

PCA

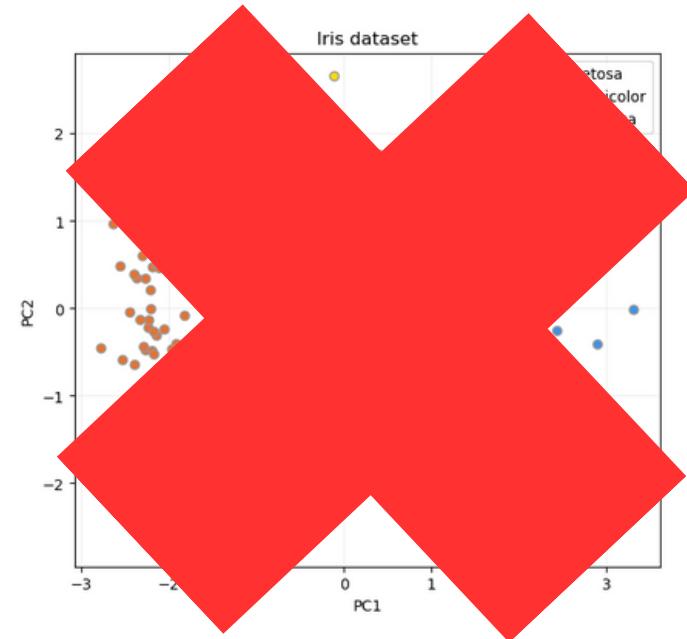
Principal Component Analysis

Viernes de Bioinformática

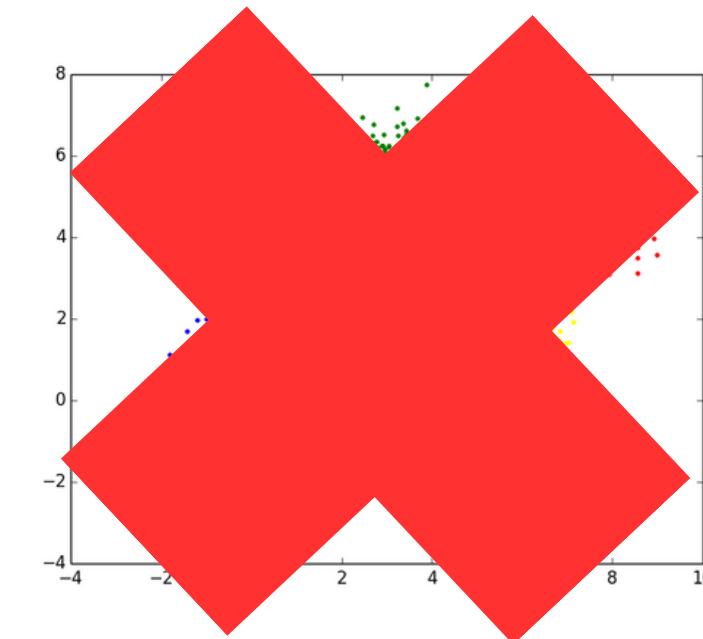


“Solo le quité lo que le sobraba”

¿Qué es el PCA?



¿Un gráfico?



¿método de clusterización de datos?

🌟 Método de reducción de dimensiones 🌟

¿Qué es el PCA?

El análisis de componentes principales, o PCA, **reduce el número de dimensiones** de grandes conjuntos de datos a **componentes principales** (PC) que conservan la mayor parte de la información original.

Es el abstract de tus datos.

¿Cómo funciona el PCA?

- ¿Puede el PCA comprender qué parte de nuestros datos es importante? ¿Podemos cuantificar matemáticamente la cantidad de información contenida en los datos? La **varianza** sí puede.
- Cuanto mayor sea la varianza, mayor será la información. Y viceversa.

Varianza

$$Var(X) = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

$$Var(X) = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \cdots + (x_n - \bar{X})^2}{n}$$

La varianza mide el grado medio en que cada punto difiere de la media.

¿Cómo que la varianza es información?



¿Puedes adivinar quién es quién?
Es difícil cuando son muy parecidos en altura.

Del mismo modo, cuando nuestros datos tienen una mayor varianza, contienen más información.



La **varianza** es información.

Pensemos en genes...

Genes	Sujeto 1	Sujeto 2	Sujeto 3	Sujeto 4	Sujeto 5	Sujeto 6	...
Gen 1	10	11	8	13	1	2	
Gen 2	6	4	5	3	2	1	
Gen 3	120	104	112	187	240	220	

...

Pero también podemos pensar en...

Variables	Sujeto 1	Sujeto 2	Sujeto 3	Sujeto 4	...
Edad	10	11	18	21	
Peso	60	45	85	81	
Conteo leucocitario	50	148	485	541	

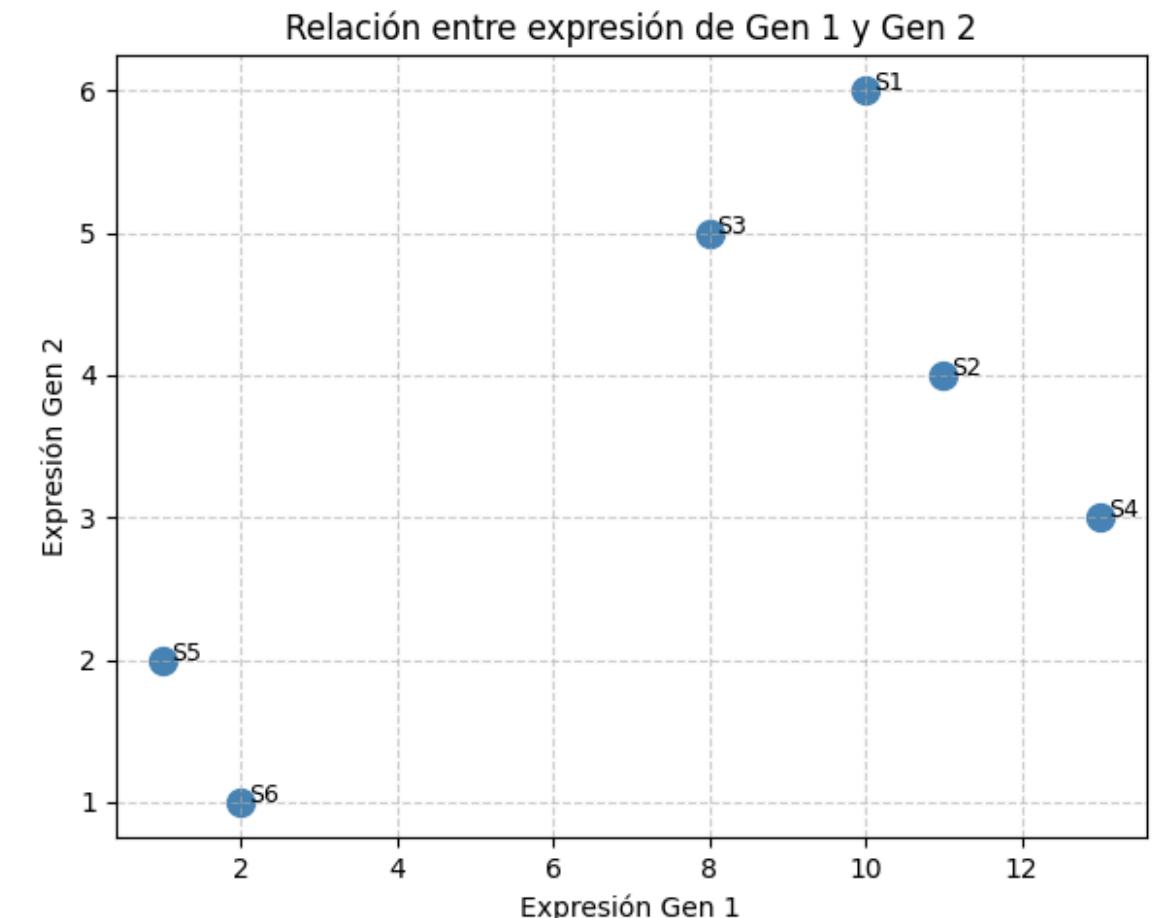
...

Pensemos en genes...

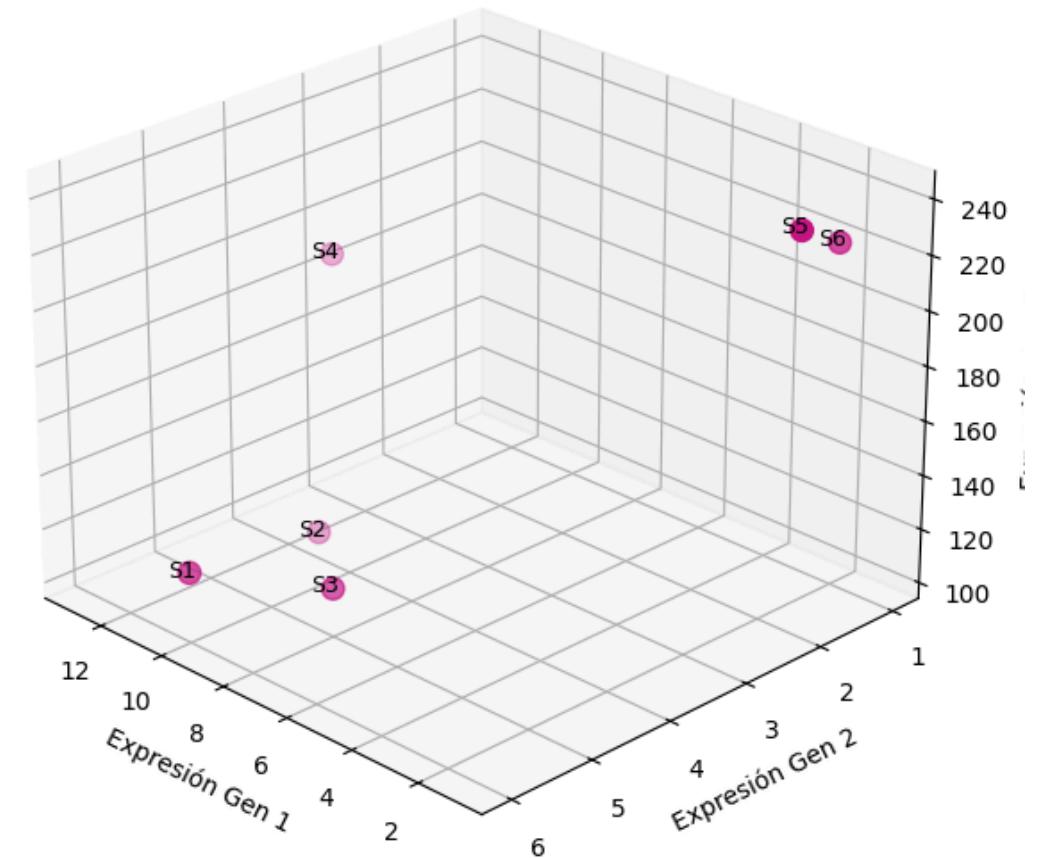
Genes	Sujeto 1	Sujeto 2	Sujeto 3	Sujeto 4	Sujeto 5	Sujeto 6	...
Gen 1	10	11	8	13	1	2	
Gen 2	6	4	5	3	2	1	
Gen 3	120	104	112	187	240	220	

...

