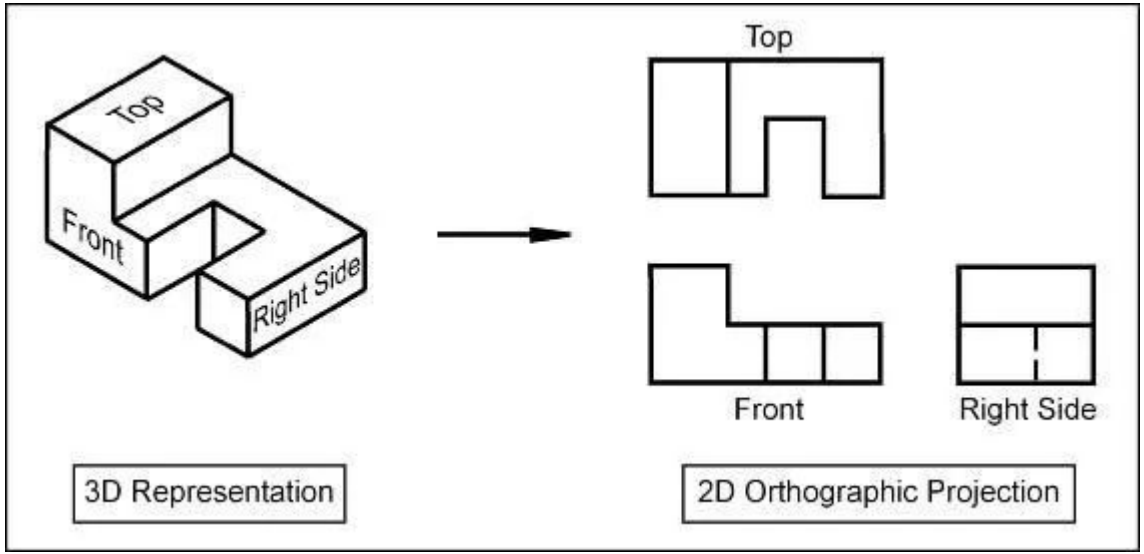
Fundamentos de Aprendizaje de Máquina

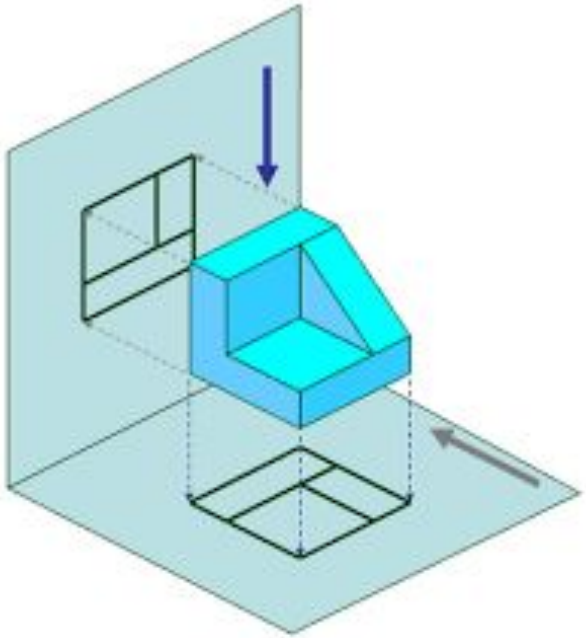
