THE BRIDGE

No Supervisado

Algoritmo PCA

Algunos conceptos matemáticos

Matriz de Covarianza (I)

La covarianza es el valor que refleja en qué cuantía dos variables aleatorias varían de forma conjunta respecto a sus medias.

Nos permite saber cómo se comporta una variable en función de lo que hace otra variable. Es decir, cuando X sube ¿Cómo se comporta Y? Así pues, la covarianza puede tomar los siguientes valores:

Covarianza (X,Y) es menor que cero cuando "X" sube e "Y" baja. Hay una relación negativa.

Covarianza (X,Y) es mayor que cero cuando "X" sube e "Y" sube. Hay una relación positiva.

Si cuando "X" varia, "Y" no varía… ¿cuánto valdrá la covarianza?

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

En la variación está la información

Matriz de Covarianza (II)

La matriz de covarianza de un dataset es la matriz en la que se recogen las covarianzas de las features tomadas de dos en dos.

Es como la matriz de correlación que hemos calculado tantas veces pero en vez de recoger la correlación recoge la covarianza.

Por ejemplo para un dataset con dos features (b0 y b1):

$$\mathbf{V}(\mathbf{b}) = \begin{bmatrix} var(b_0) & cov(b_0, b_1) \\ cov(b_0, b_1) & V(b_1) \end{bmatrix}$$

Autovalores y autovectores

En álgebra lineal, los vectores propios, eigenvectores o autovectores de una matriz son los vectores no nulos que, cuando son mutiplicados por la matriz, dan lugar a un múltiplo escalar de sí mismos, con lo que no cambian su dirección. Este escalar lambda recibe el nombre valor propio, autovalor o valor característico.

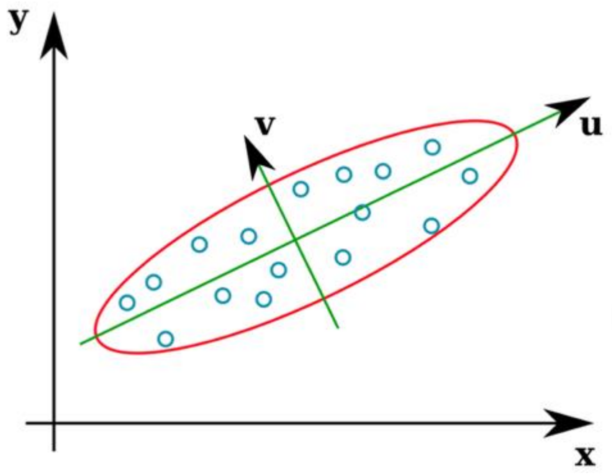
La matriz sobre la que se calculan los autovalores y autovectores tiene que ser cuadrada.

Hay tantos autovalores como dimensiones tenga la matriz. Se pueden repetir.