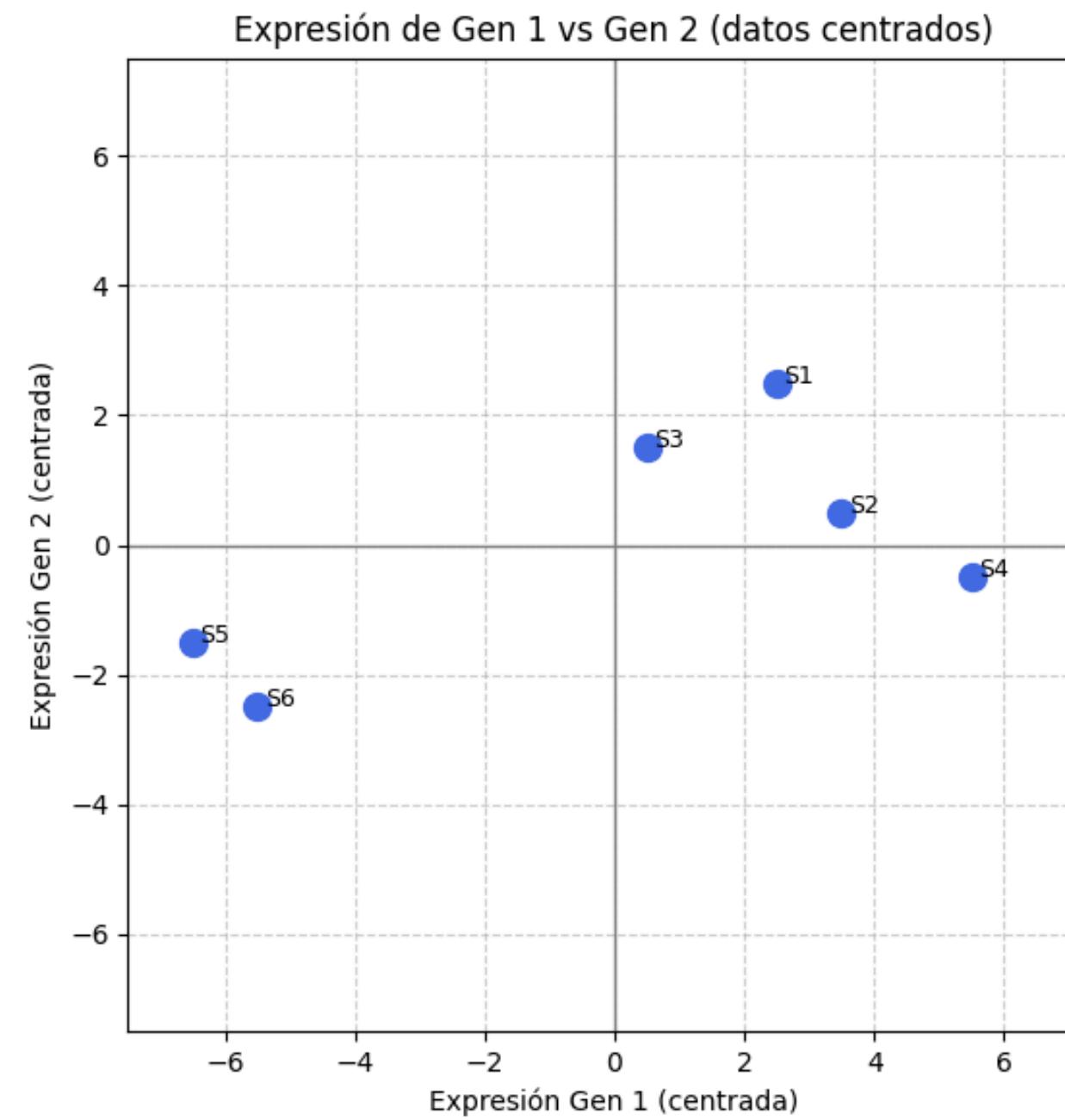
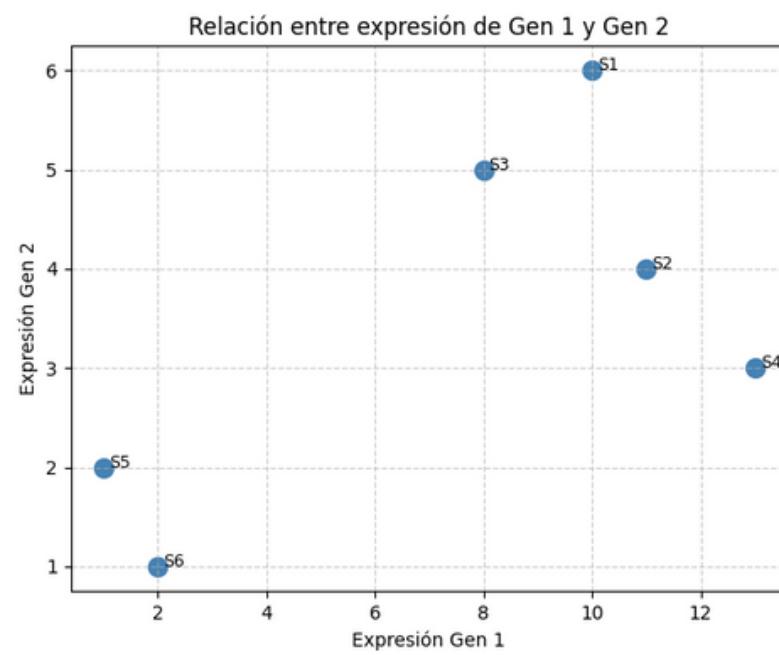
**El PCA puede tomar mas de 4 dimensiones y hacer un gráfico en 2D
Y puede decirte cual variable es más importante en términos de
explicacion de la varianza**



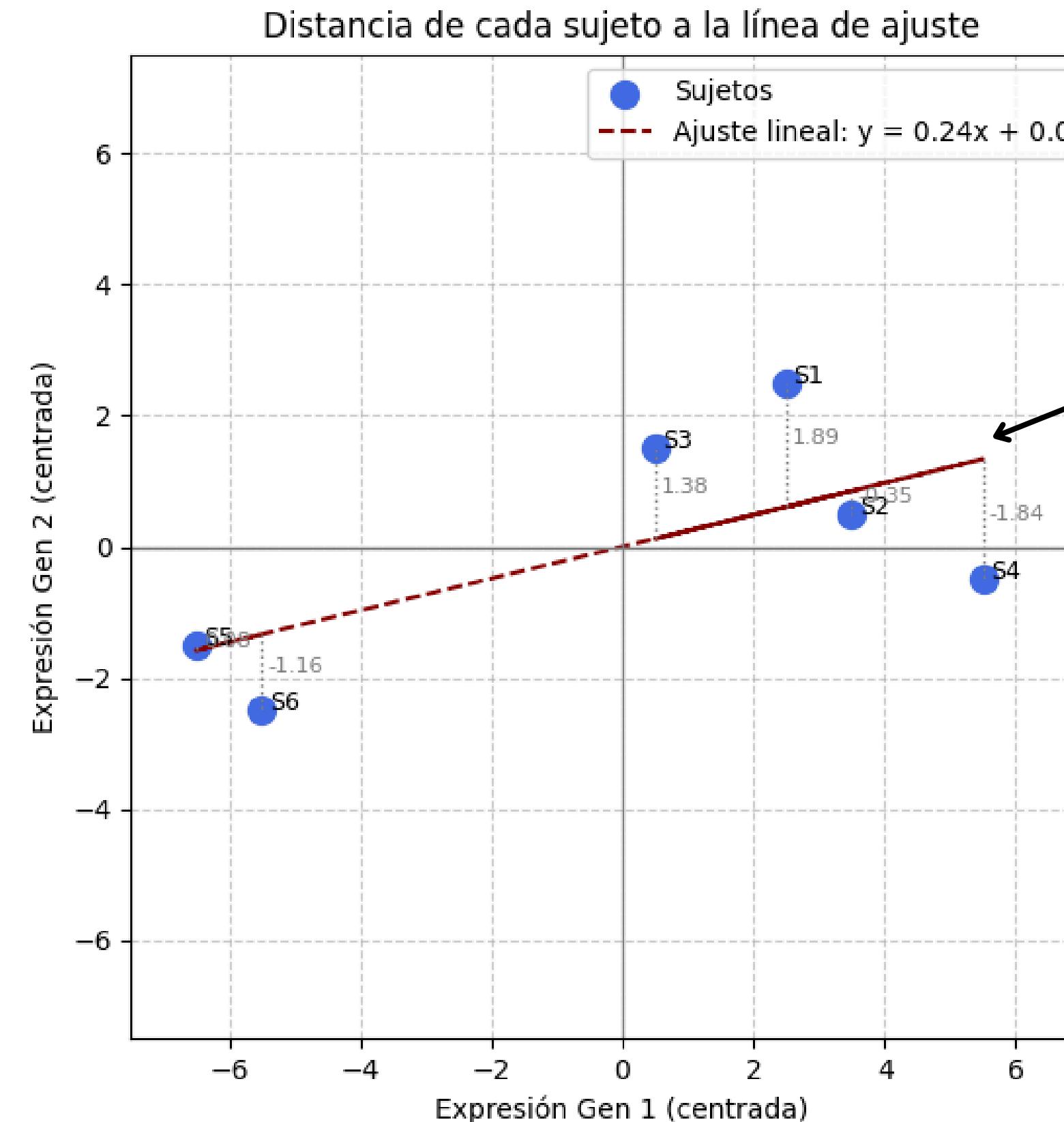
Expresión tridimensional de los genes 1-3



1. Centrar los datos — centrar los datos significa restar la media de cada gen, de modo que el punto (0,0) represente el promedio de las expresiones.



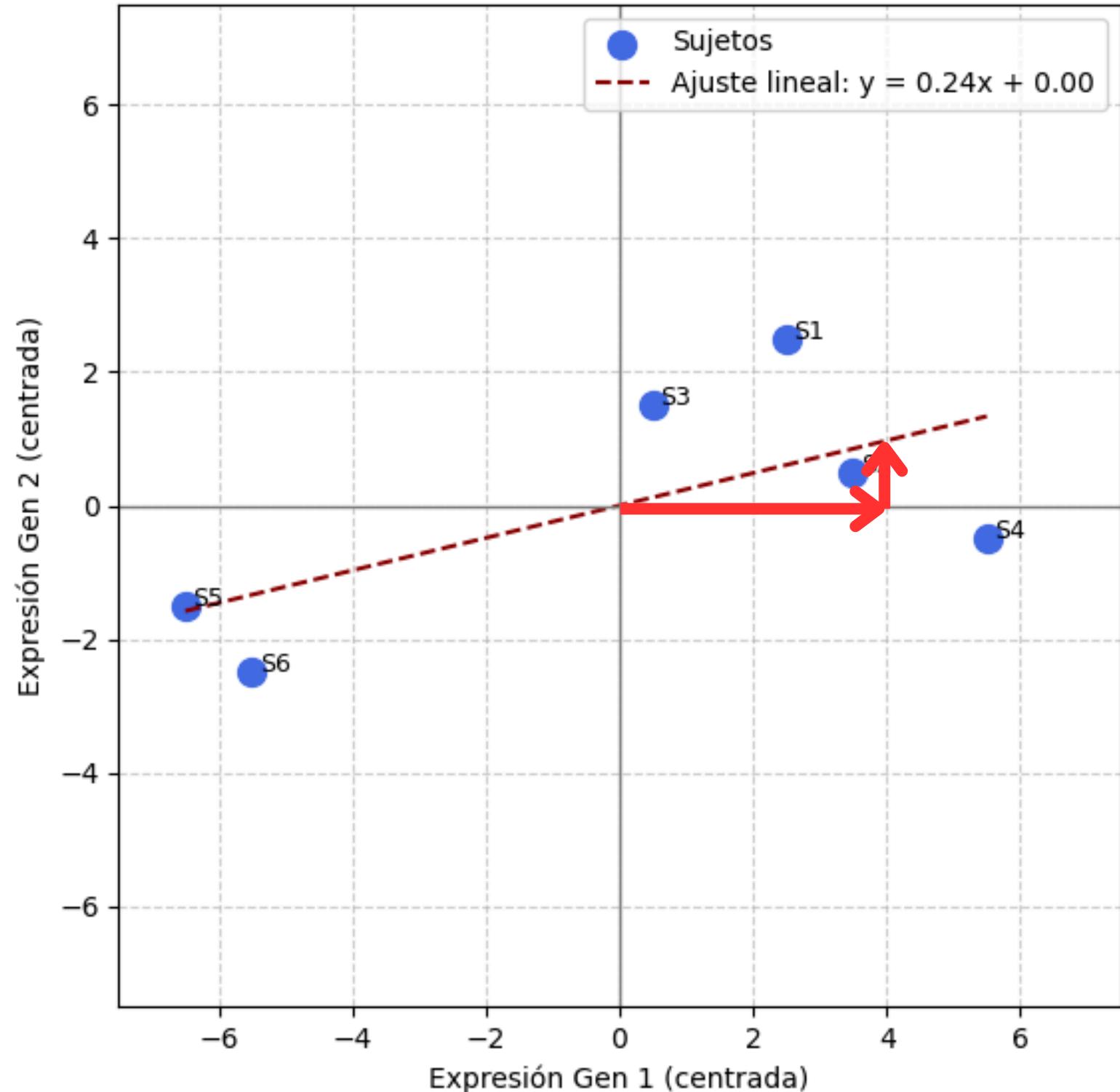
2. Ajustar una línea de tal manera que se minimice la distancia a la línea



Esto se llama PC1!