PhD Jorge Rudas

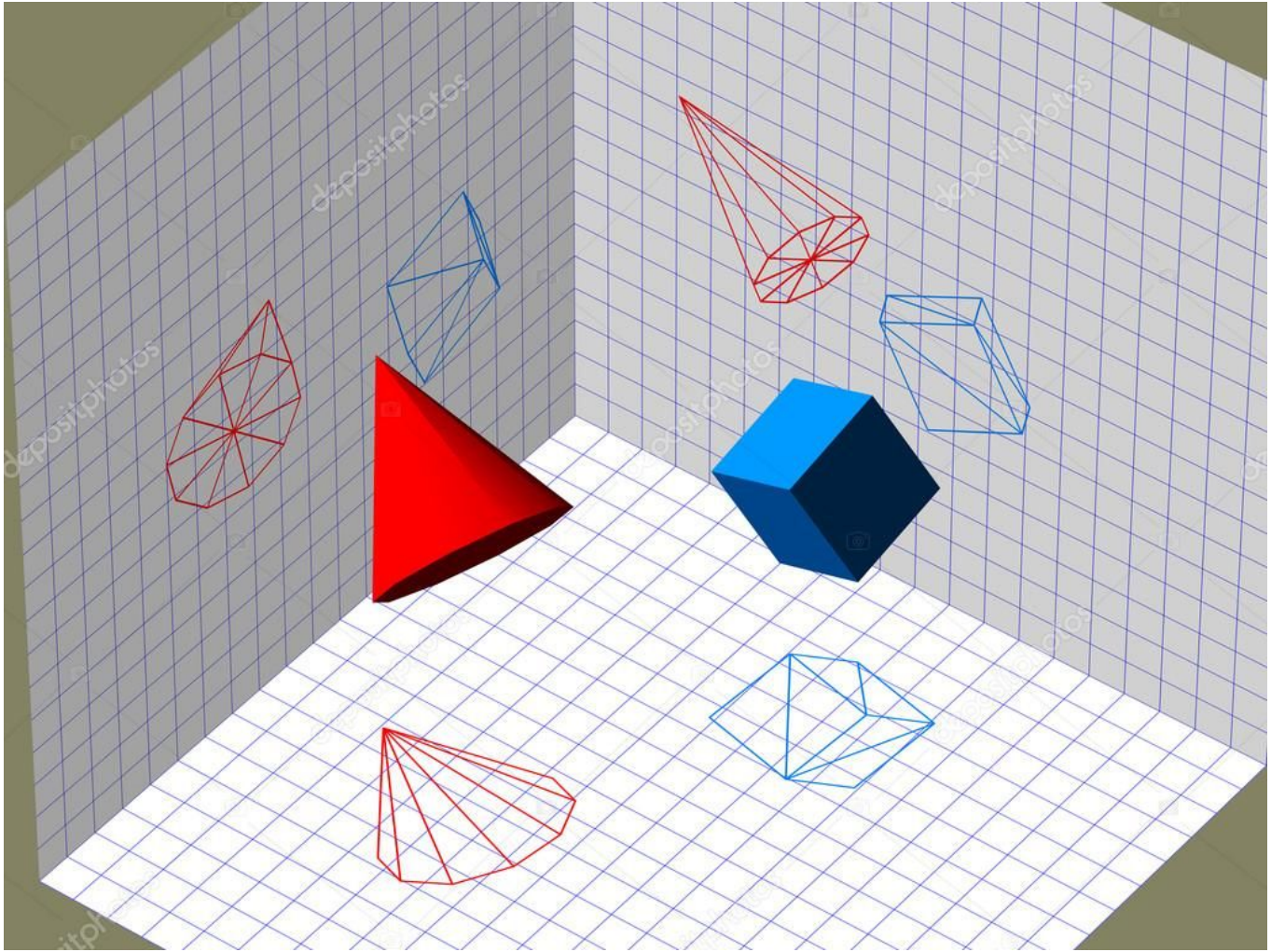
INFOTEP
INSTITUCIÓN DE EDUCACIÓN SUPERIOR

