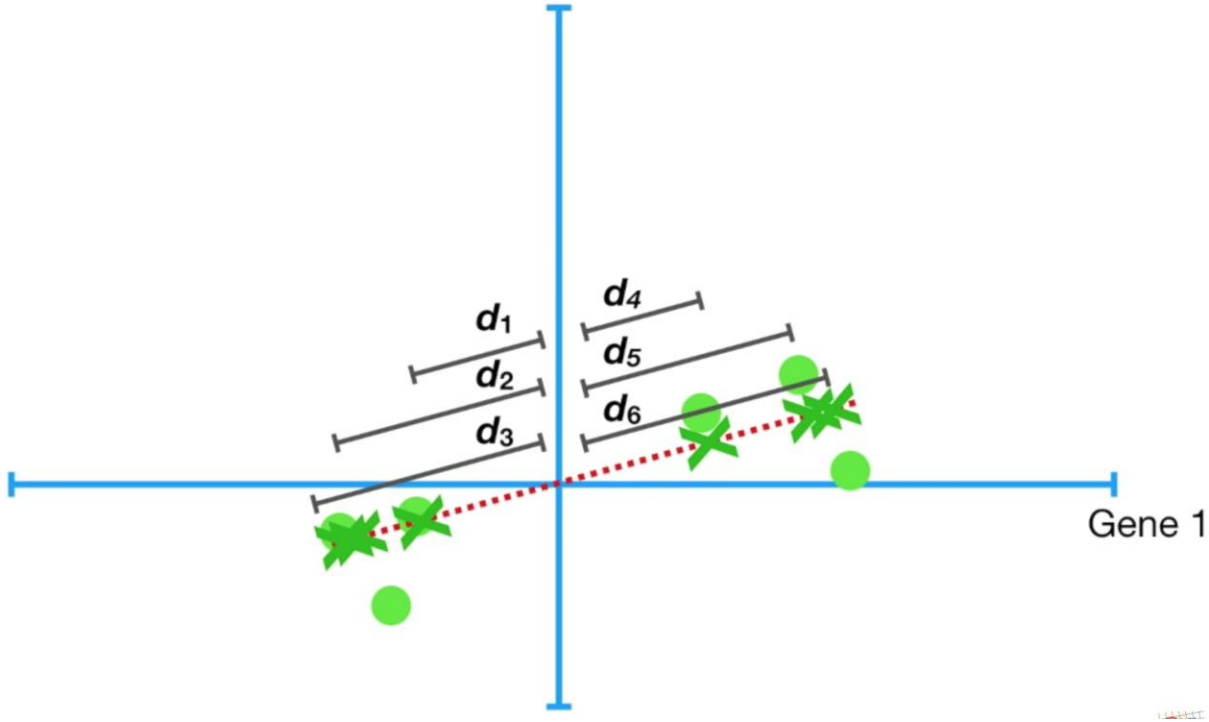
La proporción de cada Gene se denomina Carga (loading)



Este vector de una unidad consiste de 0.97 de Gene 1 y 0.242 de Gene 2 y se denomina vector singular para PC1