Todos los autovectores son perpendiculares entre sí

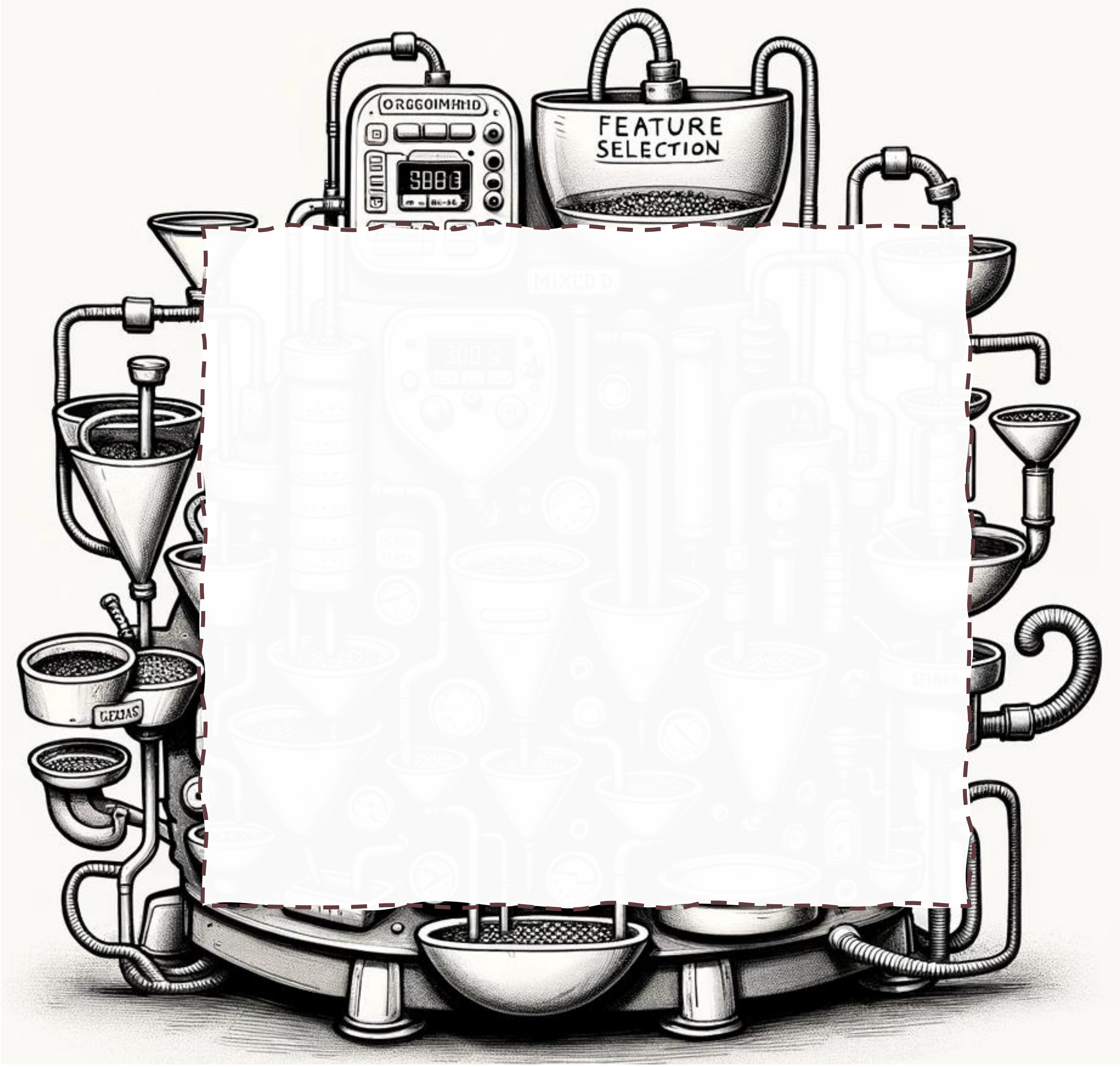
La longitud del autovector es 1 y su autovalor representa el poder de cada autovector.

$$A.v = \lambda.v \quad , \quad v \neq 0_V$$



[Explicación detallada de cálculo de autovalores y autovetores]