PC1 tiene una pendiente de 0.25

Relación entre Gen 1 y Gen 2 (centrados) con ajuste lineal



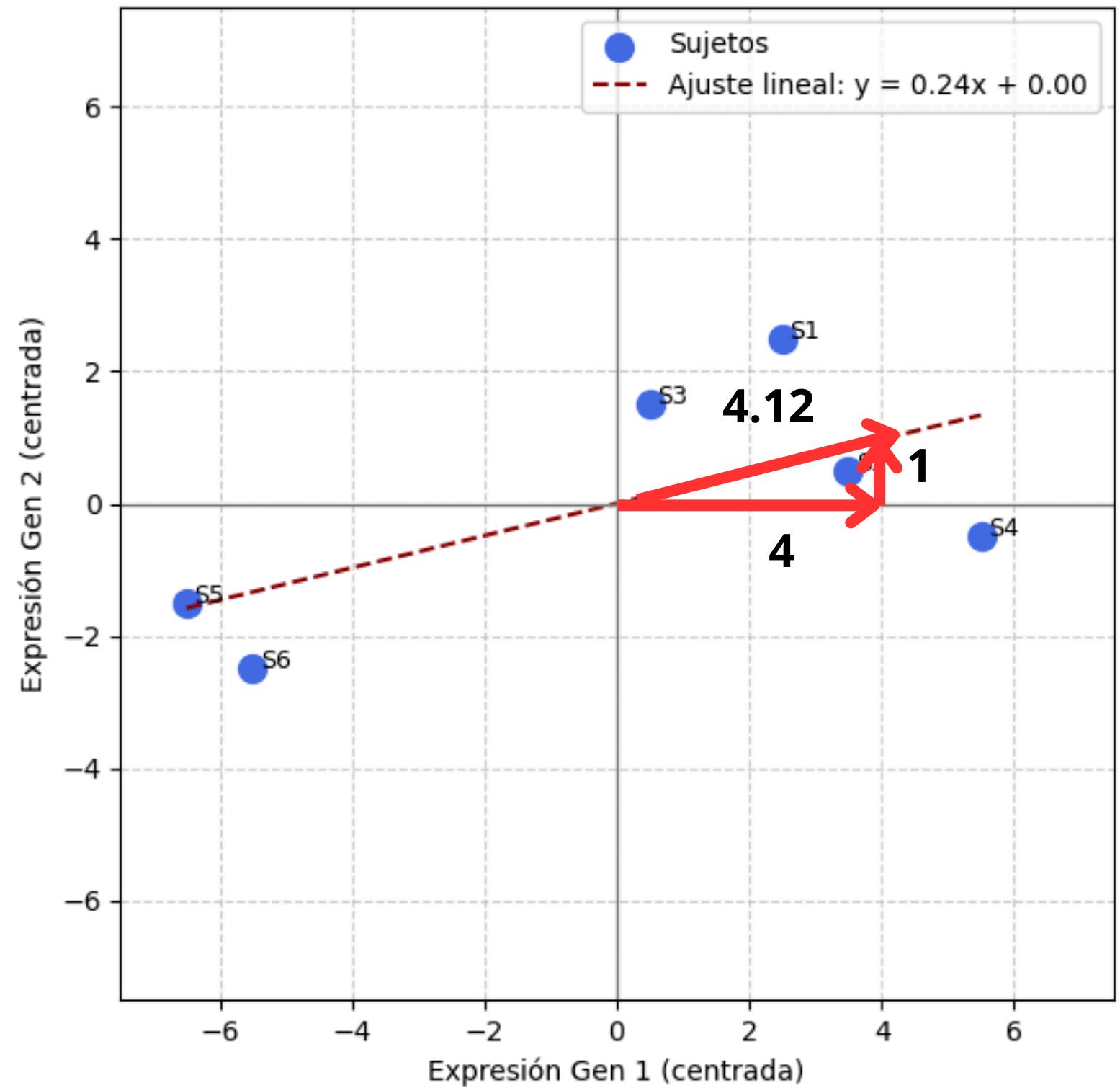
Según el teorema de Pitágoras, podemos conocer la proporción que cada eje «contribuye» a la línea.

Vemos que los datos están más distribuidos a lo largo del eje X que del eje Y.

Para hacer este PC1, necesitamos 4 partes del Gen 1 y 1 parte del Gen 2

¿Cuál gen es más importante para describir cómo está distribuida la data?

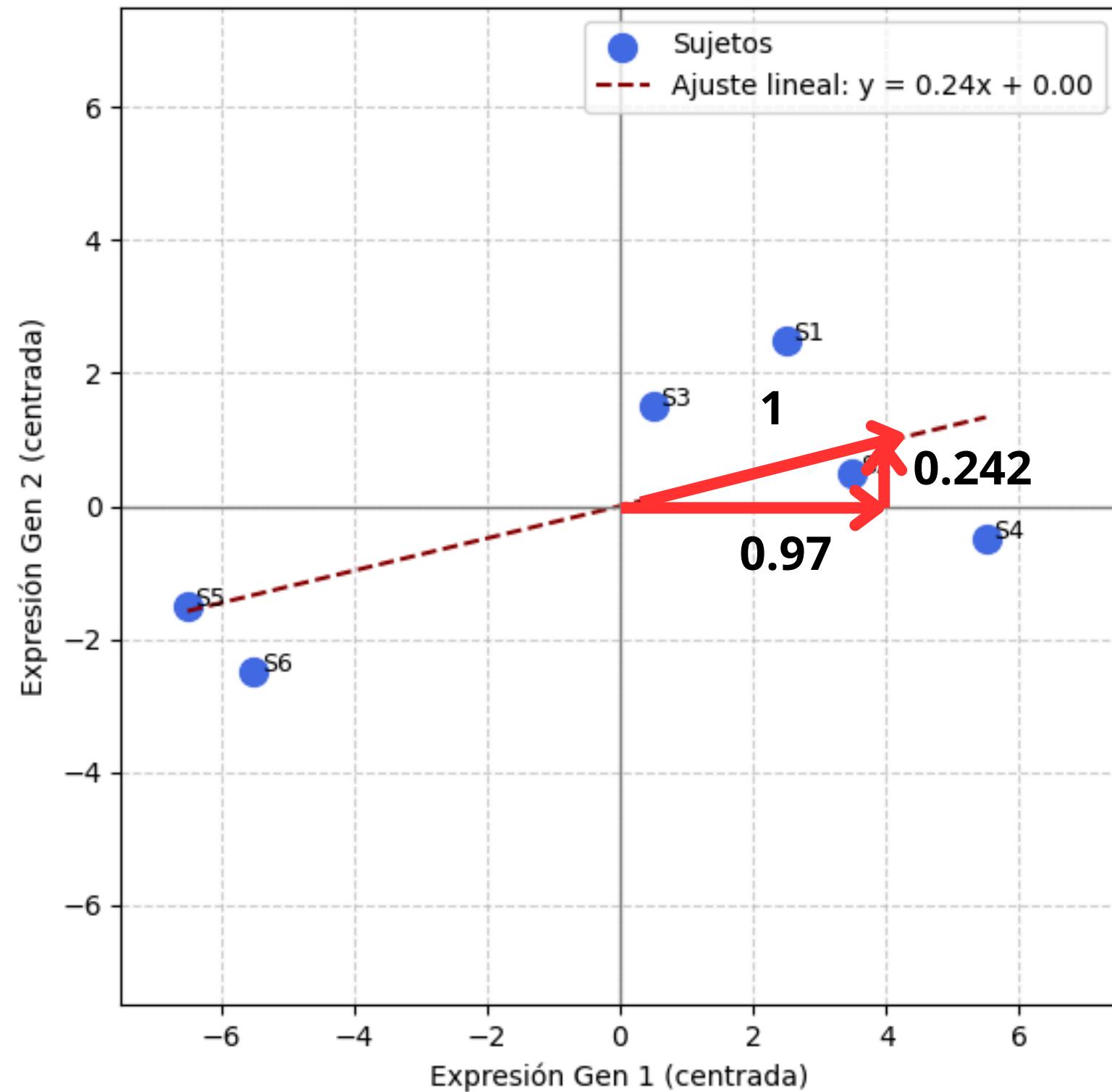
Relación entre Gen 1 y Gen 2 (centrados) con ajuste lineal



A esto le llamamos una
⭐ Combinación lineal ⭐

Esta línea tiene un tamaño de 4.12,
pero si lo escalamos (lo dividimos todo
entre 4.12) tendremos el

Relación entre Gen 1 y Gen 2 (centrados) con ajuste lineal



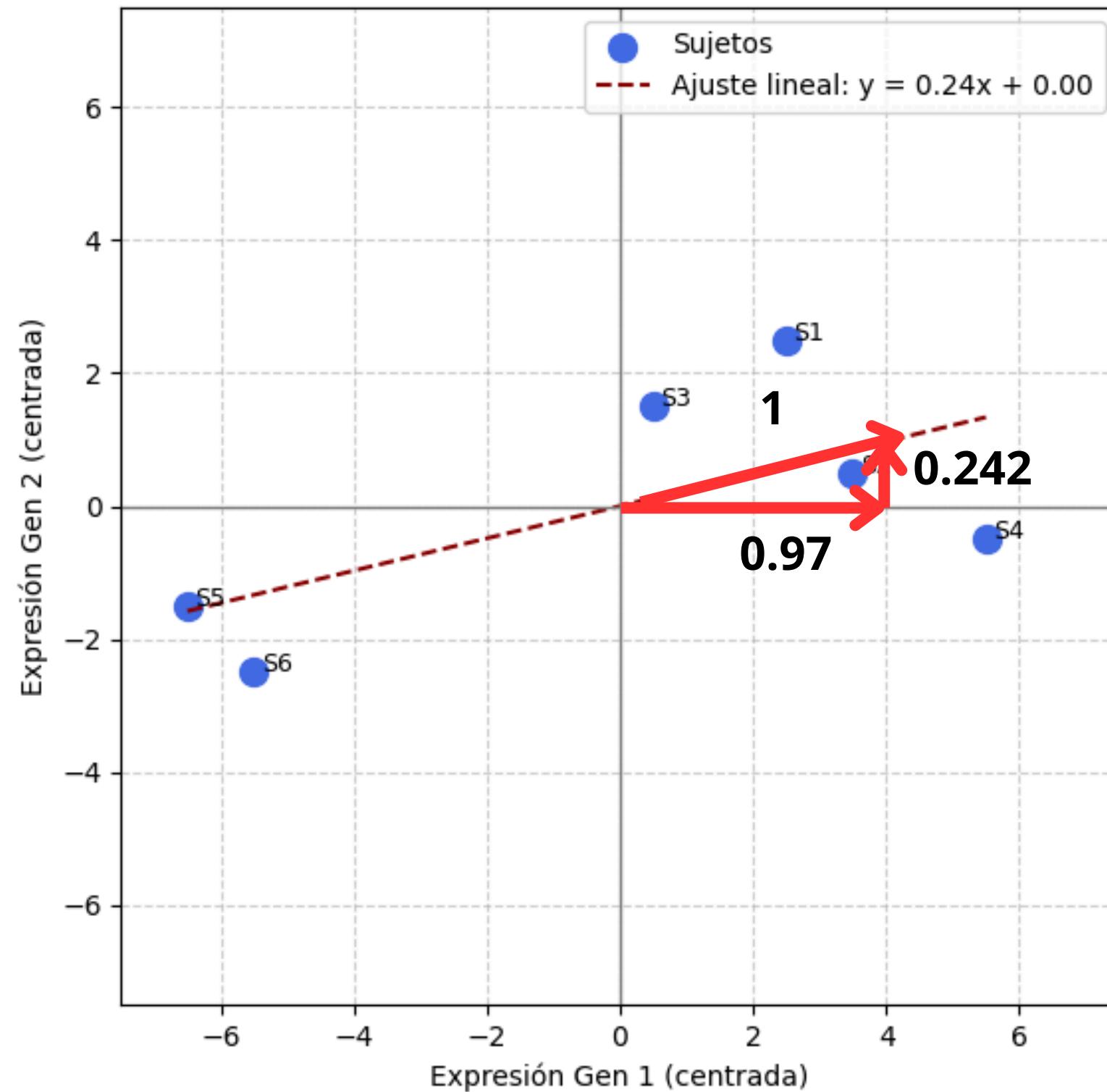
A esto le llamamos una
⭐ Combinación lineal ⭐

Esta línea tiene un tamaño de 4.12,
pero si lo escalamos (lo dividimos todo
entre 4.12) tendremos el

⭐ Vector singular o
Eigenvector ⭐

Para hacer el PC1 tenemos que
mezclar 0.97 partes del Gen 1 y 0.242
partes del Gen 2

Relación entre Gen 1 y Gen 2 (centrados) con ajuste lineal



Para hacer el PC1 tenemos que mezclar 0.97 partes del Gen 1 y 0.242 partes del Gen 2