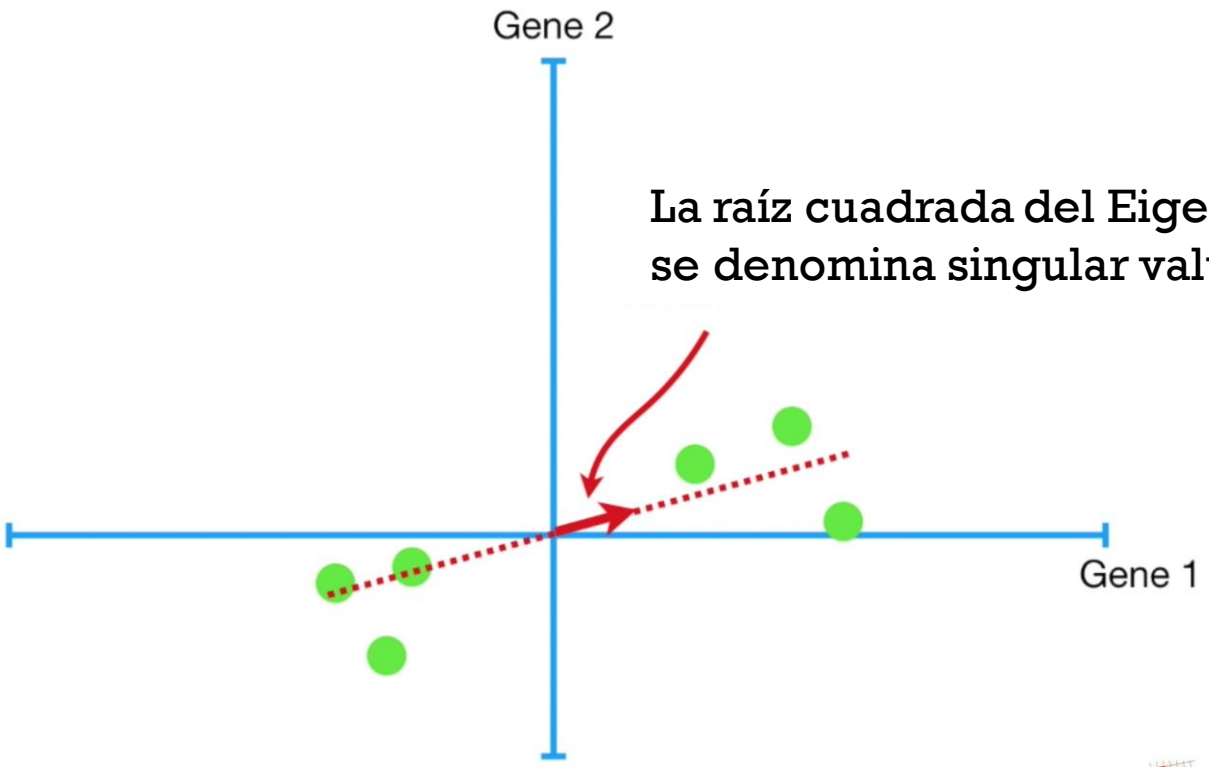
COMPONENTES PRINCIPALES

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



UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

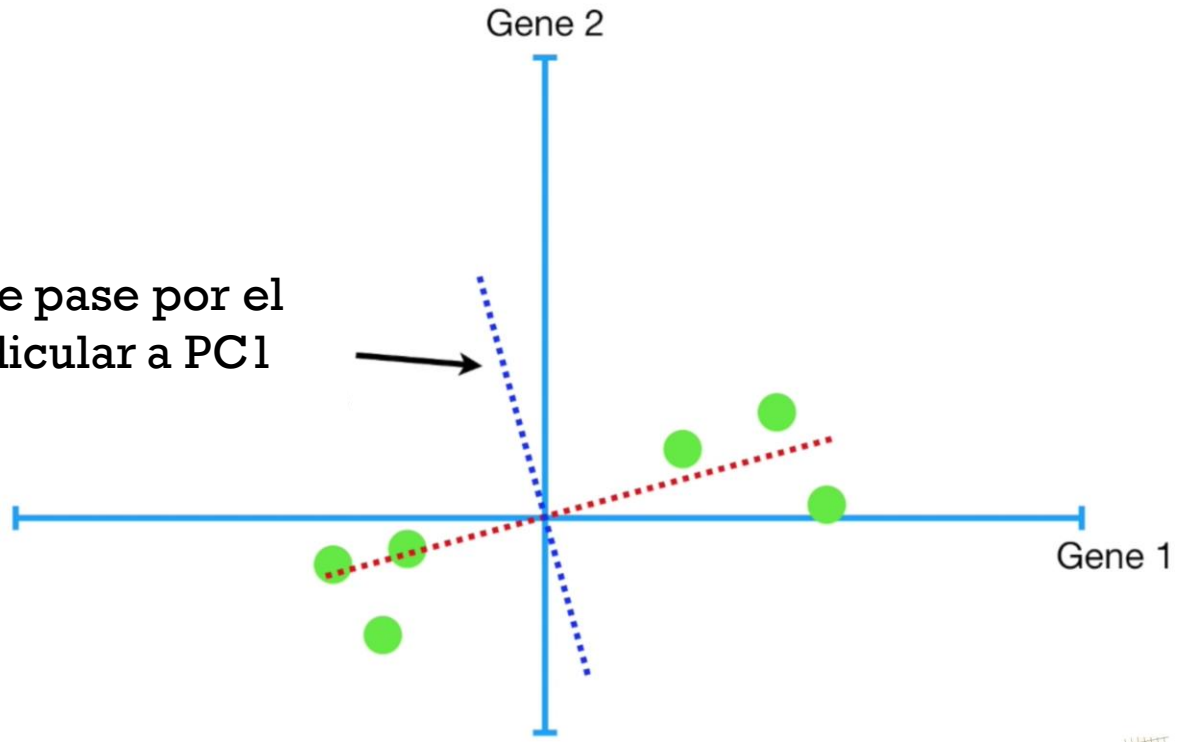


La raíz cuadrada del Eigenvalue para PC1 se denomina singular value para PC1

UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

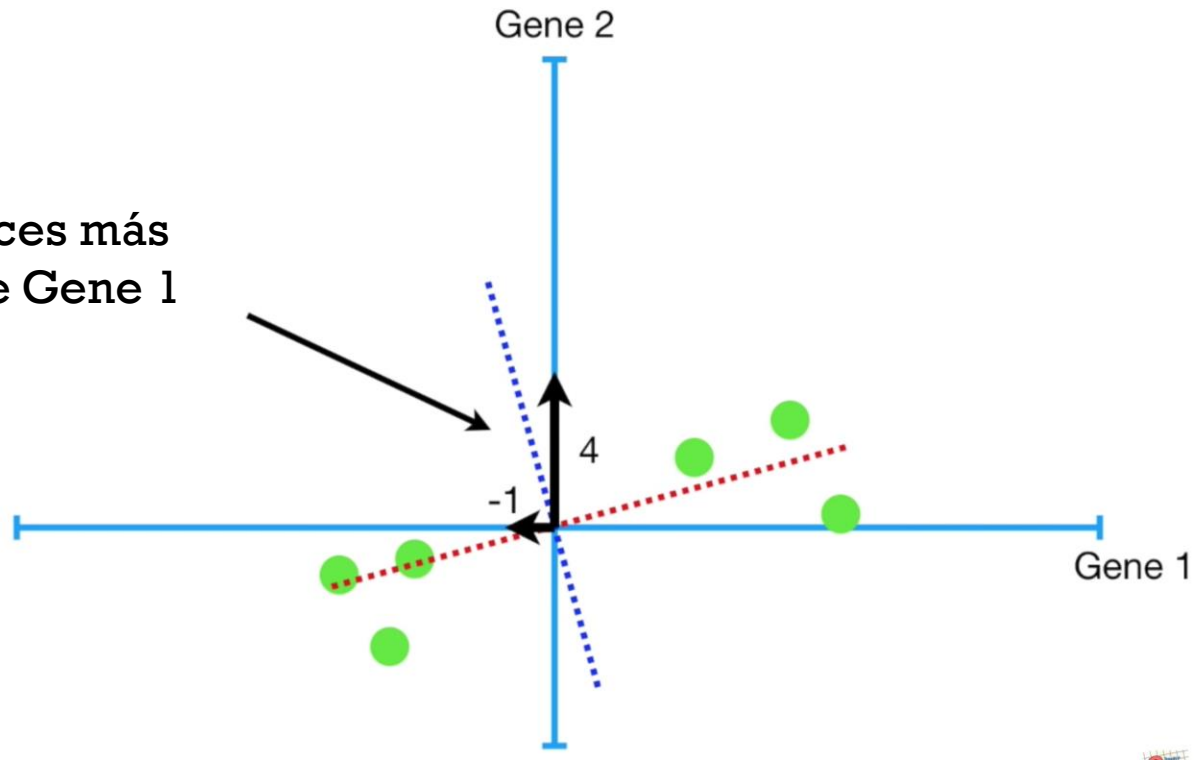
Se escoge la línea que pase por el origen y sea perpendicular a PC1



UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

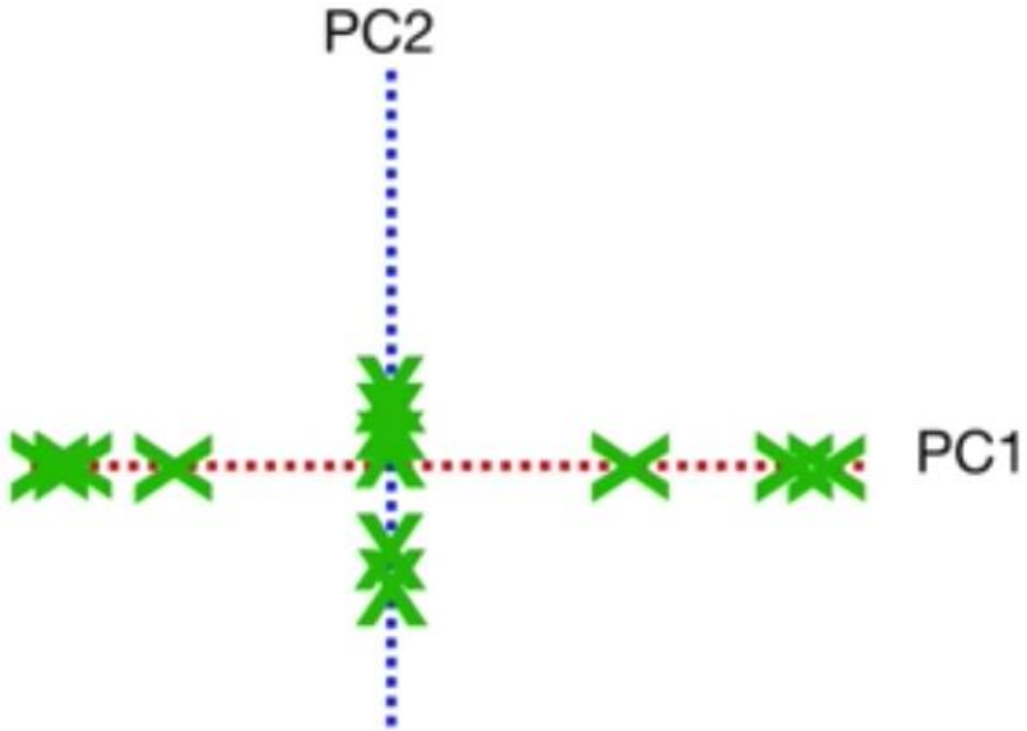
Gene 2 es 4 veces más importante que Gene 1