PCA



Algoritmo de la PCA

0 Importante, SENSIBLE A LA ESCALA, hay que estandarizar

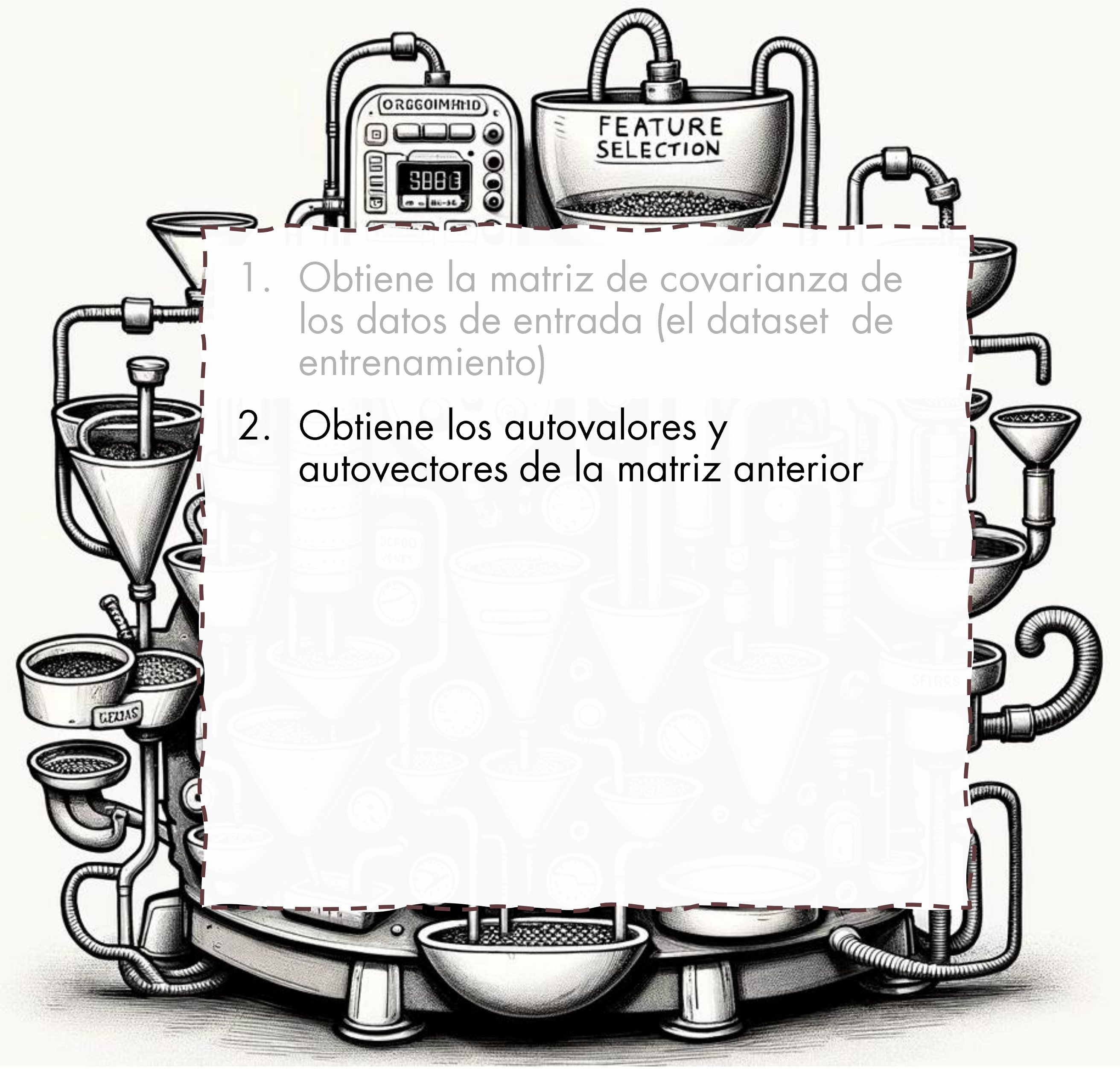
Algoritmo de la PCA

1. Obtiene la matriz de covarianza de los datos de entrada (el dataset de entrenamiento)

1

$$\mathbf{V}(\mathbf{b}) = \mathbf{V}\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{bmatrix} var(b_0) & cov(b_0, b_1) \\ cov(b_0, b_1) & V(b_1) \end{bmatrix}$$

Algoritmo de la PCA



1. Obtiene la matriz de covarianza de los datos de entrada (el dataset de entrenamiento)
2. Obtiene los autovalores y autovectores de la matriz anterior