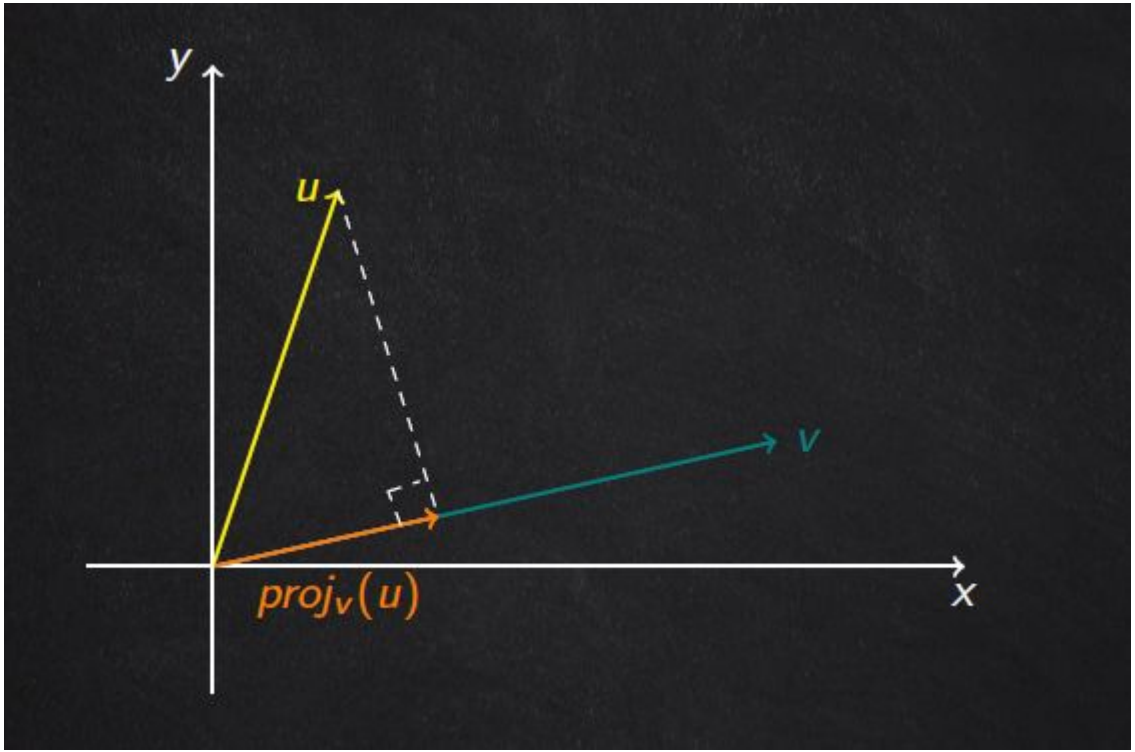
Análisis de Componentes Principales (PCA)

Es un método de proyección que proyecta las observaciones de un espacio p-dimensional con p variables a un espacio k-dimensional (donde k < p) para conservar la máxima cantidad de información (la información se mide aquí a través de la varianza total del conjunto de datos) de las dimensiones iniciales.


Top
Front
Right Side
Top
Front
Right Side
3D Representation
2D Orthographic Projection






y
u
v
$proj_v(u)$
x

PCA puede considerarse un método de minería de datos, ya que permite extraer fácilmente información de grandes conjuntos de datos

3 puntos claves sobre PCA

- Reducir la dimensión de los segmentos de datos con el fin de tener un menor número de variables, manteniendo de esta manera la información más importante

- El foco principal de este método de reducción de dimensionalidad es que el segmento de datos se vuelva más simple mientras retiene la información más importante.