UNIVERSIDAD ICESI

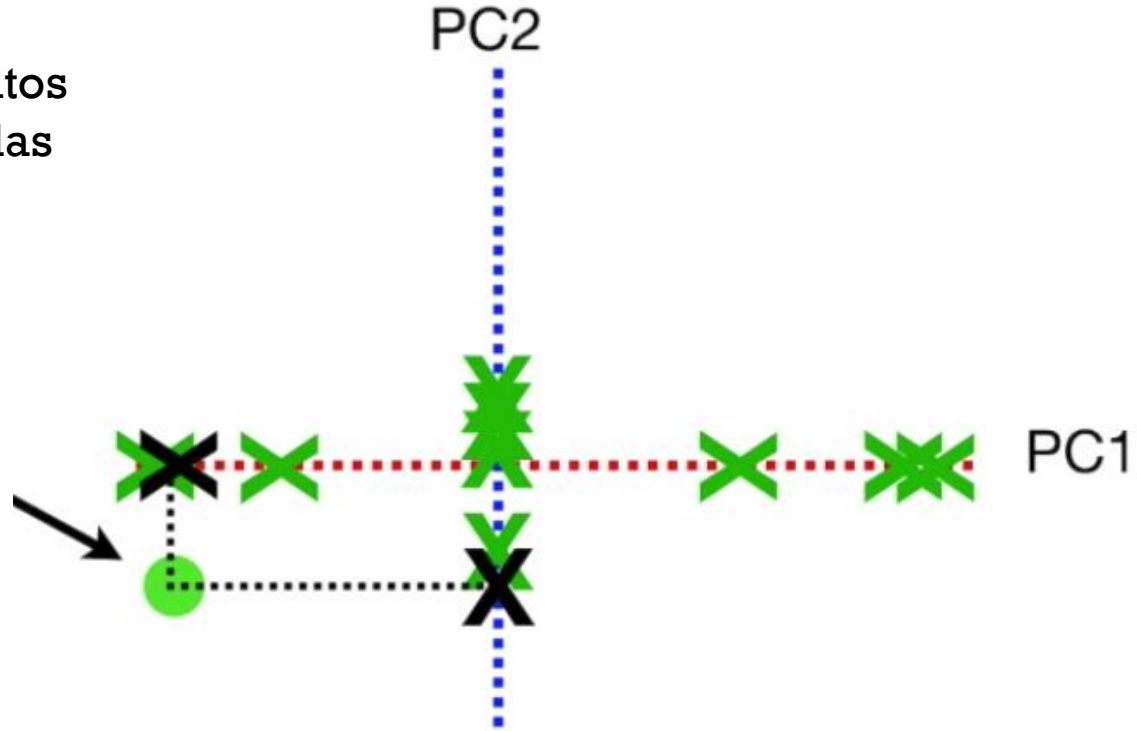
COMPONENTES PRINCIPALES

Rotamos y utilizamos los puntos proyectados para encontrar las muestras



COMPONENTES PRINCIPALES

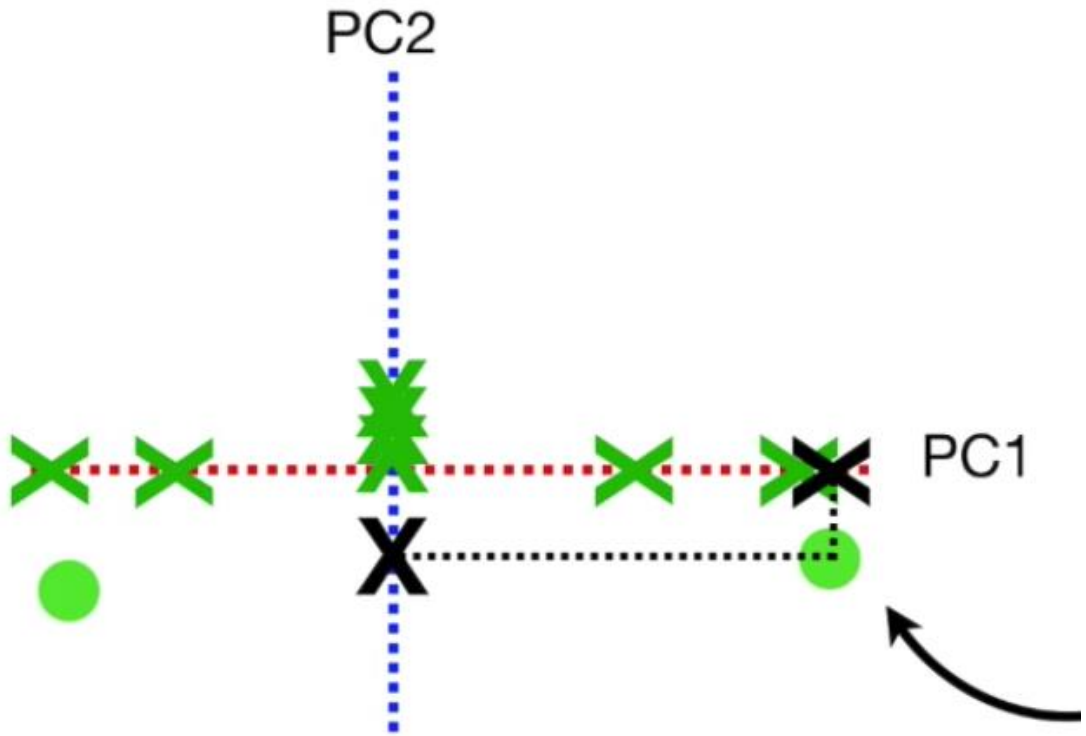
Rotamos y utilizamos los puntos proyectados para encontrar las muestras