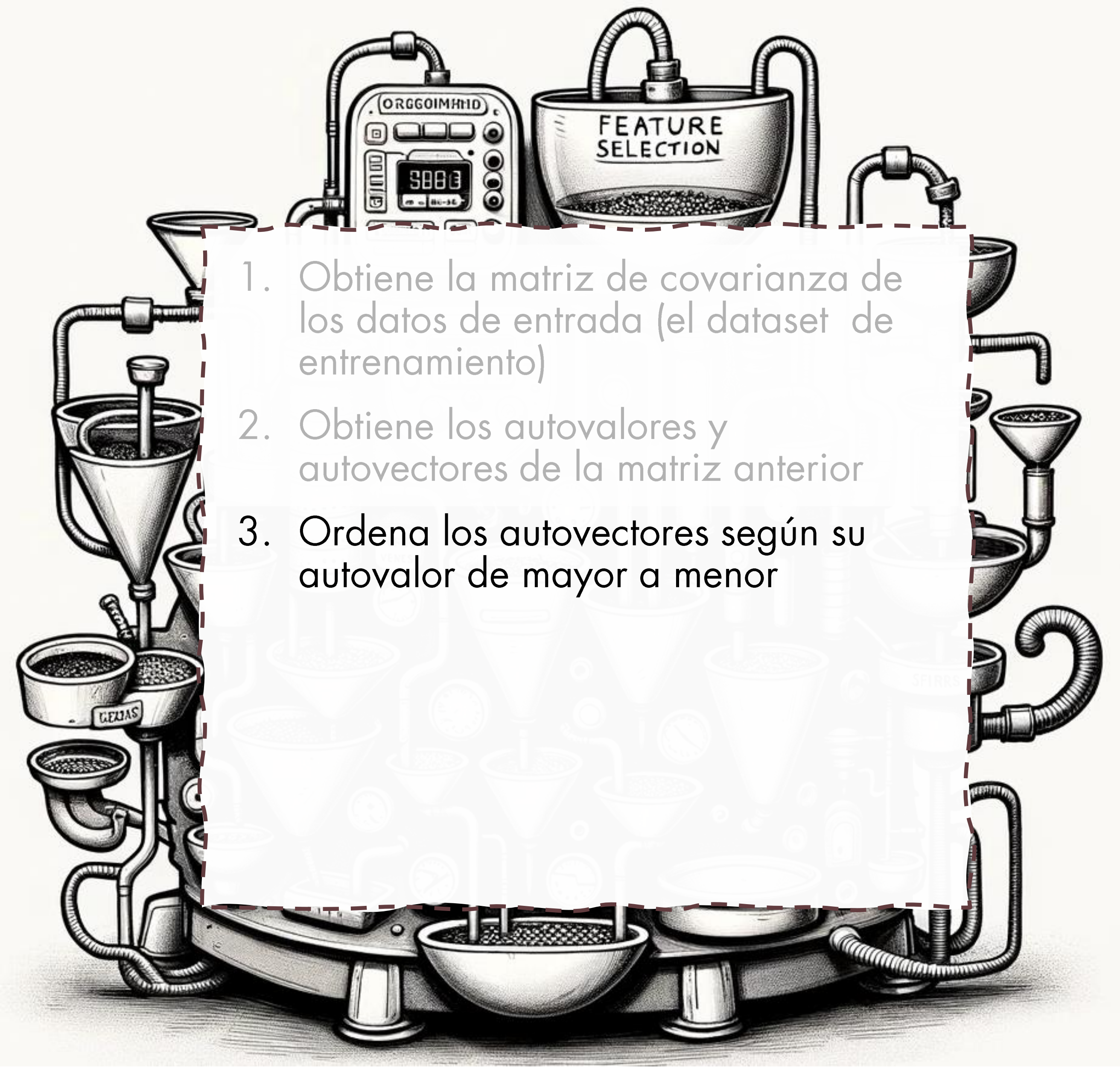
1

$$V(b) = V\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{bmatrix} var(b_0) & cov(b_0,b_1) \\ cov(b_0,b_1) & V(b_1) \end{bmatrix}$$

2

$$A.v = \lambda.v \quad , \quad v \neq 0_V$$

Algoritmo de la PCA



1. Obtiene la matriz de covarianza de los datos de entrada (el dataset de entrenamiento)
2. Obtiene los autovalores y autovectores de la matriz anterior
3. Ordena los autovectores según su autovalor de mayor a menor