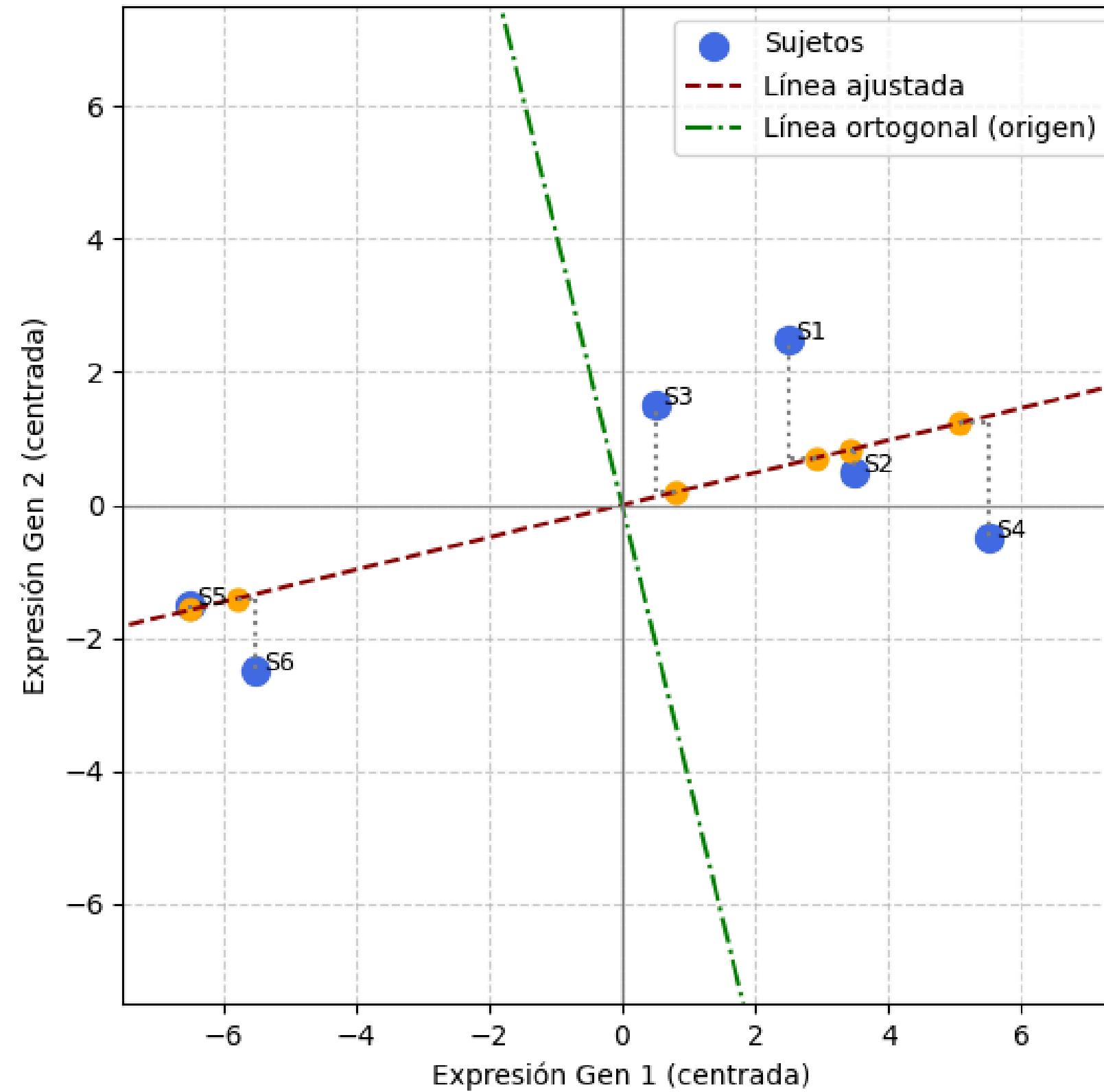
☀️Estos son los “Loading Scores”☀️

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

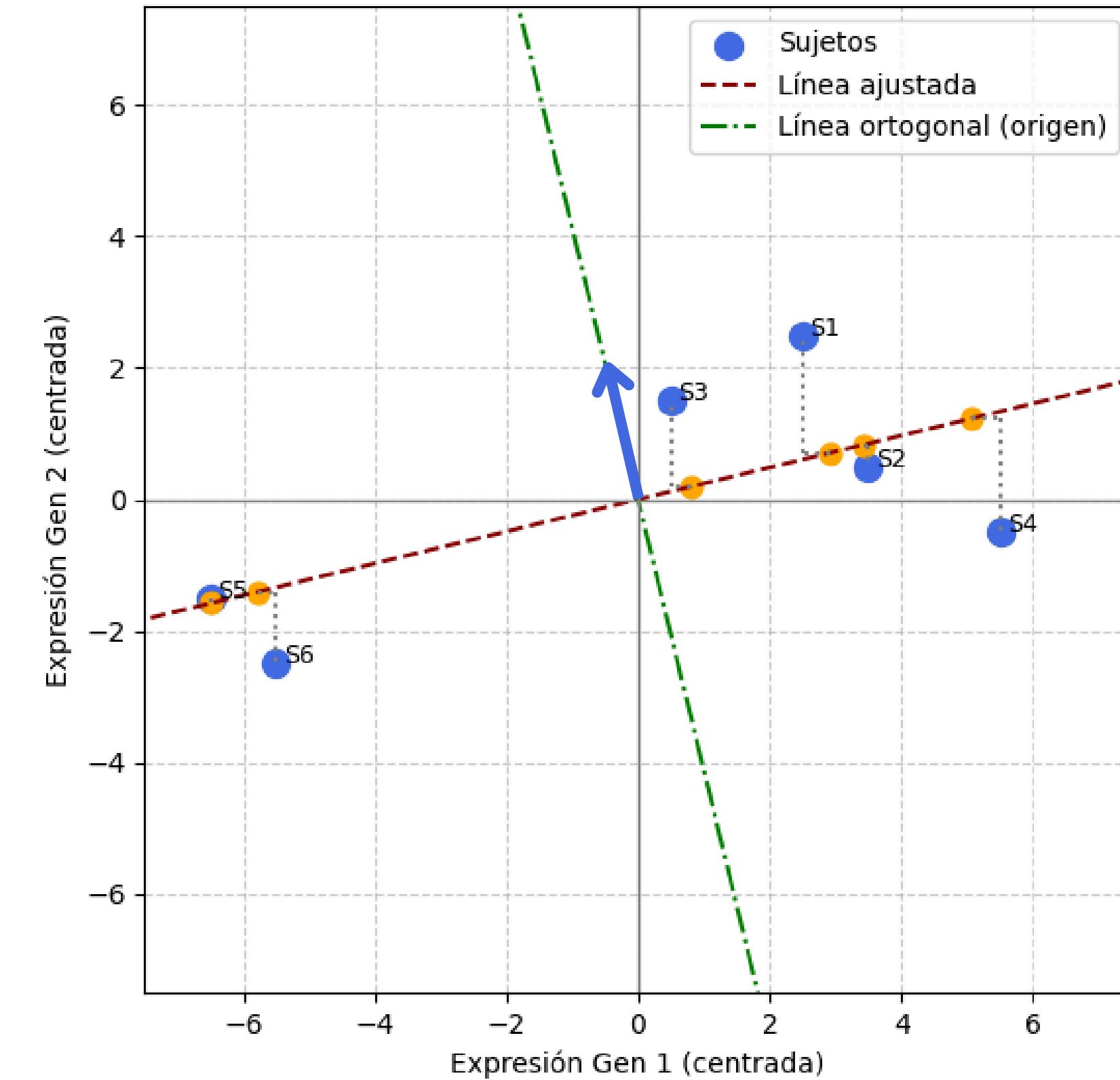
$$\sqrt{\text{SS(distances for PC1)}} = \text{Singular Value for PC1}$$

Y el PC2?

Triángulos y línea ortogonal a la línea ajustada

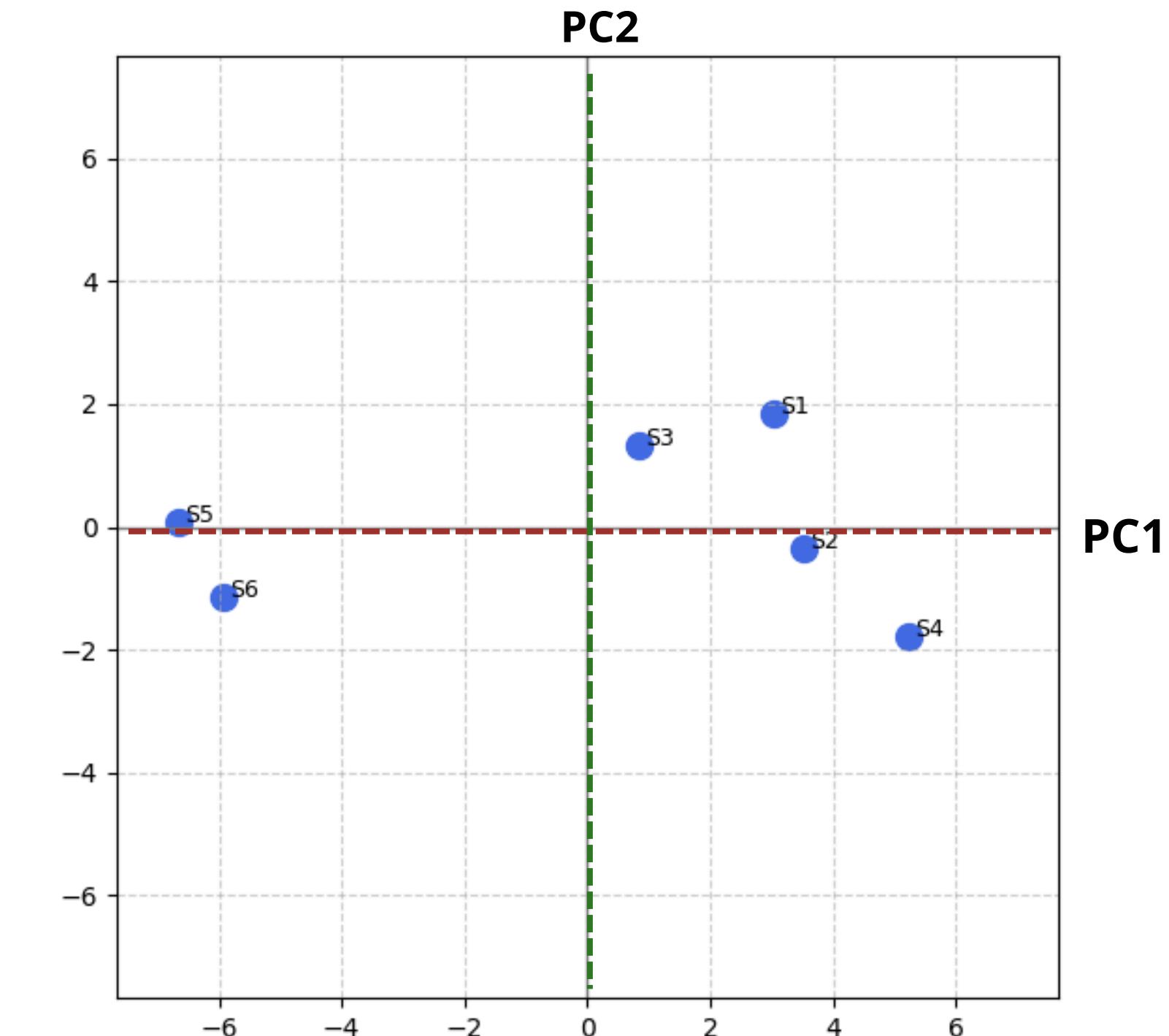
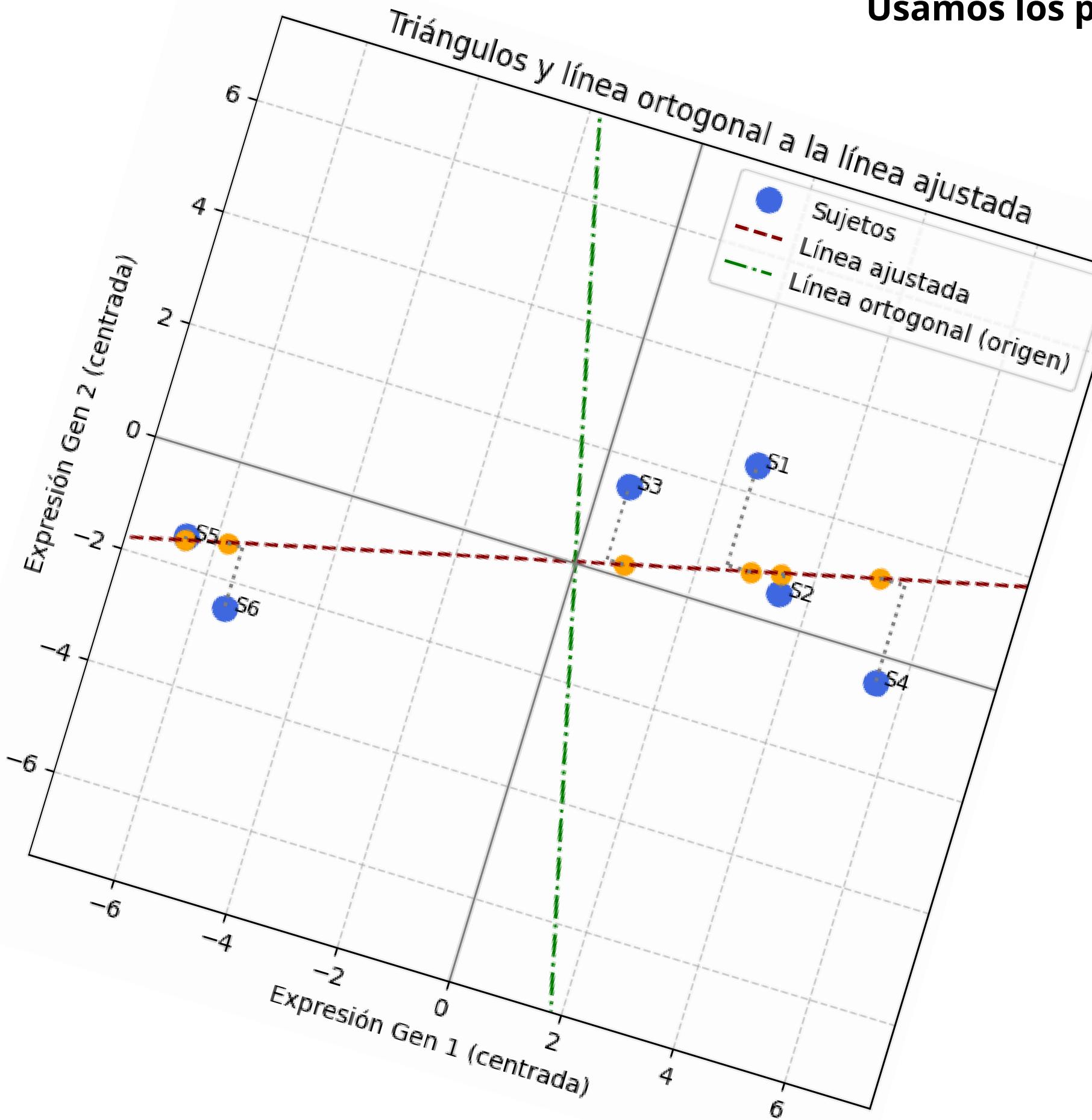