UNIVERSIDAD ICESI

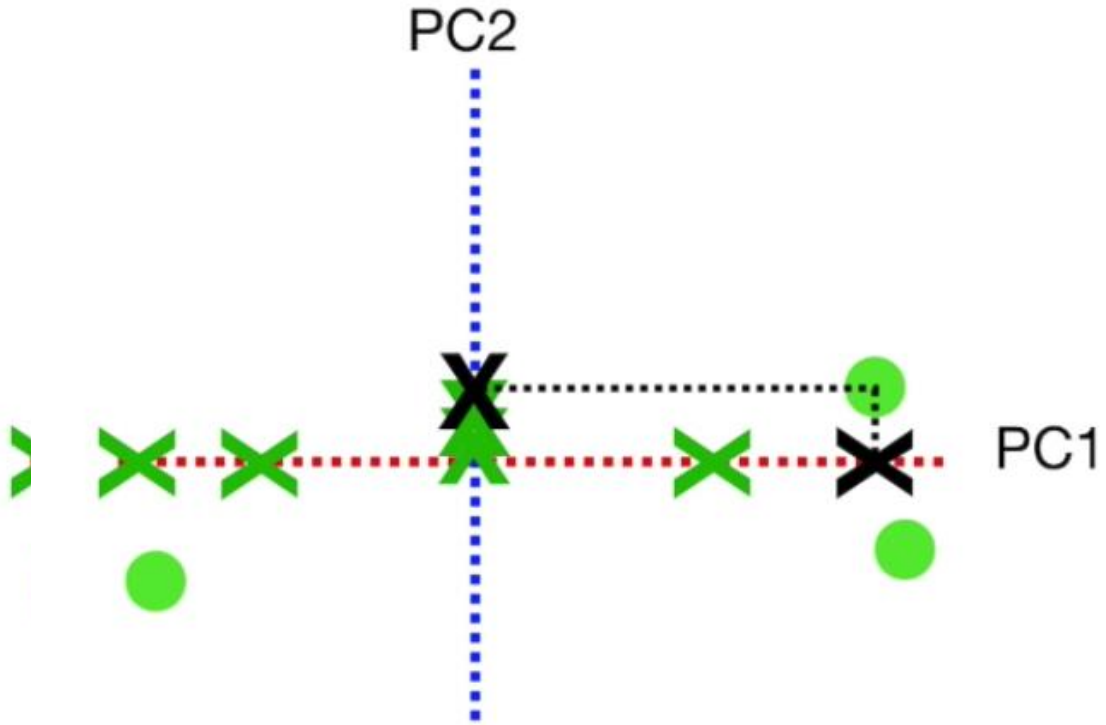
COMPONENTES PRINCIPALES

Rotamos y utilizamos los puntos proyectados para encontrar las muestras



COMPONENTES PRINCIPALES

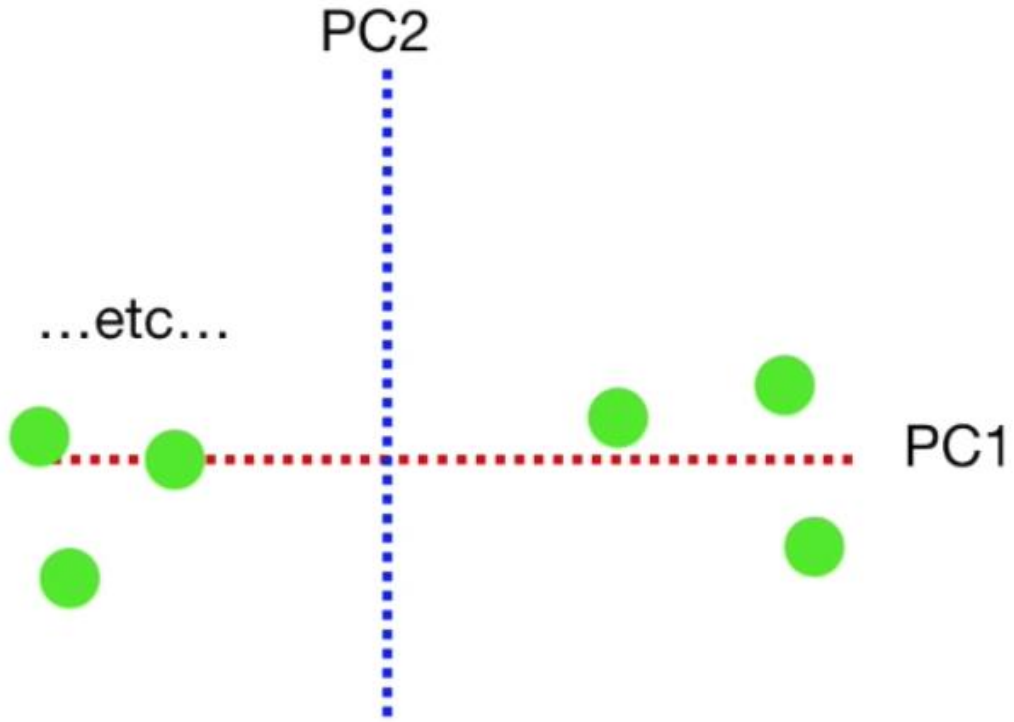
Rotamos y utilizamos los puntos proyectados para encontrar las muestras



UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

Rotamos y utilizamos los puntos proyectados para encontrar las muestras



UNIVERSIDAD ICESI

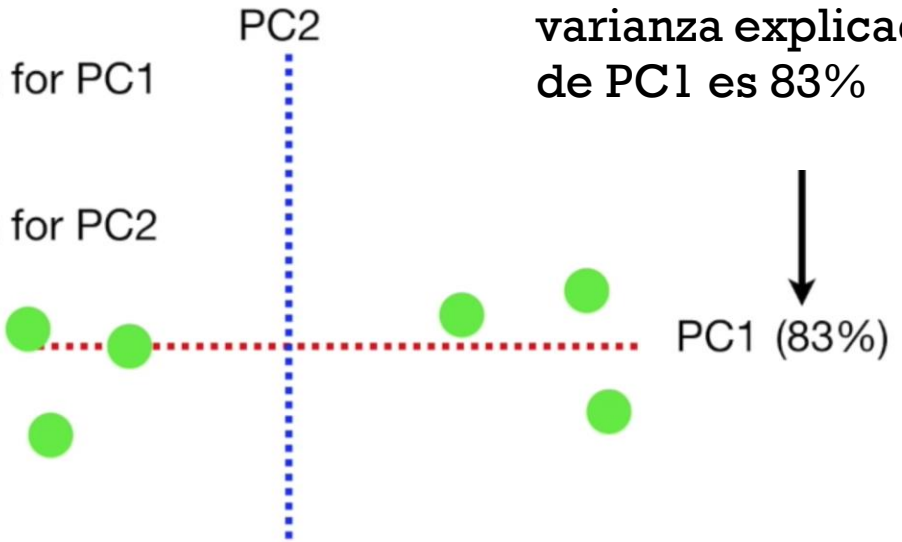
COMPONENTES PRINCIPALES

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

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

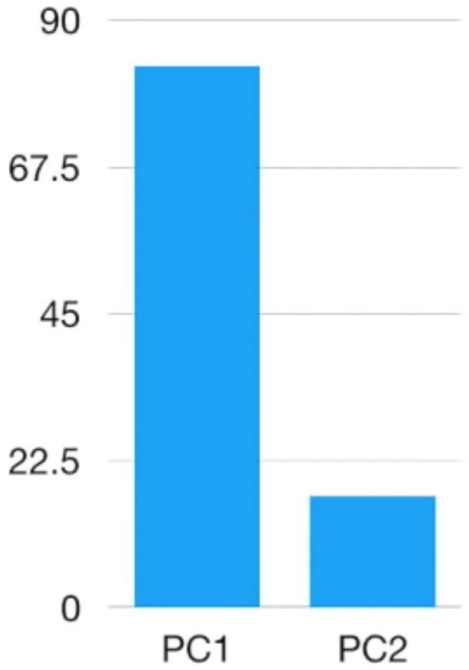
$$\frac{SS(\text{distances for PC2})}{n-1} = \text{Variation for PC2}$$

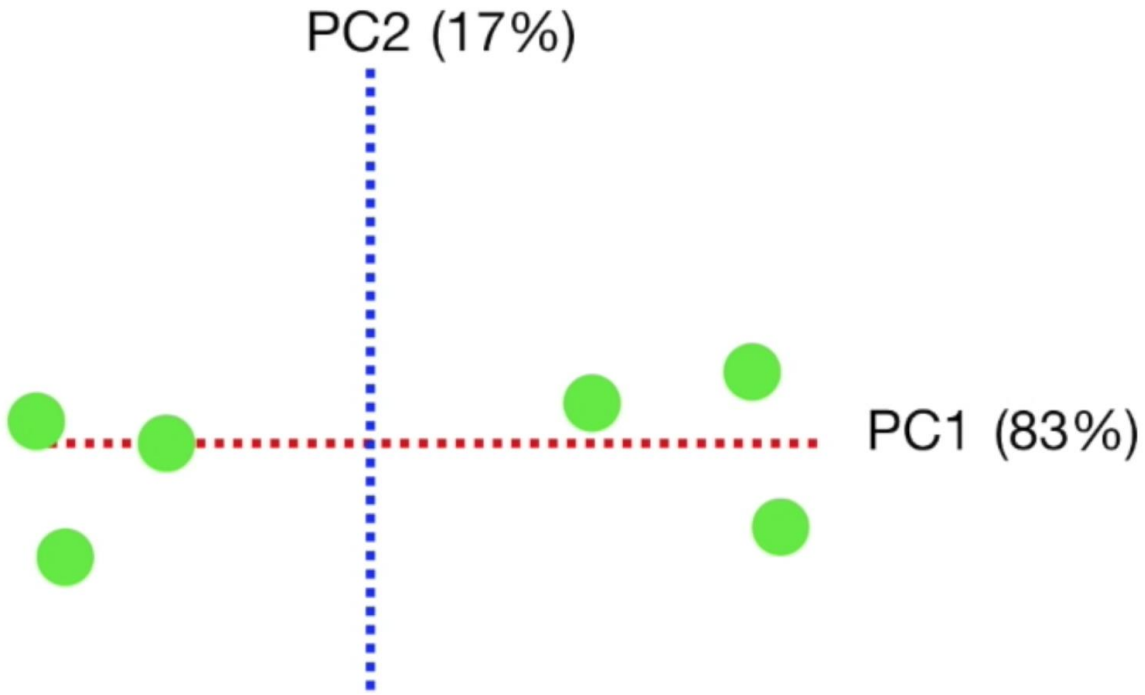
La variación total en ambos PCs es 18 y la proporción de varianza explicada de PC1 es 83%