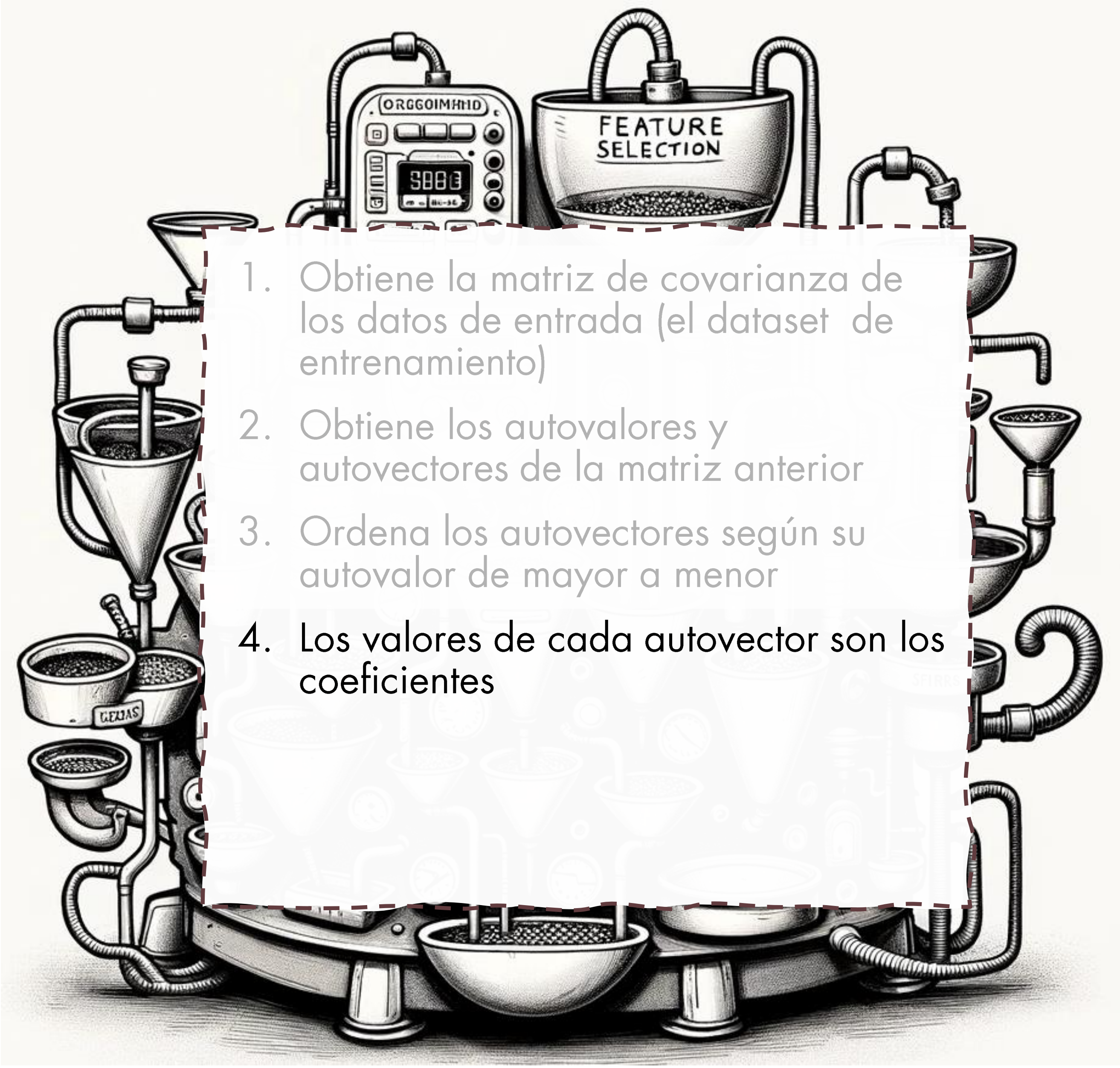
1

$$V(b) = V\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{bmatrix} var(b_0) & cov(b_0,b_1) \\ cov(b_0,b_1) & V(b_1) \end{bmatrix}$$

2,3

$$A.v = \lambda.v \quad , \quad v \neq 0_V$$

Algoritmo de la PCA



1. Obtiene la matriz de covarianza de los datos de entrada (el dataset de entrenamiento)
2. Obtiene los autovalores y autovectores de la matriz anterior
3. Ordena los autovectores según su autovalor de mayor a menor
4. Los valores de cada autovector son los coeficientes

1

$$V(b) = V\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{bmatrix} var(b_0) & cov(b_0,b_1) \\ cov(b_0,b_1) & V(b_1) \end{bmatrix}$$

2,3

$$A.v = \lambda.v \quad , \quad v \neq 0_V$$

4

$$CP_1 = c_{10}x_0 + c_{11}x_1 + c_{12}x_2 + c_{13}x_3 + ... + c_{1m}x_m$$
$$CP_2 = c_{20}x_0 + c_{21}x_1 + c_{22}x_2 + c_{23}x_3 + ... + c_{2m}x_m$$
$$CP_3 = c_{30}x_0 + c_{31}x_1 + c_{32}x_2 + c_{33}x_3 + ... + c_{3m}x_m$$
$$...$$
$$CP_m = c_{m0}x_0 + c_{m1}x_1 + c_{m2}x_2 + c_{m3}x_3 + ... + c_{mm}x_m$$