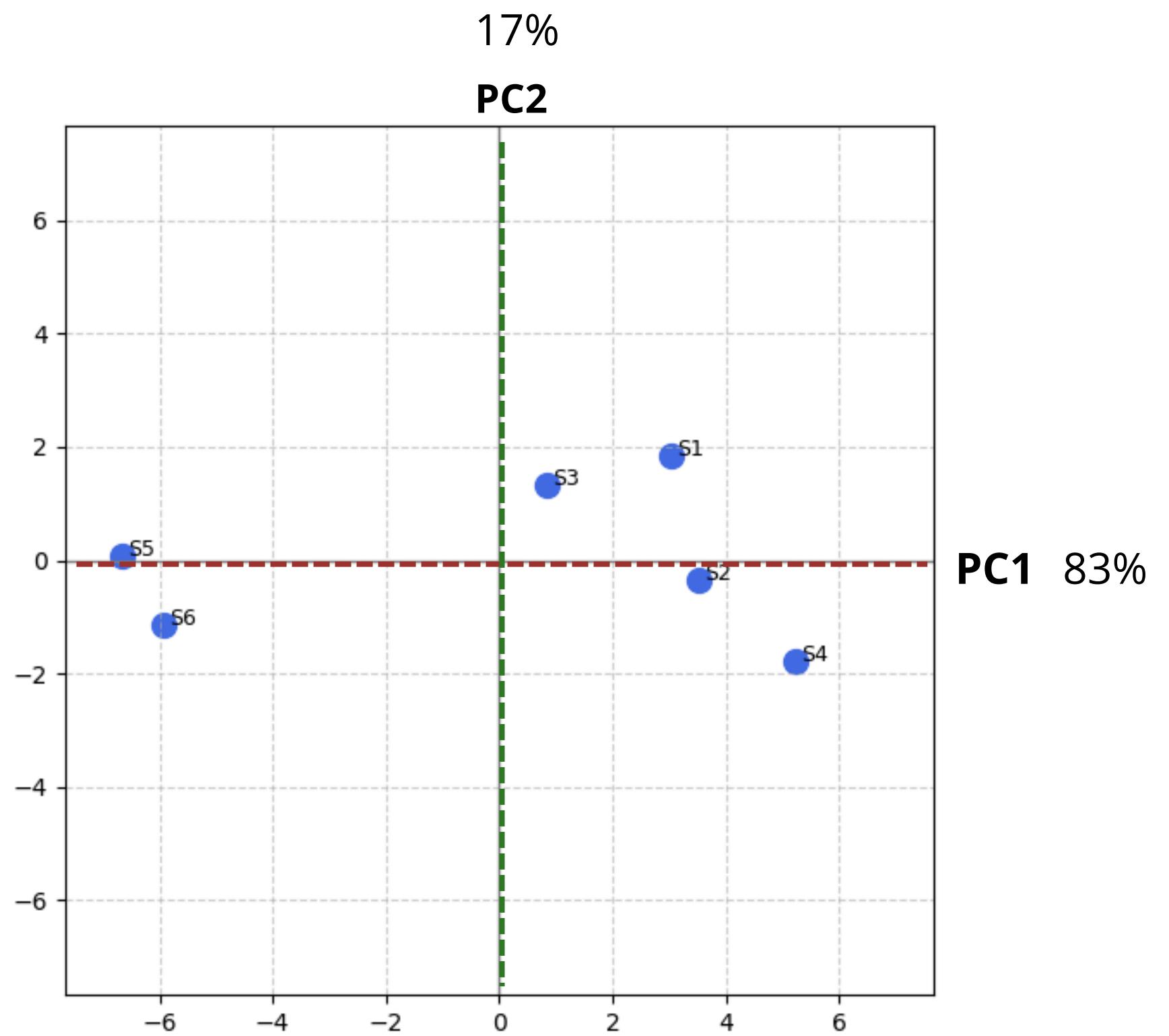
Triángulos y línea ortogonal a la línea ajustada



Y luego?

Usamos los puntos proyectados para encontrar donde van los puntos en el PCA, con nuestros nuevos ejes





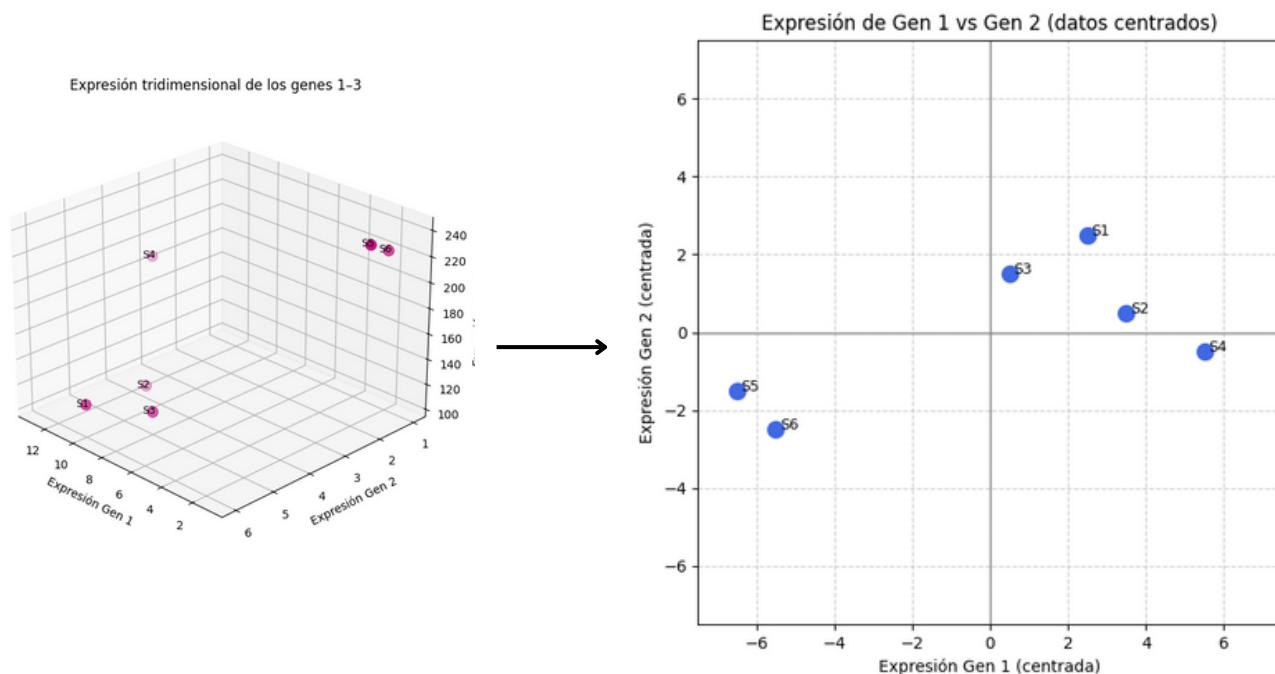
Una última cosa...

Imaginemos que la variación del PC1 es =15
Y que la variación del PC2 es = 3

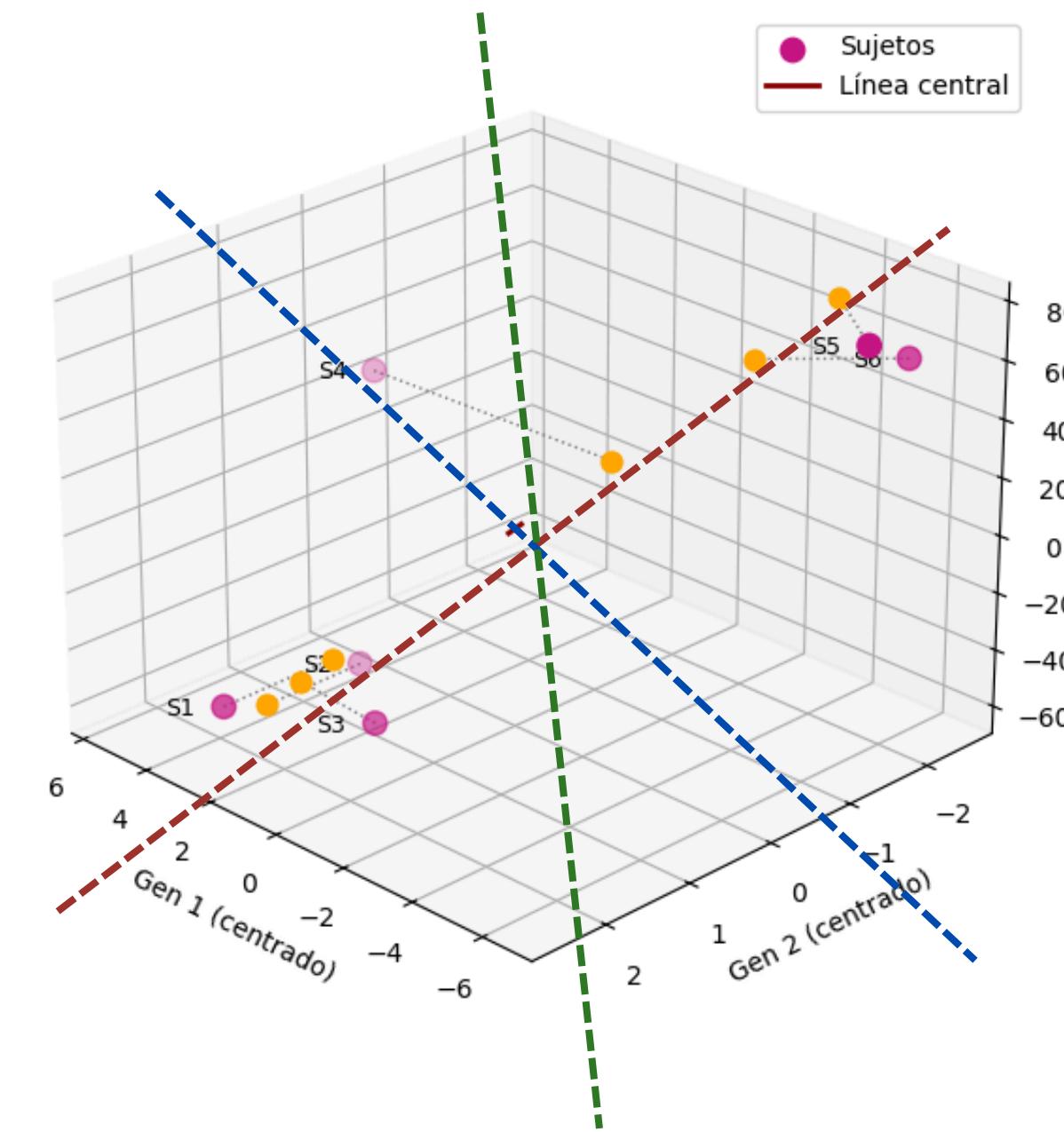
Saquemos porcentajes...

Pero tenemos mas de 2 Genes ...