UNIVERSIDAD ICESI

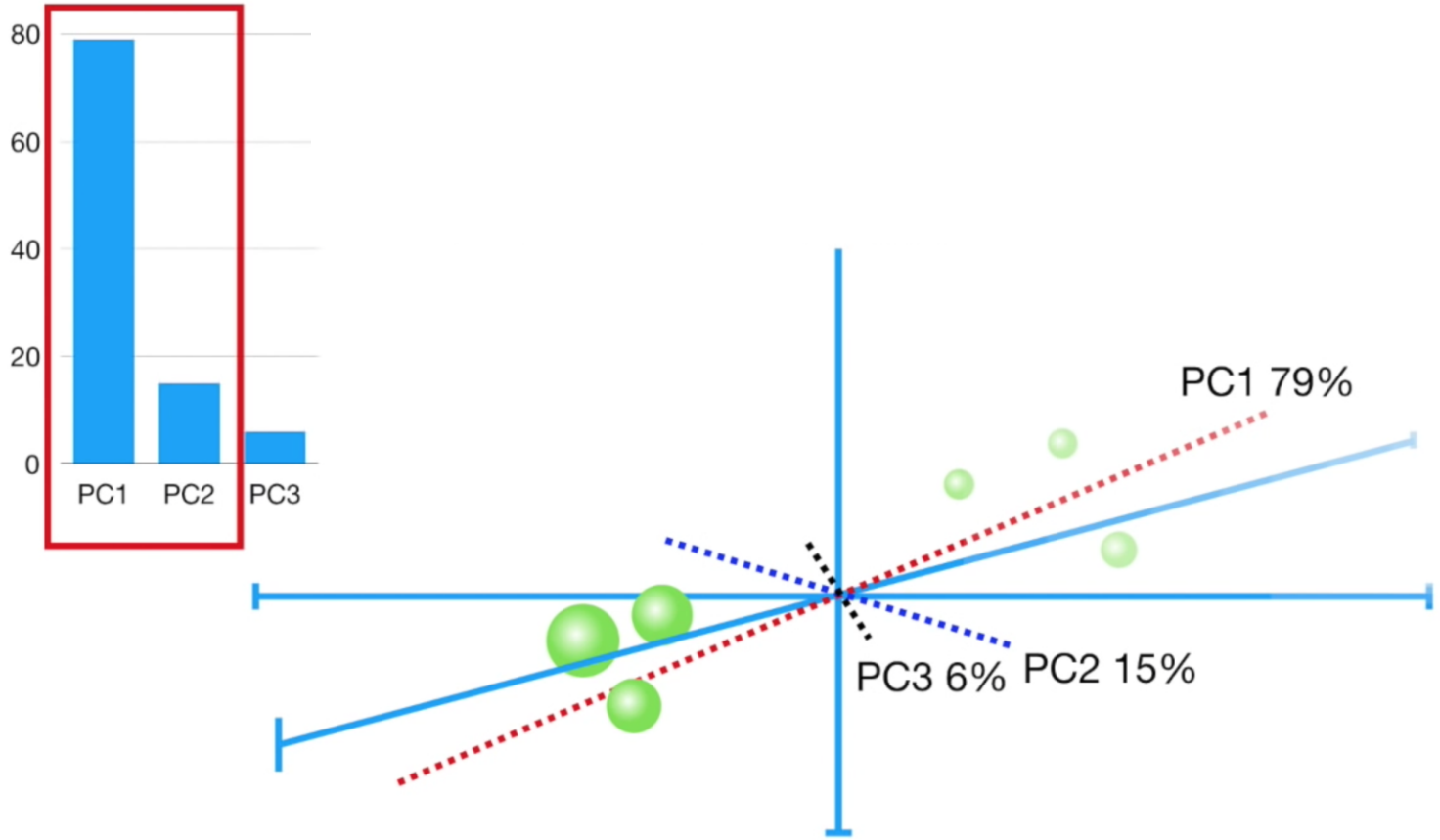
COMPONENTES PRINCIPALES





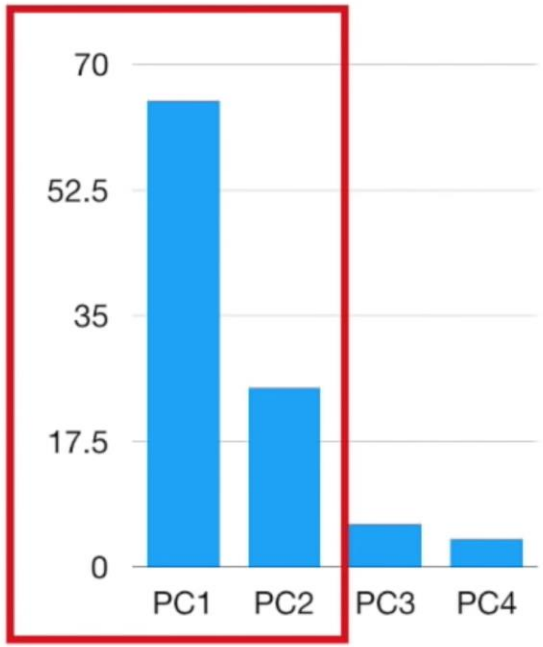
UNIVERSIDAD ICESI

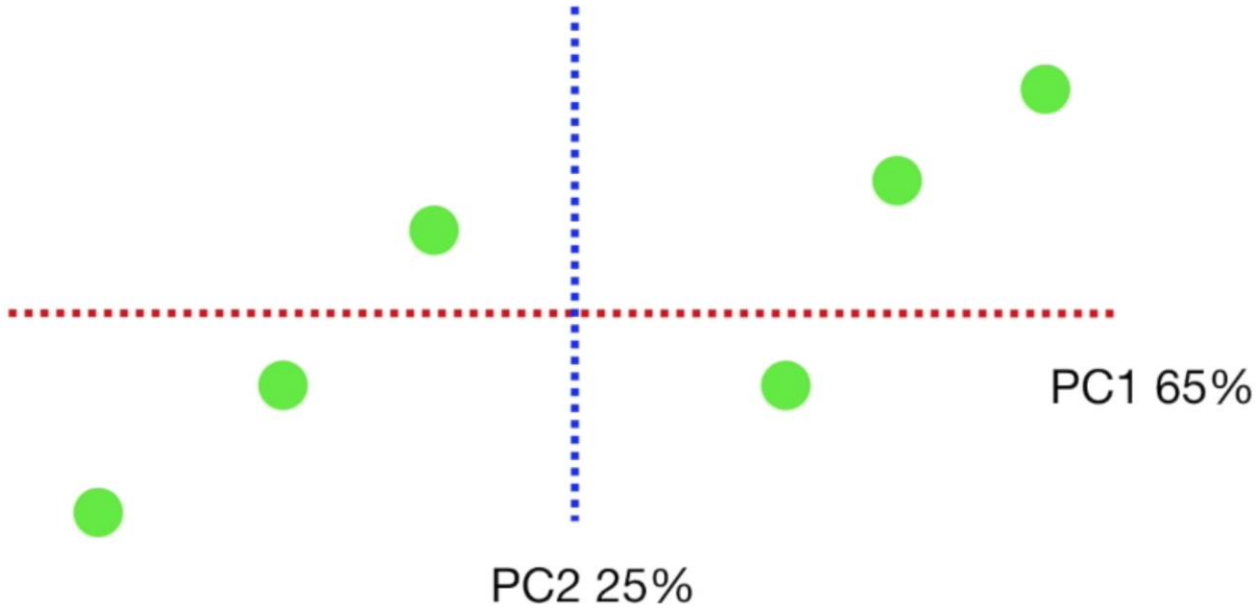
COMPONENTES PRINCIPALES



UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES





UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

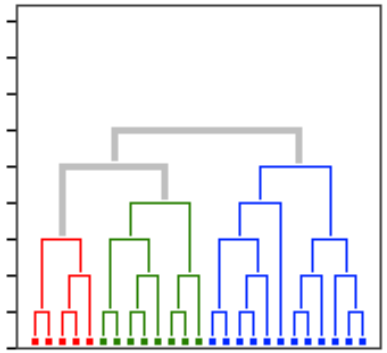
Consideraciones

- La varianza de cada uno de los atributos (dimensiones) originales depende de su escala, por lo que se deben **normalizar** los datos originales
- El número de dimensiones originales puede ser superior al número de instancias del dataset, pero se limita 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
- Considerar solo las componentes principales más importantes permite reducir la influencia del ruido en los datos.
- A partir de una transformación inversa del espacio de los PCs hacia el espacio original podemos entonces **filtrar el ruido**.

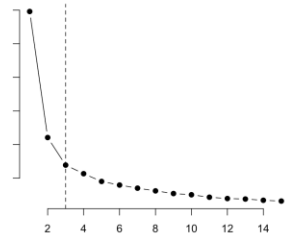
TALLER: COMPONENTES PRINCIPALES

1. Realizar el taller de clustering de clientes de supermercado 10-SUPERMERCADOS-PCA-STUD.html con la parte dedicada a la reducción de dimensionalidad a partir de PCA
2. Continuar el taller de clustering de clientes que desertan aplicando reducción de dimensionalidad a partir de PCA