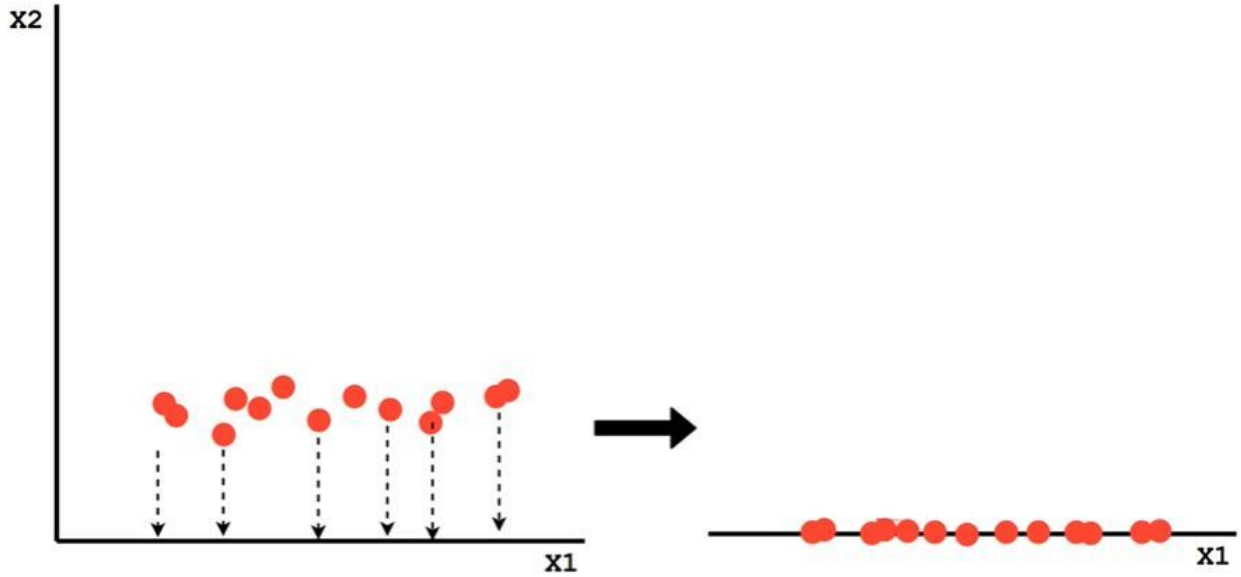
- Esto hace que sea más fácil visualizar y manipular los datos, lo cual ayuda a un análisis más rápido

Aplicaciones de PCA

- Reducción de dimensionalidad de datos
- Visualizar observaciones n-dimensionales en un espacio bidimensional o tridimensional para identificar grupos uniformes o atípicos de observaciones.
- Determinación de la dirección de objetos de imágenes
- Compresión de datos