Algoritmo de la PCA

1. Obtiene la matriz de covarianza de los datos de entrada (el dataset de entrenamiento)
2. Obtiene los autovalores y autovectores de la matriz anterior
3. Ordena los autovectores según su autovalor de mayor a menor
4. Los valores de cada autovector son los coeficientes
5. Aplica los coeficientes a las features originales y obtiene el nuevo dataset

1

$$\mathrm{V(b)} = \mathrm{V}\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{bmatrix} var(b_0) & cov(b_0,b_1) \\ cov(b_0,b_1) & V(b_1) \end{bmatrix}$$

2,3

$$A.v = \lambda.v \quad , \quad v \neq 0_V$$

4

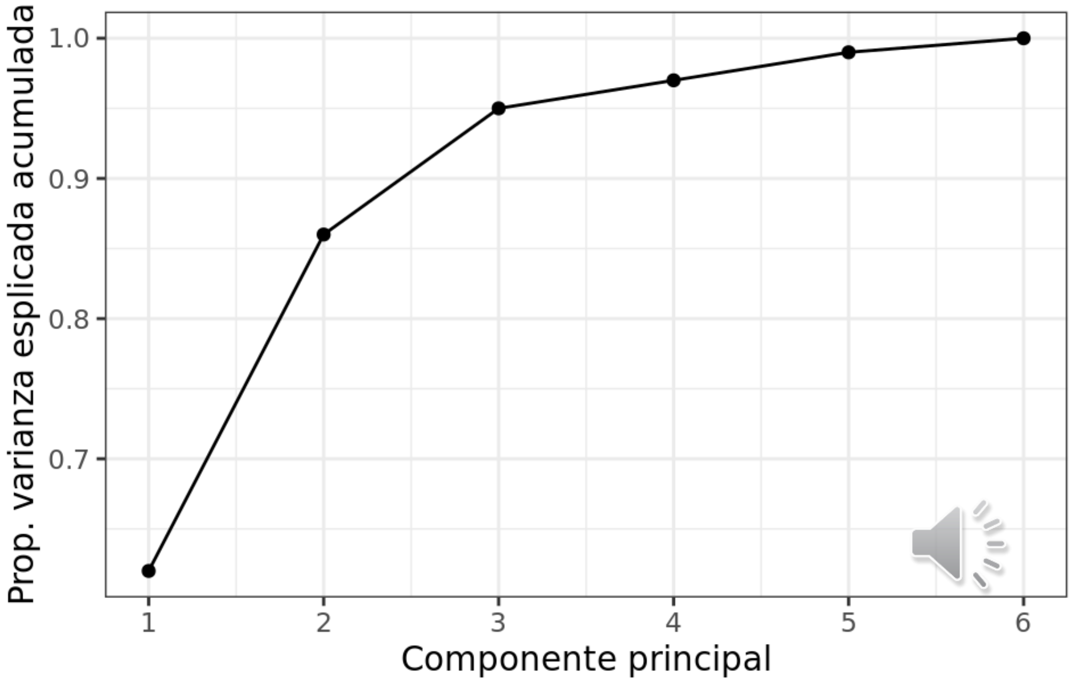
$$CP_1 = c_{10}x_0 + c_{11}x_1 + c_{12}x_2 + c_{13}x_3 + ... + c_{1m}x_m$$
$$CP_2 = c_{20}x_0 + c_{21}x_1 + c_{22}x_2 + c_{23}x_3 + ... + c_{2m}x_m$$
$$CP_3 = c_{30}x_0 + c_{31}x_1 + c_{32}x_2 + c_{33}x_3 + ... + c_{3m}x_m$$
$$...$$
$$CP_m = c_{m0}x_0 + c_{m1}x_1 + c_{m2}x_2 + c_{m3}x_3 + ... + c_{mm}x_m$$