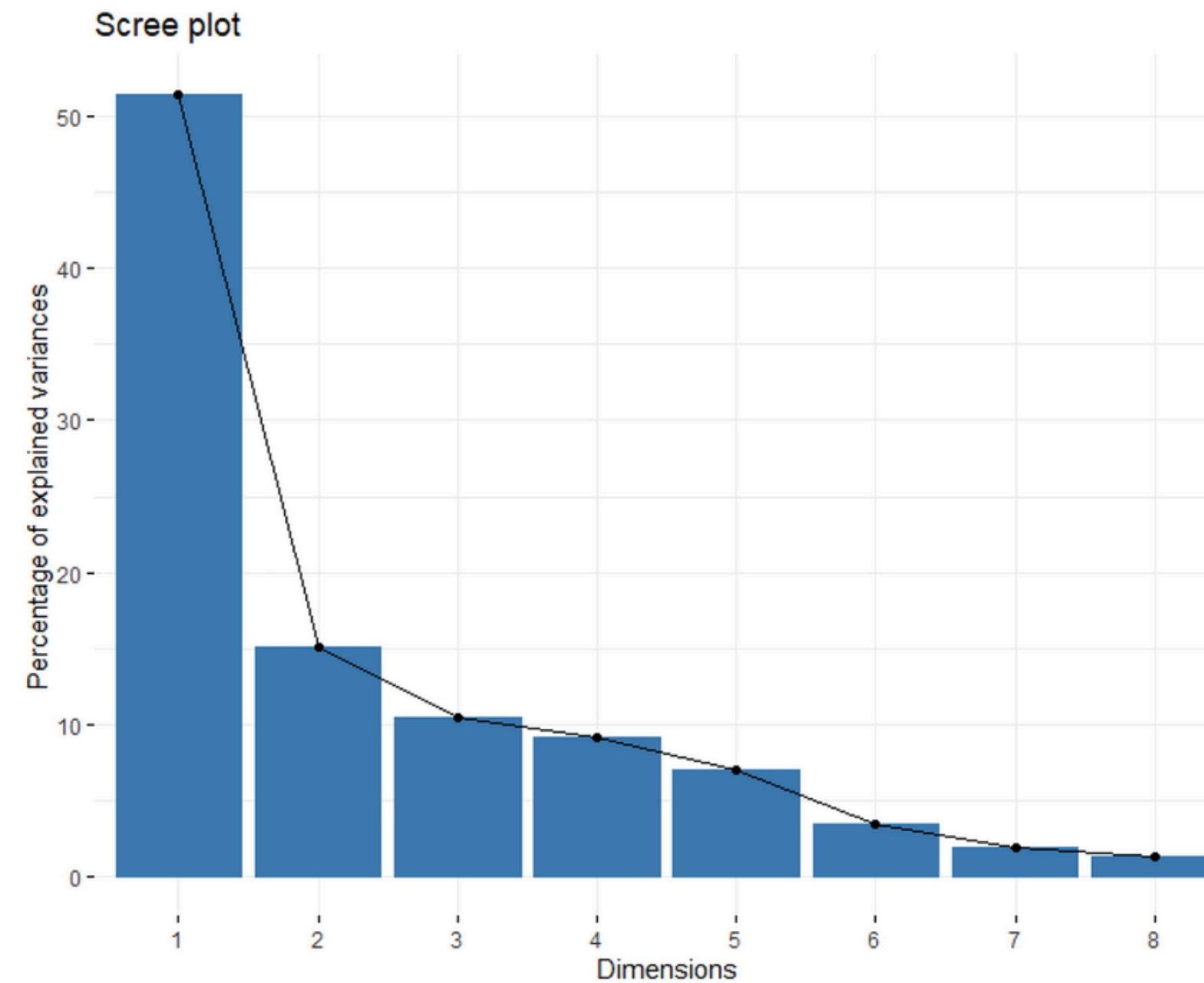
1.Centramos



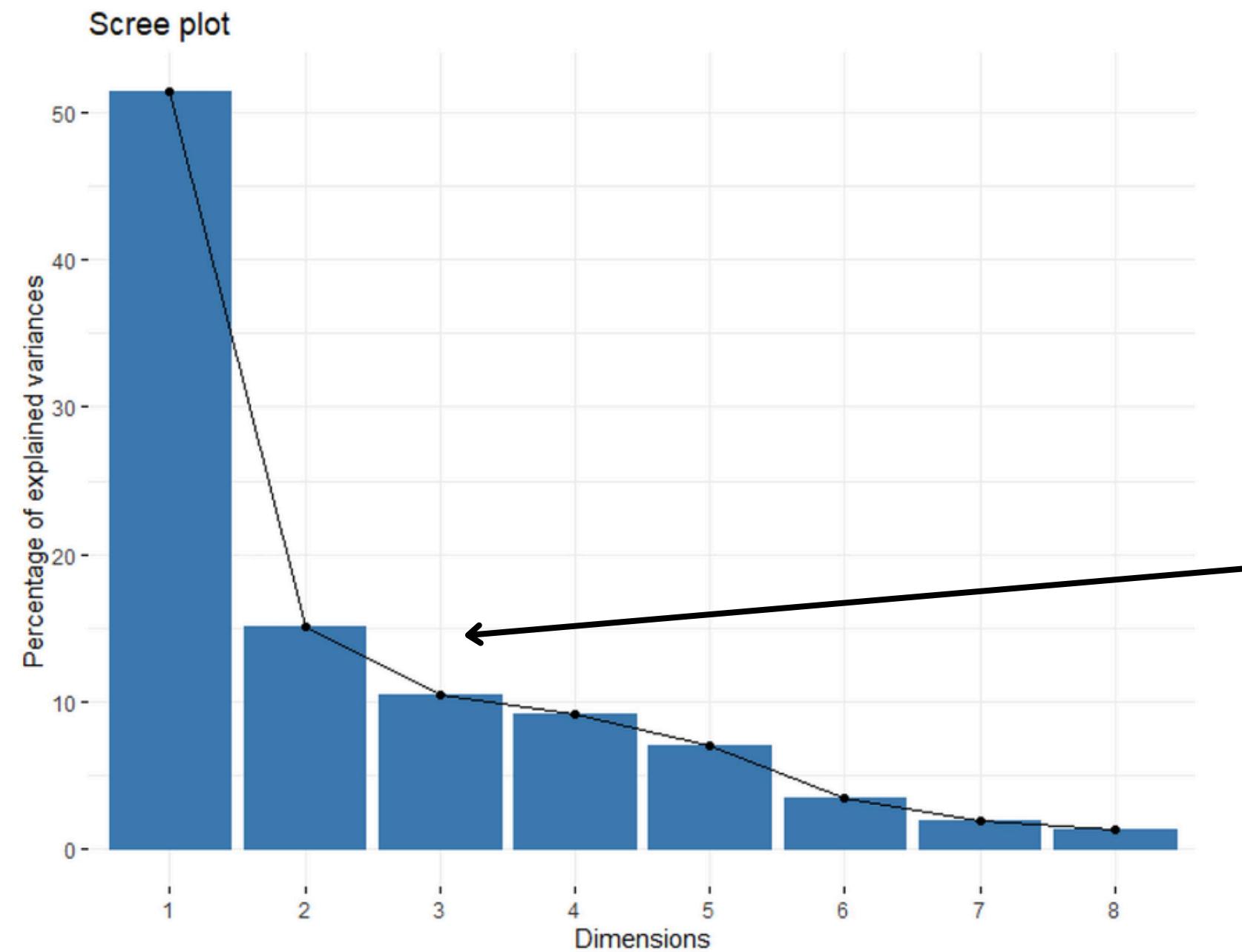
2.Encontramos la linea que mejor ajusta al origen



El PCA transforma un conjunto de datos en un conjunto de componentes **ortogonales** —los PCs— que capturan la máxima varianza en los datos.



Al realizar un PCA, obtendremos N componentes principales, donde N es igual a la dimensionalidad de nuestros datos originales.



**El PCA es la estructura de la covarianza de los vectores de expresión de genes.
Cosas que covarian juntas van en el mismo PC.**

¿Por qué/para qué usamos el PCA?

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...
...
...
GeneM	25	0	.	0

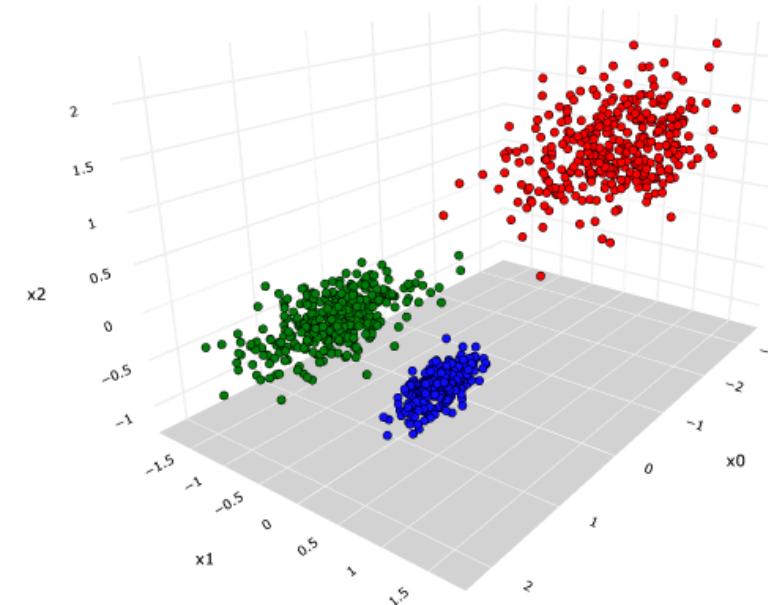
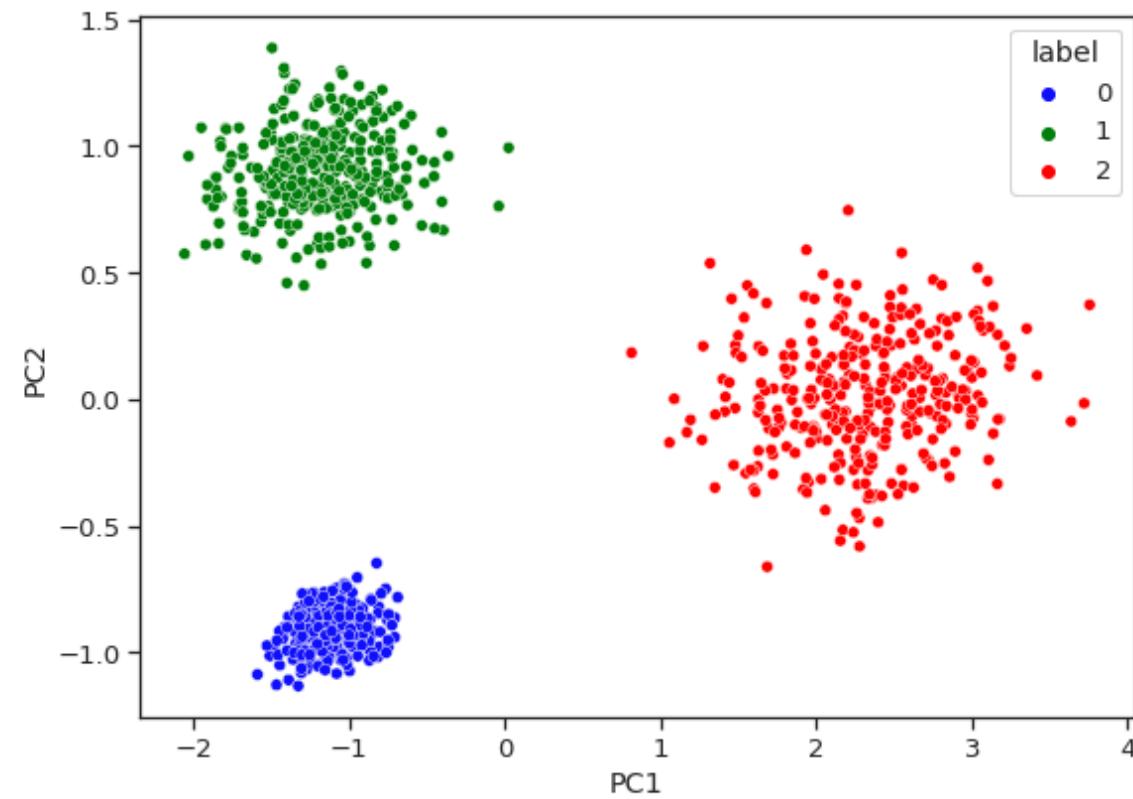
Tenemos una enorme cantidad de datos

No todas estas variables serán críticas/importantes.

Tenemos que aprender qué es lo que le sobra

M genes x N muestras

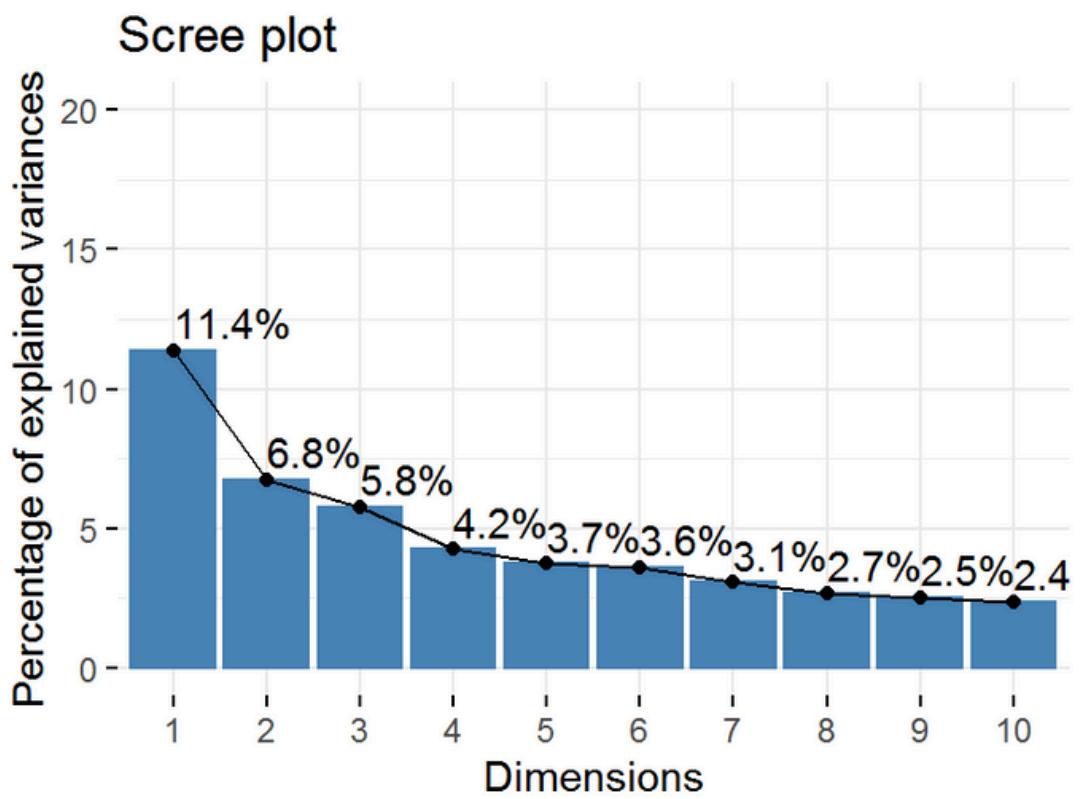
¿Por qué/para qué usamos el PCA?