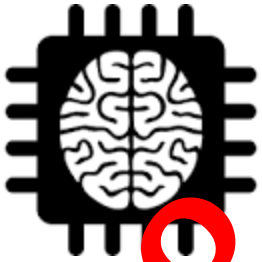
AGENDA



Clustering jerárquico

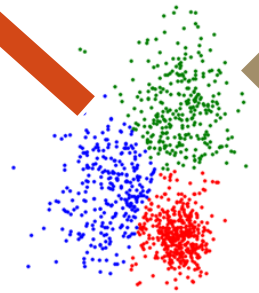


Evaluación del clustering

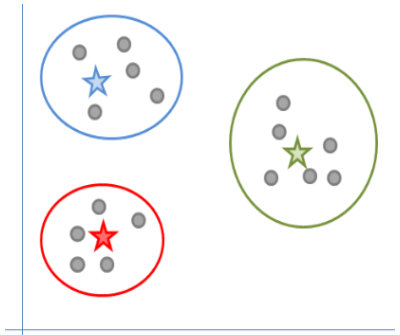


Aprendizaje automático

REPASO



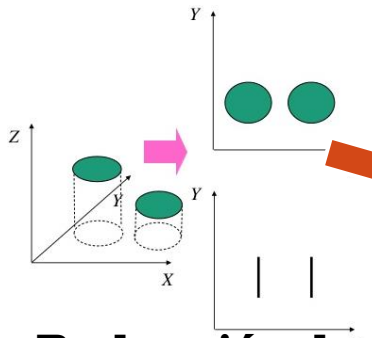
Clustering



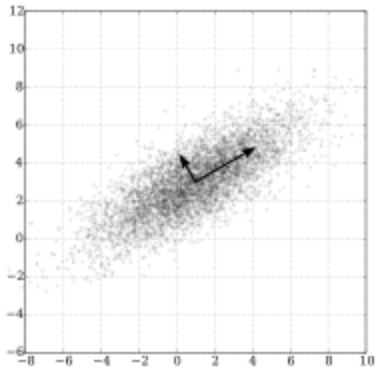
K-means

Aprendizaje supervisado

Aprendizaje no supervisado



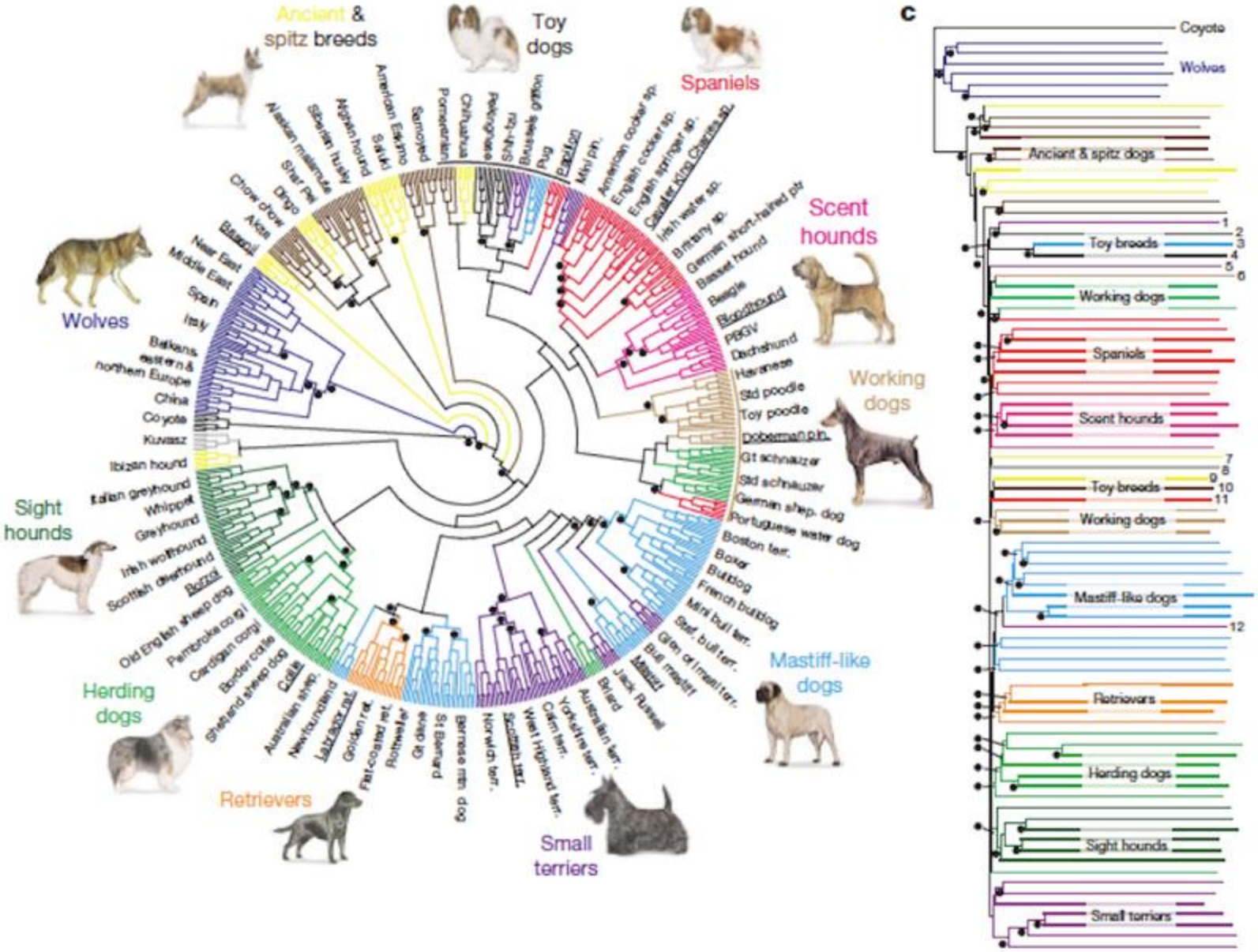
Reducción de dimensionalidad



Componentes principales

UNIVERSIDAD ICESI

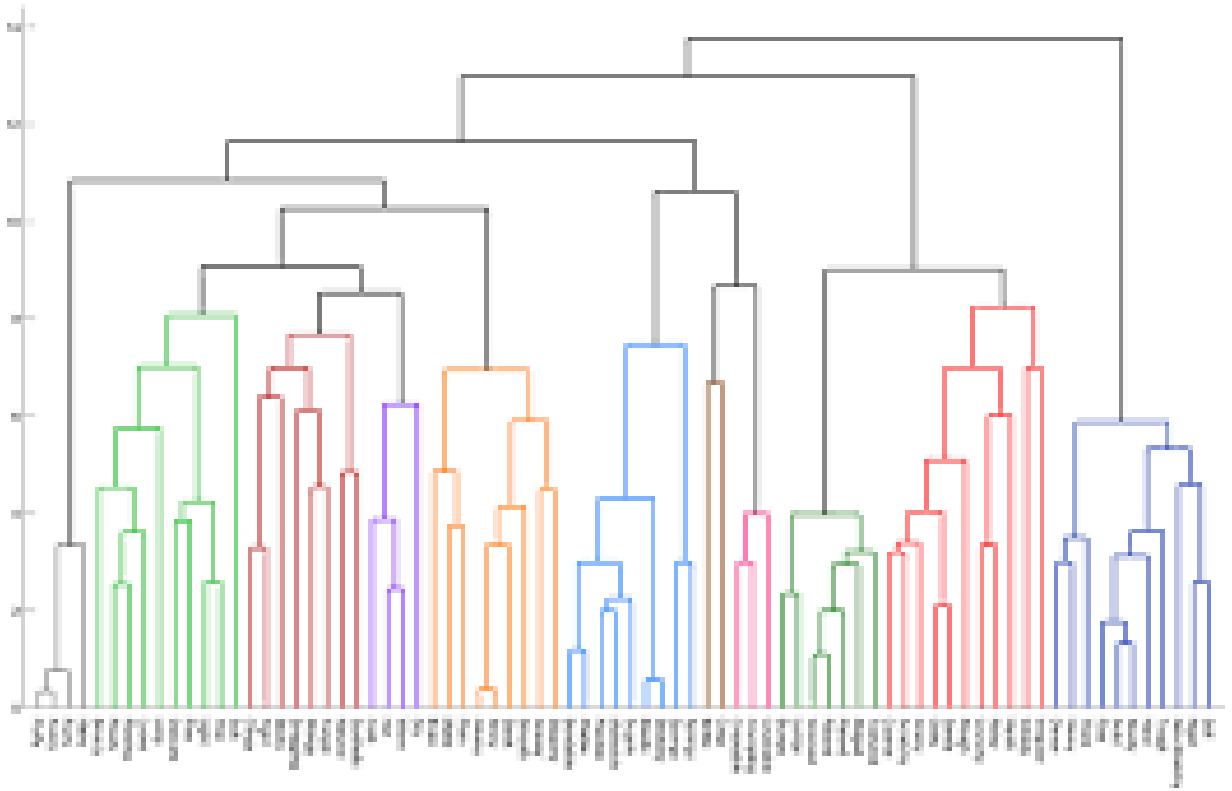
CLUSTERING JERÁRQUICO



UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO

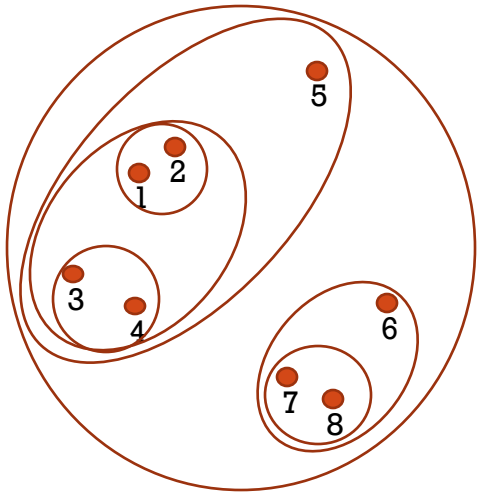
- Aproximación **bottom-up**
- Produce como resultado un **dendrograma**
  - Basado en las **distancias** entre instancias y entre clusters