Ejemplo 1

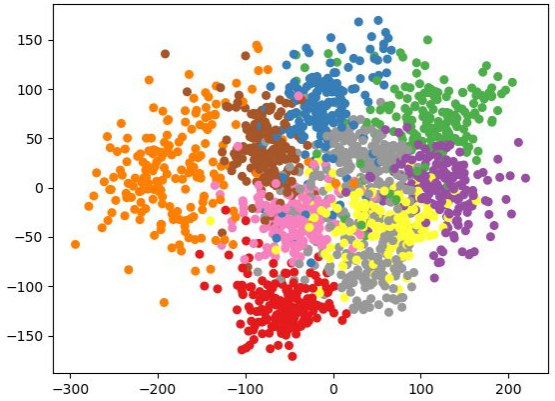
Reducción de dimensionalidad de datos



Ejemplo 2

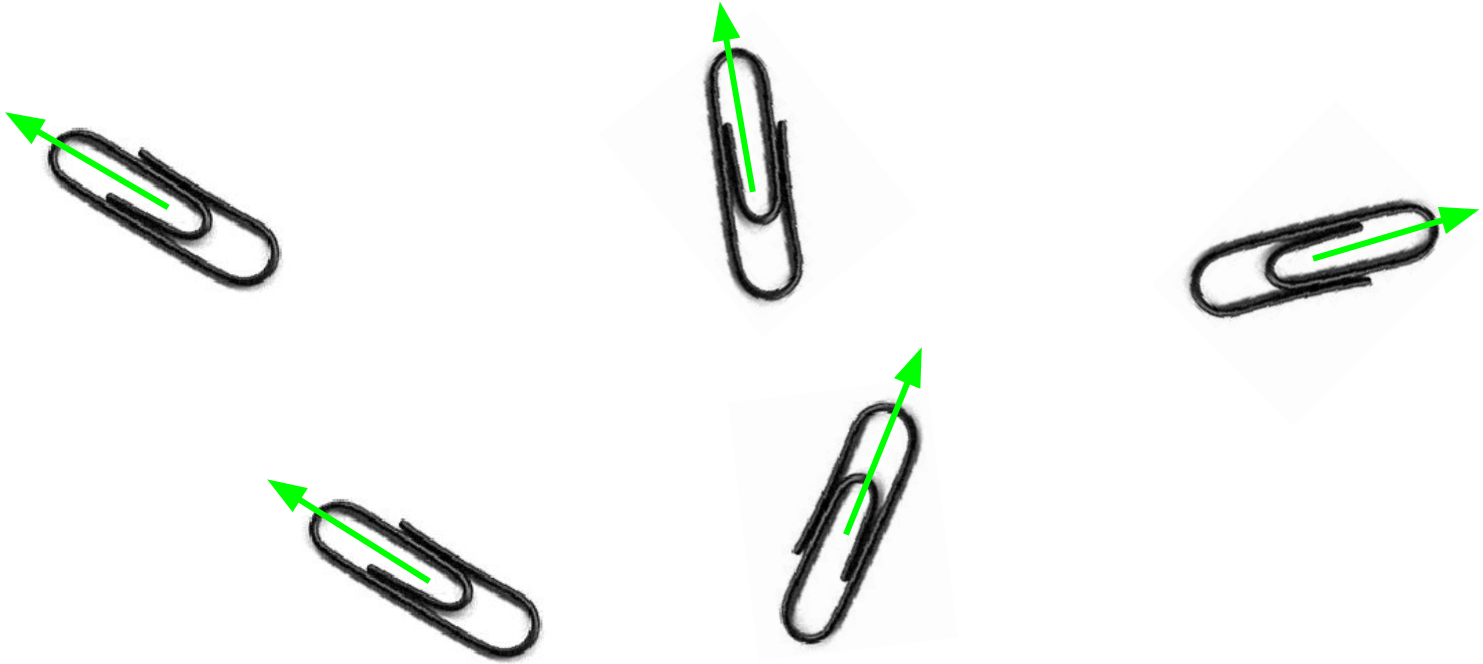
"Visualización" de datos n-dimensionales

bird_id	sex	bill_depth_mm	bill_width_mm	bill_length_mm	head_length_mm	body_mass_g	skull_size_mm
0000-00000	M	8.26	9.21	25.92	56.58	73.30	30.66
1142-05901	M	8.54	8.76	24.99	56.36	75.10	31.38
1142-05905	M	8.39	8.78	26.07	57.32	70.25	31.25
1142-05907	F	7.78	9.30	23.48	53.77	65.50	30.29
1142-05909	M	8.71	9.84	25.47	57.32	74.90	31.85
1142-05911	F	7.28	9.30	22.25	52.25	63.90	30.00
1142-05912	M	8.74	9.28	25.35	57.12	75.10	31.77
1142-05914	M	8.72	9.94	30.00	60.67	78.10	30.67
1142-05917	F	8.20	9.01	22.78	52.83	64.00	30.05
1142-05920	F	7.67	9.31	24.61	54.94	67.33	30.33
1142-05930	M	8.78	8.83	25.72	56.54	76.40	30.82
1142-05941	F	8.15	8.67	24.66	54.69	71.50	30.03
1142-05957	M	8.62	9.28	24.50	56.48	78.20	31.98