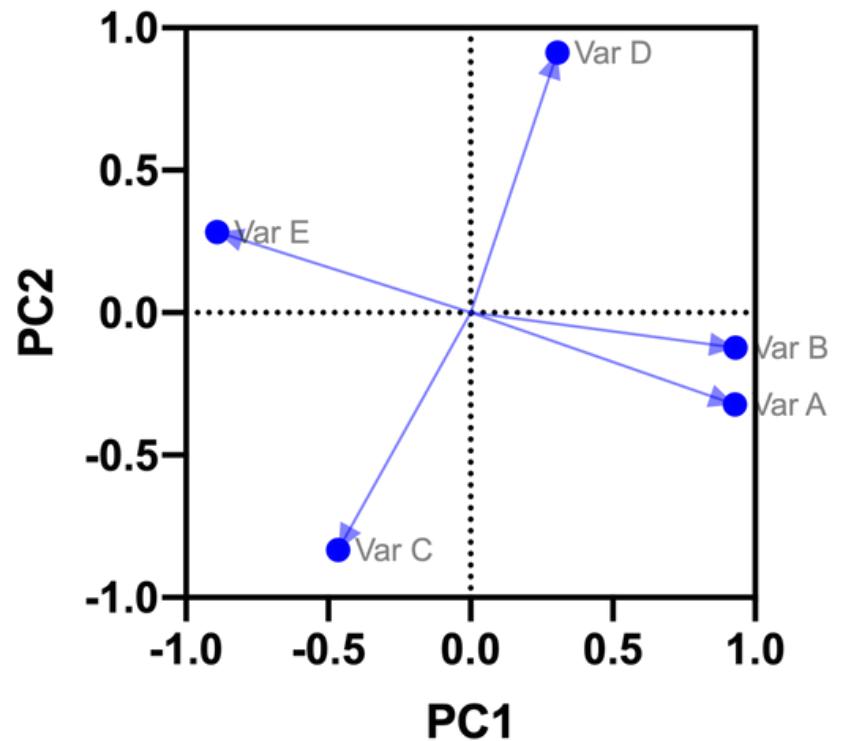
- Análisis exploratorio de datos
- Preprocesamiento de datos
- “Denoising”
- Feature Selection
- Encontrar similitudes y diferencias entre muestras.
- Visualización

Plots del PCA

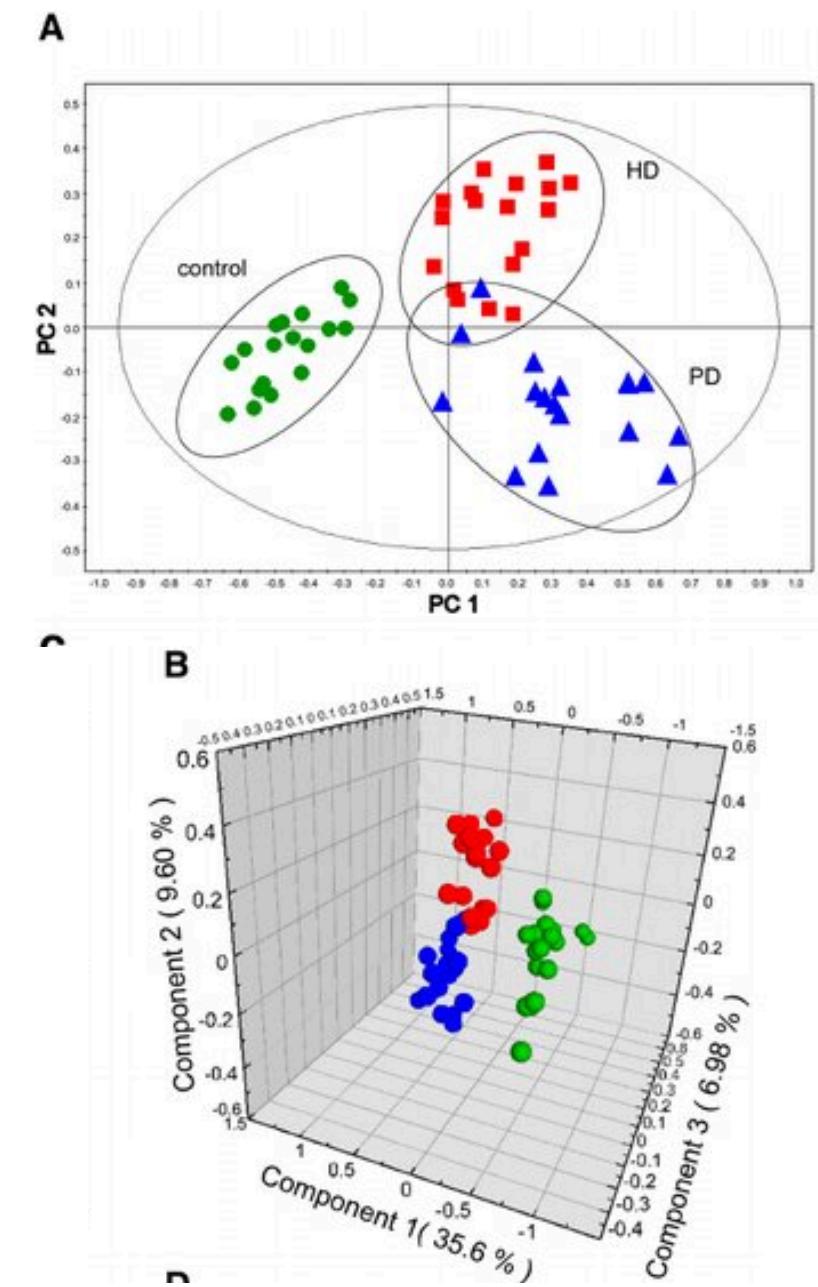
Scree plot (gráfico de varianza explicada)



Loading plot



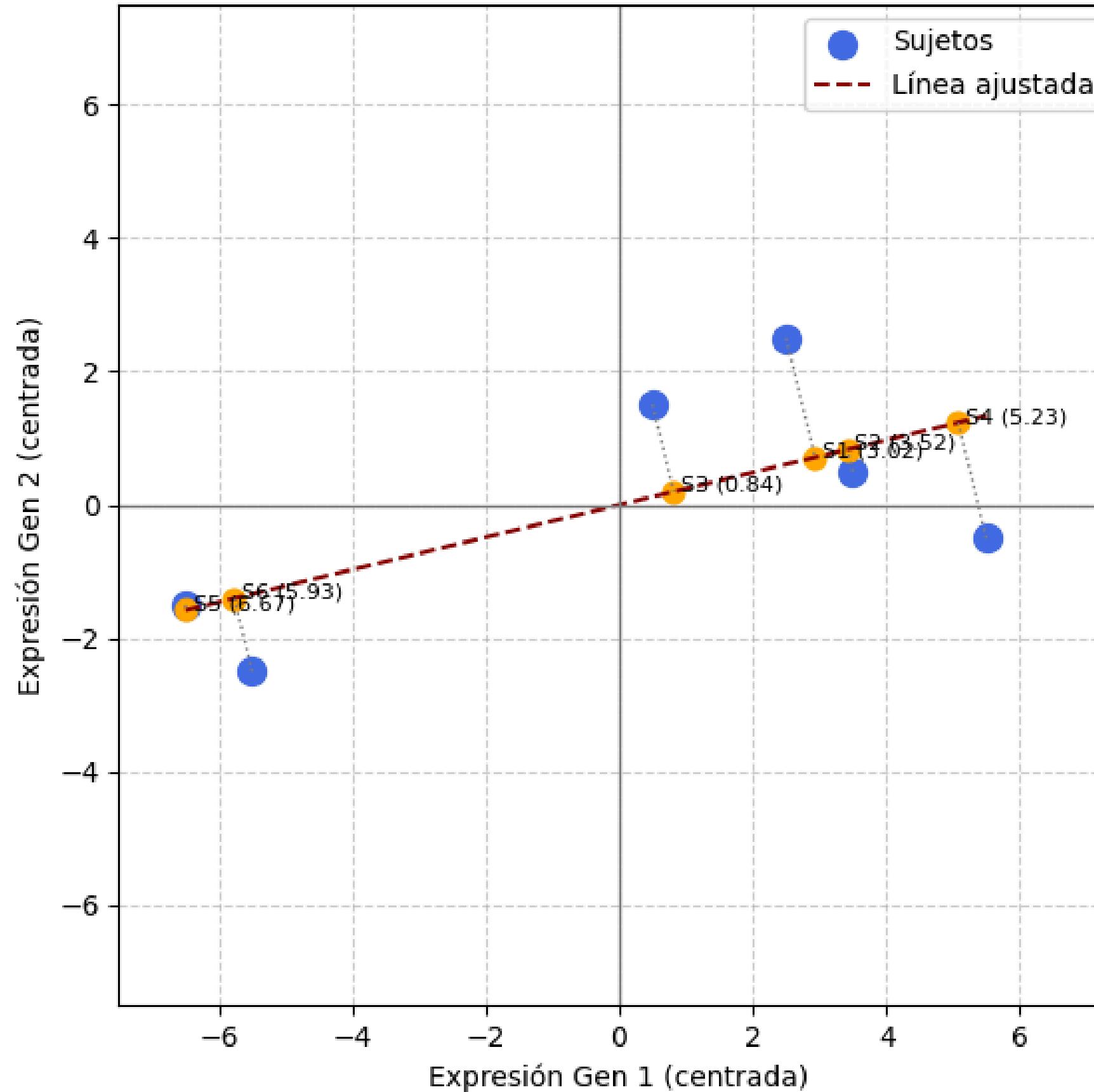
Score plot
(proyección de muestras)



¡Vamos a programar!



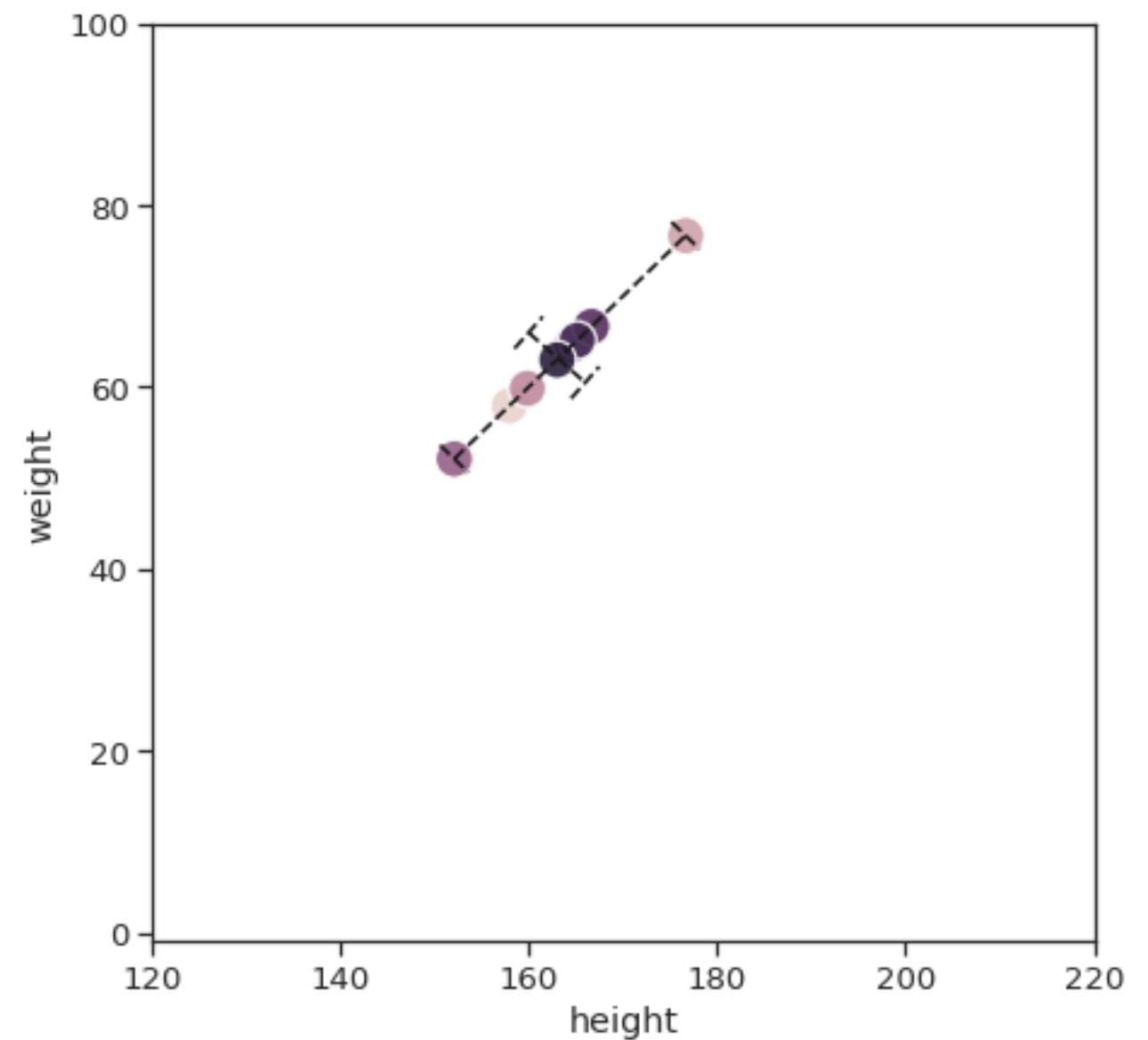
Distancia al origen de las proyecciones sobre la línea ajustada