5

$$newDataset^T = featureVector^T . dataset^T$$

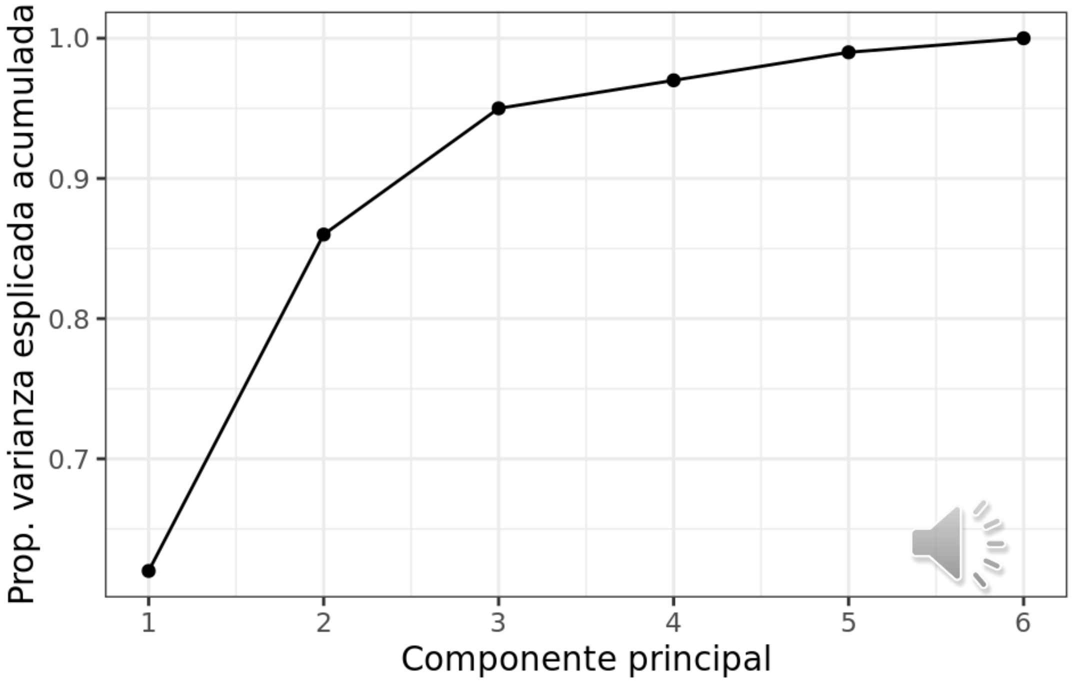
A TENER EN CUENTA

- Escalado de variables: PCA identifica direcciones cuya varianza es mayor. Por ello deberemos tener los datos en la misma escala. StandardScaler.