Ejemplo 3

Determinación de la orientación de objetos



Introducción al Álgebra de PCA

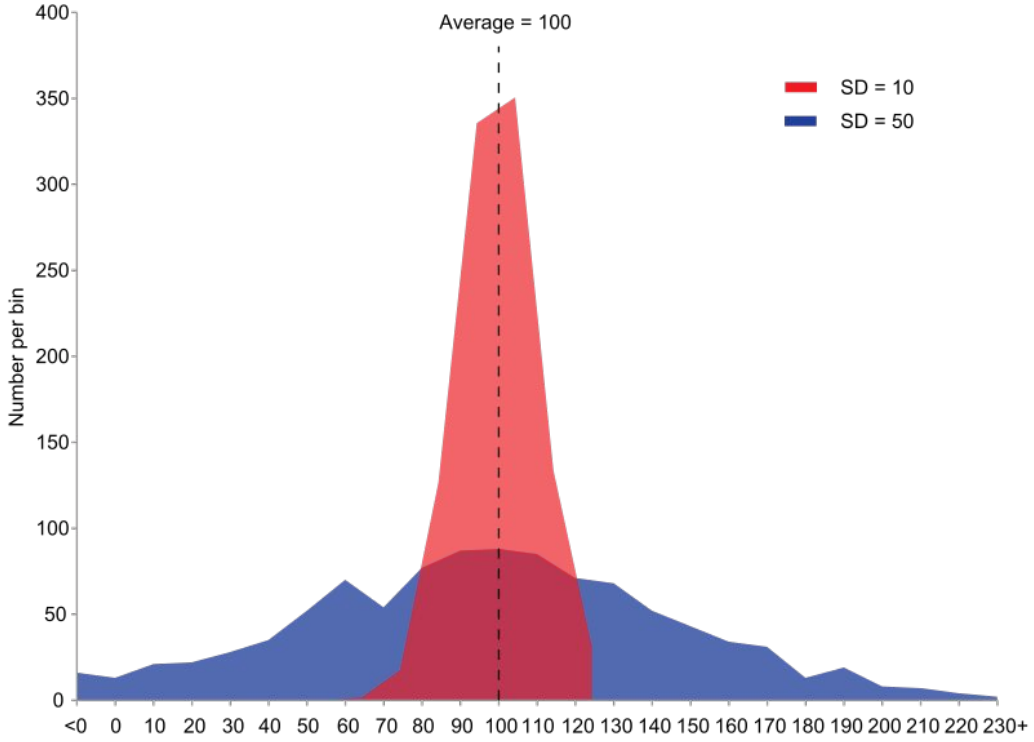
Dispersión de datos

$$\mathrm{Var}(X) = \frac{1}{n}\sum_j (x - x_j)^2 = \sigma_X^2$$

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_j (x - x_j)(y - y_j) = \sigma_{XY}$$

$$\mathrm{Var}(X) = \frac{1}{n}\sum_j (x - x_j)^2 = \sigma_X^2$$

Media y varianza de datos



Matriz de Covarianza

Example: blue jays dataset

bird_id	sex	bill_depth_mm	bill_width_mm	bill_length_mm	head_length_mm	body_mass_g	skull_size_mm
0000-00000	M	8.26	9.21	25.92	56.58	73.30	30.66
1142-05901	M	8.54	8.76	24.99	56.36	75.10	31.38
1142-05905	M	8.39	8.78	26.07	57.32	70.25	31.25
1142-05907	F	7.78	9.30	23.48	53.77	65.50	30.29
1142-05909	M	8.71	9.84	25.47	57.32	74.90	31.85
1142-05911	F	7.28	9.30	22.25	52.25	63.90	30.00
1142-05912	M	8.74	9.28	25.35	57.12	75.10	31.77
1142-05914	M	8.72	9.94	30.00	60.67	78.10	30.67
1142-05917	F	8.20	9.01	22.78	52.83	64.00	30.05
1142-05920	F	7.67	9.31	24.61	54.94	67.33	30.33
1142-05930	M	8.78	8.83	25.72	56.54	76.40	30.82
1142-05941	F	8.15	8.67	24.66	54.69	71.50	30.03
1142-05957	M	8.62	9.28	24.50	56.48	78.20	31.98

Covariance matrix of n variables $X_1 \dots X_n$

$$\mathrm{C} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn}^2 \end{pmatrix}$$