Distancia al origen de las proyecciones sobre la línea ajustada

Sujeto	Distancia al origen	Distancia al origen ²
S1	3.019	9.114
S2	3.519	1.237
S3	840	706
S4	5.227	2.732
S5	6.670	4.449
S6	5.935	3.521

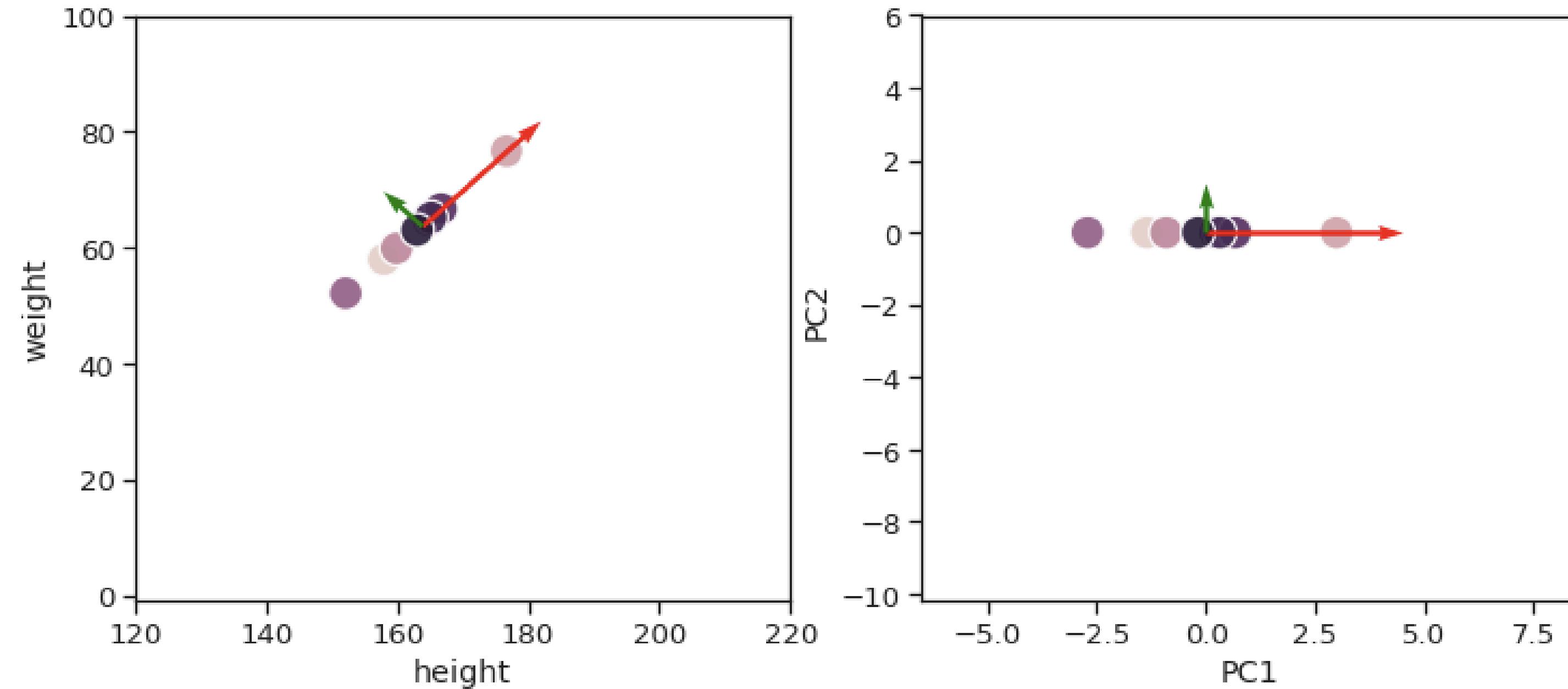
Los elevamos al cuadrado para que los valores positivos no cancelen a los negativos



Vemos que la mayor variación no se encuentra en el eje x ni en el eje y, sino en una línea diagonal que los atraviesa.

La segunda mayor variación sería una línea de 90 grados que corta la primera.

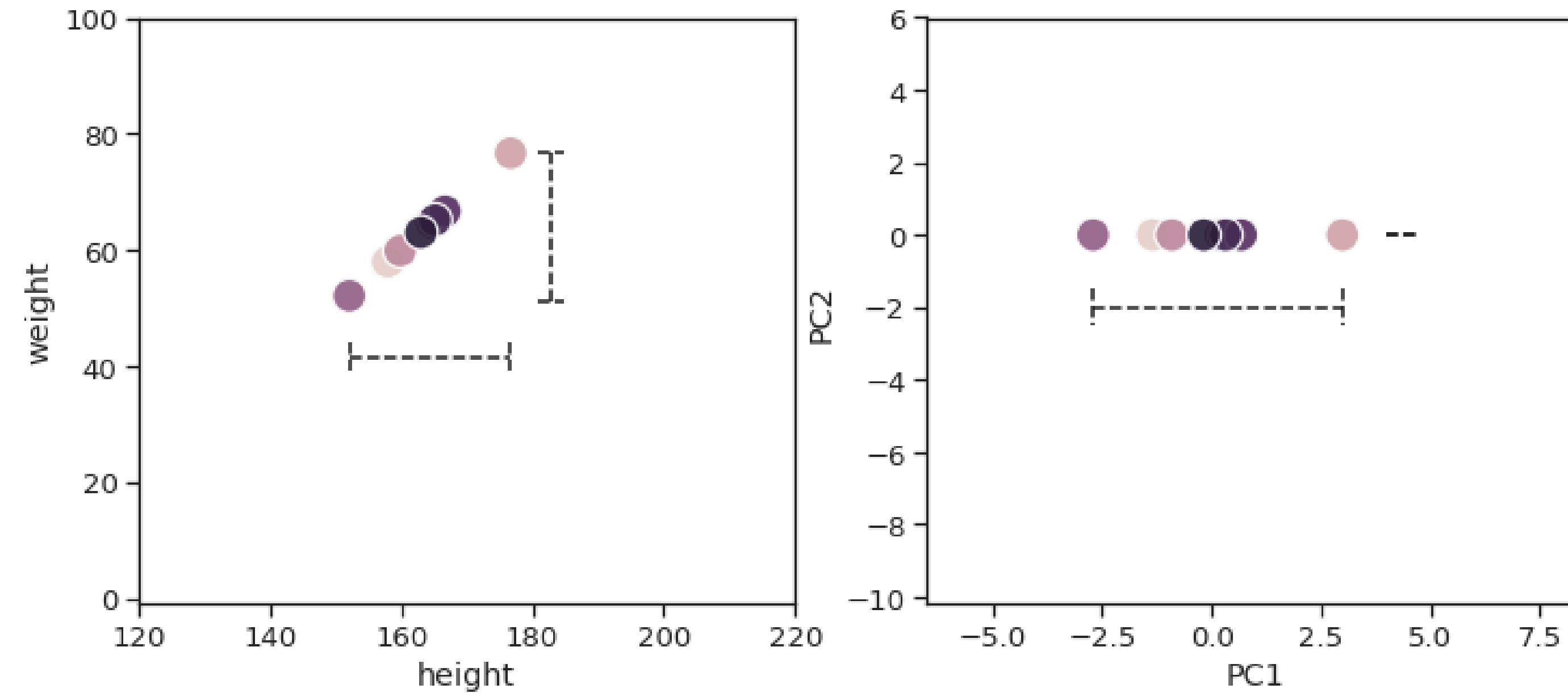
Feature	Variance
Height	1.11
Weight	1.11
TOTAL	2.22



Estas dos nuevas variables se denominan primer componente principal (PC1) y segundo componente principal (PC2).

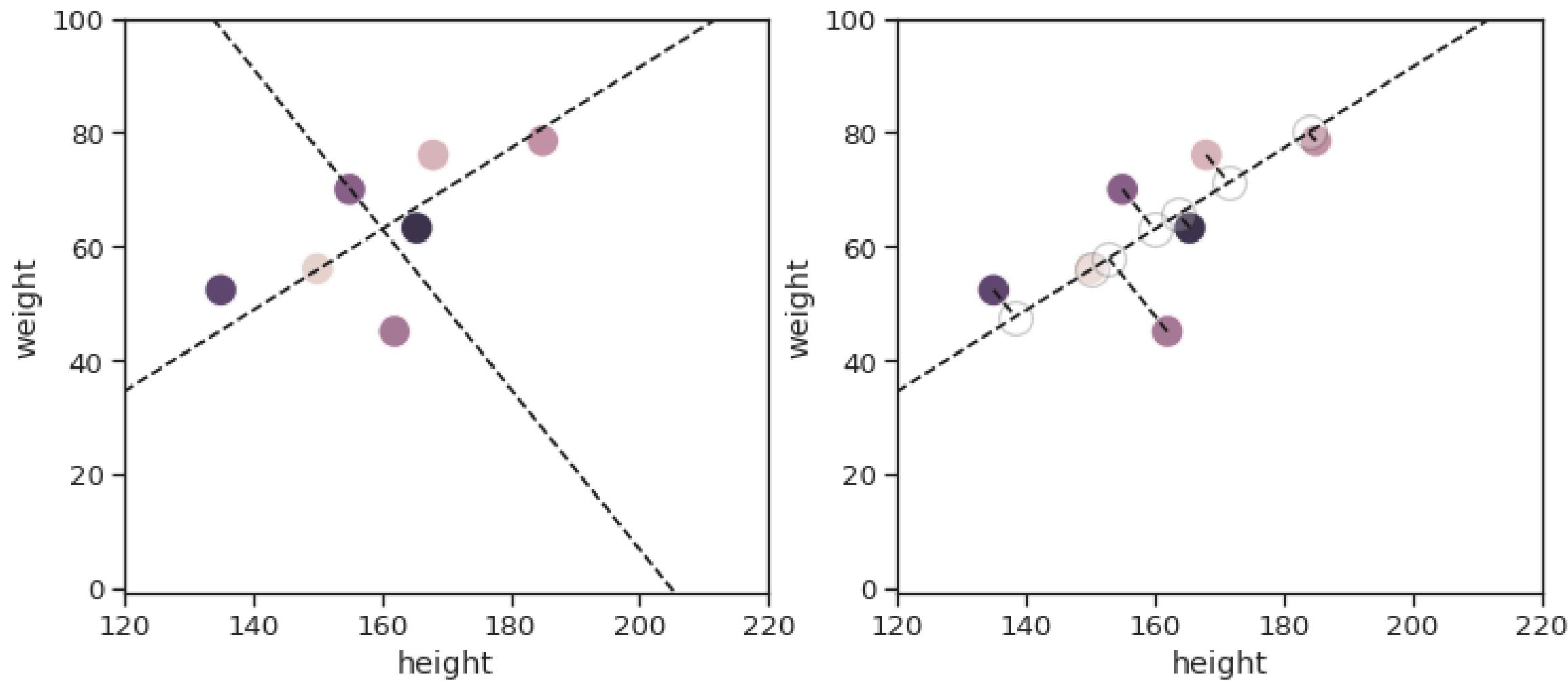
En lugar de utilizar la altura y el peso en los dos ejes, podemos utilizar PC1 y PC2 respectivamente.

Podría ser un 30 % de altura y un 70 % de peso, o un 87,2 % de altura y un 13,8 % de peso, o cualquier otra combinación dependiendo de los datos que tengamos.



Feature	Variance	Feature	Variance
Height	1.11	PC1	2.22
Weight	1.11	PC2	0.00
TOTAL	2.22	TOTAL	2.22

El PC1 por sí solo puede capturar la varianza total de la altura y el peso combinados.



Sin embargo, supongamos que hemos decidido conservar solo el primer componente principal, tendríamos que proyectar todos nuestros puntos de datos sobre el primer componente principal, ya que ya no tenemos el eje y :(.

Cuando se trata de datos reales, la mayoría de las veces no obtendremos un componente principal que capture el 100 % de las variaciones.

- https://cran.r-project.org/web/packages/LearnPCA/vignettes/Vig_03_Step_By_Step_PCA.pdf
- https://www.nature.com/articles/nmeth.4346?error=cookies_not_supported&code=e624924bab4c-456f-9809-3ee1c29da29f
- <https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d/#0226>
- <https://www.youtube.com/watch?v=FgakZw6K1QQ&t=61s>