  - Determina todas las segmentaciones posibles, permitiendo su visualización
- No se necesita repetir el proceso para diferentes valores de **K**
- Las instancias **excepcionales** pueden ser rápidamente identificadas

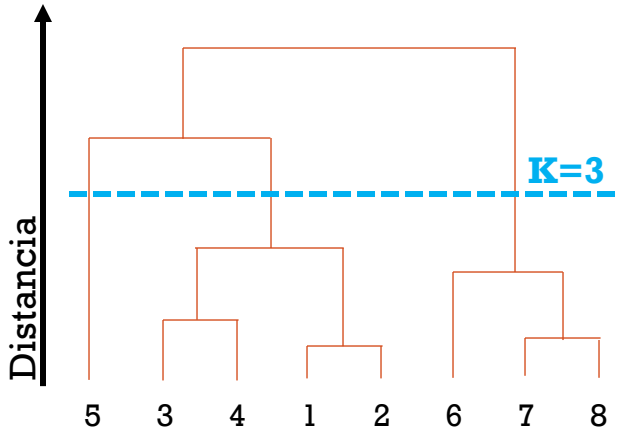


UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO

- Algoritmo (iterativo):
  1. Al inicio cada instancia es un cluster (n clusters)
  2. Se identifica el par de clusters más cercanos y se fusionan (n-1 clusters)
  3. Se repite el paso anterior hasta que queda un solo cluster con todas las instancias
  4. Se escoge un punto de corte
- Los clusters se pueden organizar en forma de **dendrograma**
- Es necesario definir como **fusionar** clusters y la **distancia** a utilizar