Taller en clase

$$\mathrm{Var}(X) = \frac{1}{n}\sum (x - x_i)^2 = \sigma_x^2$$

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum (x - x_j)(y - y_j) = \sigma_{XY}$$

$$\mathrm{C} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}^2 \end{pmatrix}$$

1. Descargue el dataset de Blue Jays
2. Importe como csv desde Google Sheet o desde Excel
3. Calcule la media, la varianza y la desviación estándar de cada característica para cada muestra
4. Calcule la covarianza para cada posible pareja de caracteristicas
5. Calcule la matriz de covarianzas

Nota: No se pueden utilizar funciones predefinidas por la hoja de cálculo

Los 2 grupos “voluntarios” para presentar sus resultados, sus análisis y sus interpretaciones

PCA es la factorización de la matriz de covarianza según:

$$\mathrm{C} = \mathrm{U}\mathrm{D}\mathrm{U}^{\mathrm{T}}$$

$$= \mathrm{U} \begin{pmatrix} \lambda_1^2 & 0 & \dots & 0 \\ 0 & \lambda_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n^2 \end{pmatrix} \mathrm{U}^{\mathrm{T}}$$

U: rotation matrix

D: diagonal matrix

λ_j^2: eigenvalues (= variance explained by each component)

The covariances between components are all 0

Components are uncorrelated

- Los vectores propios representan las direcciones de los ejes donde la máxima información está guardada, estos vectores son conocidos como Componentes Principales (Estos vectores son **SIEMPRE** ortogonales)
- Los valores propios, son valores unidos a los vectores propios, estos determinan la variación que tiene cada componente principal.
- Ambos vienen en pares y su número es igual al número de dimensionalidad de la matriz.
- Si clasificamos los vectores propios en orden de cada valor propio, de mayor a menor, podemos obtener los componentes principales en orden de importancia.

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

- Debemos utilizar el vector de características para reorientar los ejes a aquellos representados por los componentes principales de nuestra matriz.

- Puede realizarse multiplicando la transpuesta de la matriz original estandarizada por la matriz transpuesta del vector de características.

$$DatosFinales = MatrizOEst^{T} * VectorCaract^{T}$$

Resumen - PCA

1. Solo utilice datos numéricos
2. Estandarice sus datos numéricos para media 0 y varianza 1
3. Calcule la matriz de covarianza
4. Calcule PCA para la matriz de covarianza
5. Organice sus vectores propios por orden de importancia según sus valores propios
6. Defina cuál será su vector de características
7. Proyecte sus datos a lo largo de los ejes de los componentes principales

Ejemplo 1- Reducción de dimensionalidad de datos

Example: blue jays dataset

bird_id	sex	bill_depth_mm	bill_width_mm	bill_length_mm	head_length_mm	body_mass_g	skull_size_mm
0000-00000	M	8.26	9.21	25.92	56.58	73.30	30.66
1142-05901	M	8.54	8.76	24.99	56.36	75.10	31.38
1142-05905	M	8.39	8.78	26.07	57.32	70.25	31.25
1142-05907	F	7.78	9.30	23.48	53.77	65.50	30.29
1142-05909	M	8.71	9.84	25.47	57.32	74.90	31.85
1142-05911	F	7.28	9.30	22.25	52.25	63.90	30.00
1142-05912	M	8.74	9.28	25.35	57.12	75.10	31.77
1142-05914	M	8.72	9.94	30.00	60.67	78.10	30.67
1142-05917	F	8.20	9.01	22.78	52.83	64.00	30.05
1142-05920	F	7.67	9.31	24.61	54.94	67.33	30.33
1142-05930	M	8.78	8.83	25.72	56.54	76.40	30.82
1142-05941	F	8.15	8.67	24.66	54.69	71.50	30.03
1142-05957	M	8.62	9.28	24.50	56.48	78.20	31.98

Análisis BiVariados - Todos contra todos