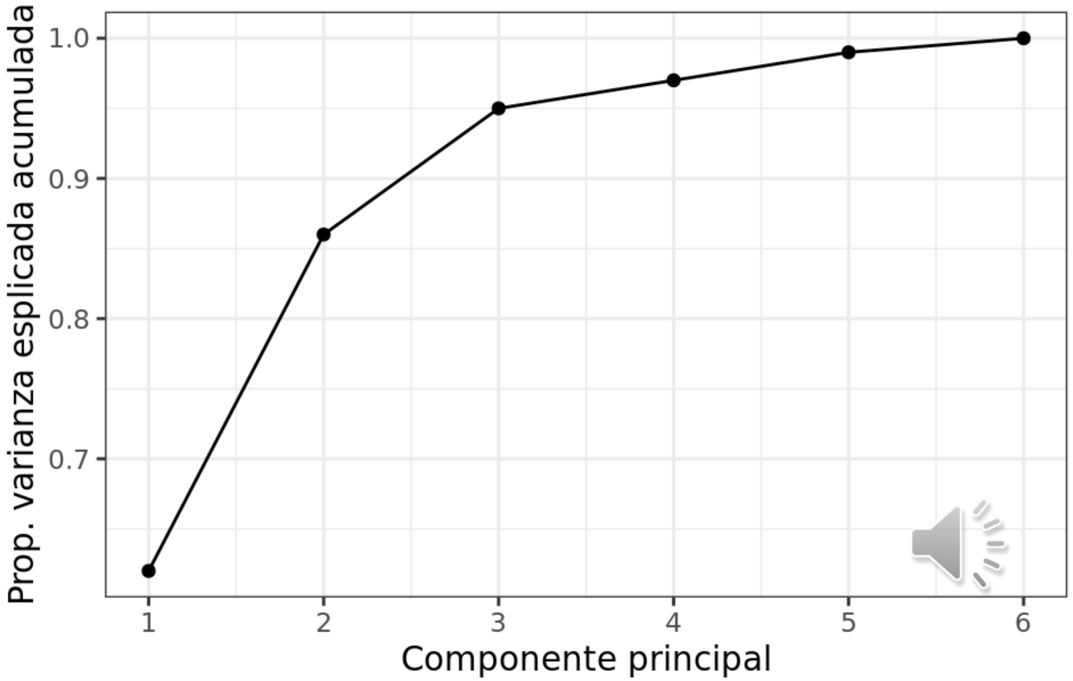
A TENER EN CUENTA

- Escalado de variables: PCA identifica direcciones cuya varianza es mayor. Por ello deberemos tener los datos en la misma escala. **StandardScaler**.
- Influencia de outliers: al trabajar con varianzas, PCA es altamente sensible a outliers. Es muy recomendable estudiar si los hay.



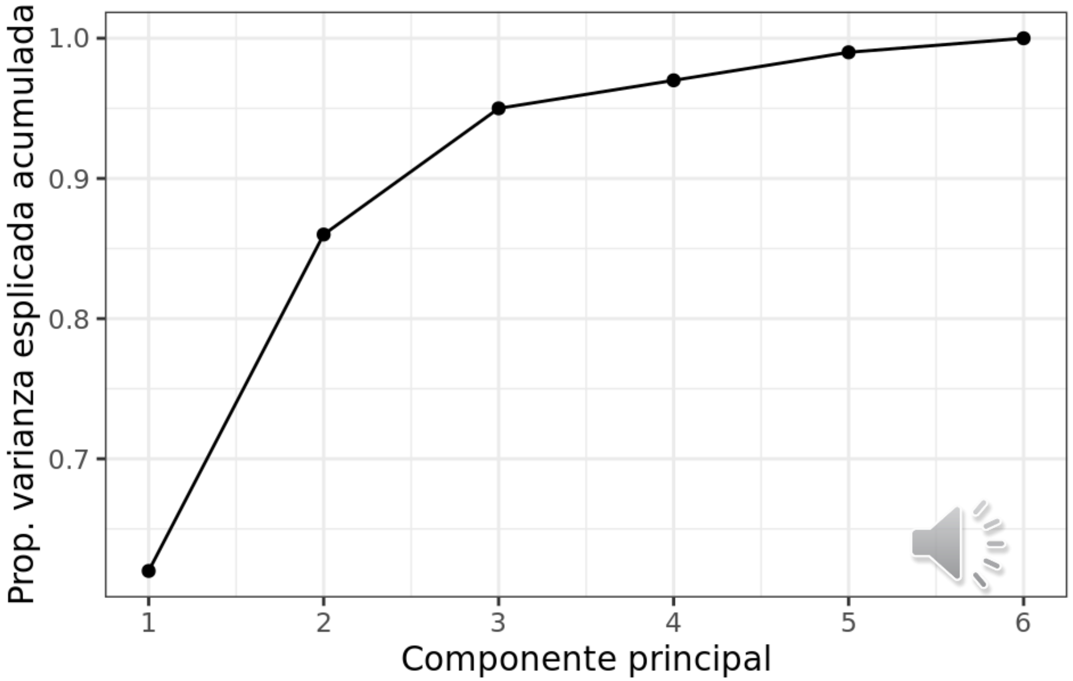
A TENER EN CUENTA

- Escalado de variables: PCA identifica direcciones cuya varianza es mayor. Por ello deberemos tener los datos en la misma escala. **StandardScaler**.
- Influencia de outliers: al trabajar con varianzas, PCA es altamente sensible a outliers. Es muy recomendable estudiar si los hay.
- ¿Cuánta información presente en el set de datos original se pierde al proyectar las observaciones en un espacio de menor dimensión? (Varianza explicada de cada componente principal).



A TENER EN CUENTA

- Escalado de variables: PCA identifica direcciones cuya varianza es mayor. Por ello deberemos tener los datos en la misma escala. **StandardScaler**.
- Influencia de outliers: al trabajar con varianzas, PCA es altamente sensible a outliers. Es muy recomendable estudiar si los hay.
- ¿Cuánta información presente en el set de datos original se pierde al proyectar las observaciones en un espacio de menor dimensión? (Varianza explicada de cada componente principal).
- Es de interés utilizar el número mínimo de componentes que resultan suficientes para explicar los datos.



¿Cómo funciona?

https://setosa.io/ev/principal-component-analysis/

THE BRIDGE