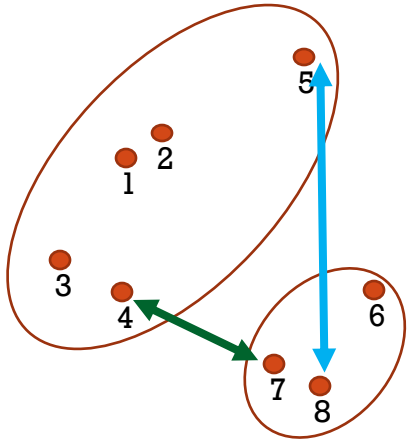




UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO

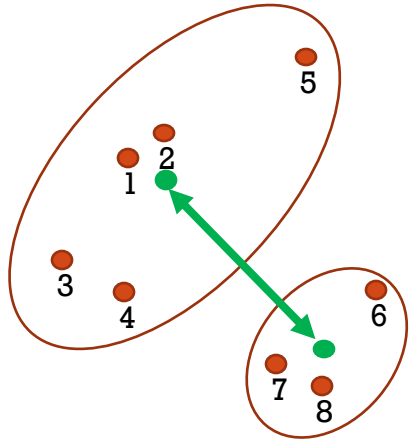
- **Fusión entre clusters** basadas en el cómputo de las distancias entre todos los pares de puntos de cada cluster:
  - **Single linkage**:
    - Distancia mínima entre dos puntos de los dos clusters.
    - Resultan clusters formados por "cadenas" de puntos, usualmente con fusiones consecutivas entre un cluster y un punto cercano
    - Sensible al ruido y a las excepciones
  - **Complete linkage**:
    - Distancia máxima entre dos puntos de los dos clusters.
    - Tiende hacia clusters esféricos con diámetros consistentes
  - **Average linkage**:
    - Promedio de las distancias entre todos los pares de puntos
    - Punto intermedio entre single y complete linkage
    - Menos afectado por las excepciones



→ Complete y average se prefieren sobre single linkage

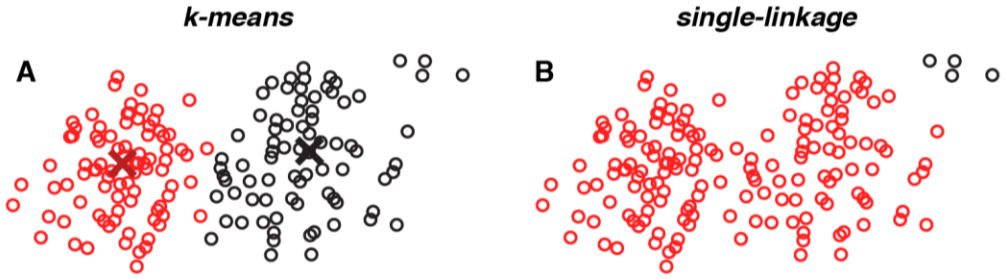
CLUSTERING JERÁRQUICO

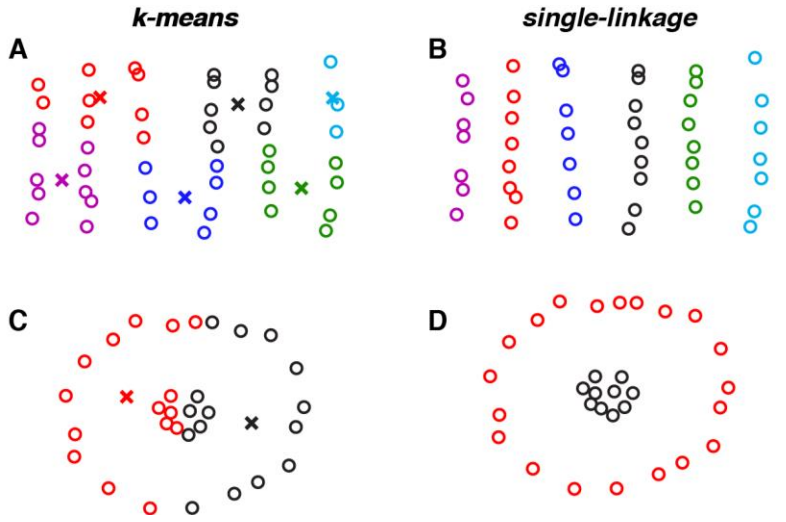
- Otros tipos de **fusión entre clusters**
  - **Centroide**:
    - Distancia entre los centroides de los clusters
    - Sufre de *inversiones*, cuando el punto de fusión de dos clusters en el dendrograma es inferior al de alguno de los clusters fusionados



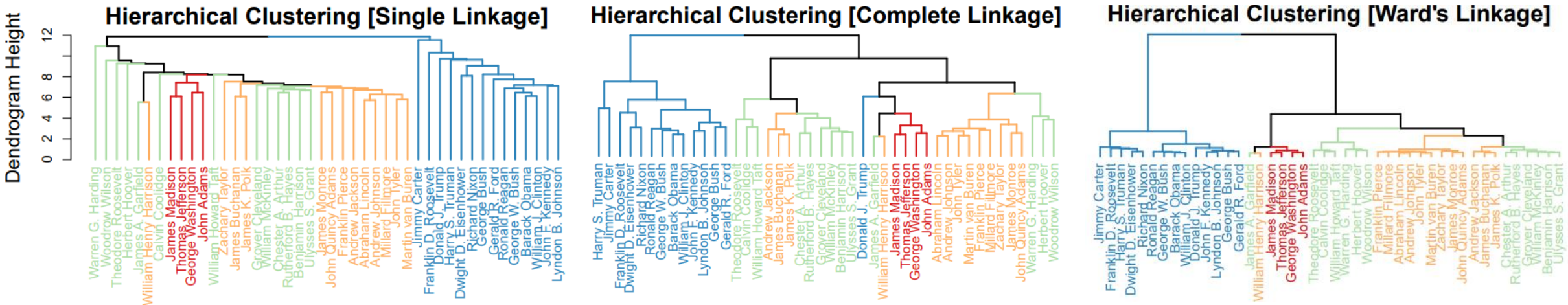
  - **Ward**: Para cada fusión se analiza el cambio en la varianza
    - Con cada fusión, la varianza global del conjunto de clusters aumenta
    - Se escoge la fusión cuyo aumento de varianza es mínimo

CLUSTERING JERÁRQUICO





http://alexhwilliams.info/itsneuronalblog/2015/09/11/clustering1/



https://arxiv.org/pdf/1901.01477.pdf

CLUSTERING JERÁRQUICO

- Consideraciones
  - La pertenencia de las instancias a los clusters es absoluta
  - Requiere poder de cálculo computacional grande
  - Una vez una fusión se decide, no hay vuelta atrás
  - Dependiendo de la distancia utilizada y al tipo de fusión:
    - Sensible al ruido y a excepciones
    - Dificulta gestionar clusters de tamaños diferentes o no convexos
    - Puede llegar a particionar clusters grandes
  - Influencia de las unidades de los atributos utilizados → estandarización
  - ¿Qué punto de corte (k) escoger?

TALLER: CLUSTERING JERÁRQUICO

Realizar el taller 10-SYNTH- HClust-STUD.html con datos sintéticos

UNIVERSIDAD ICESI

REFERENCIAS

- *Python Machine Learning*, Sebastian Raschka, Packt, 2015
- *Introduction to Statistical Learning with Applications in R (ISLR), G. James, D. Witten, T. Hastie & R. Tibshirani, 2014*
- EMC2, "Data science and big data analytics", 2015, John Wiley & Sons
- *Data Science for Business*, Foster Provost & Tom Fawcett, O'Reilly, 2013
- *Practical Data Science with R*, Nina Zumel & John Mount, 2014