IMT2200

INTRODUCCIÓN A CIENCIA DE DATOS

2022-2

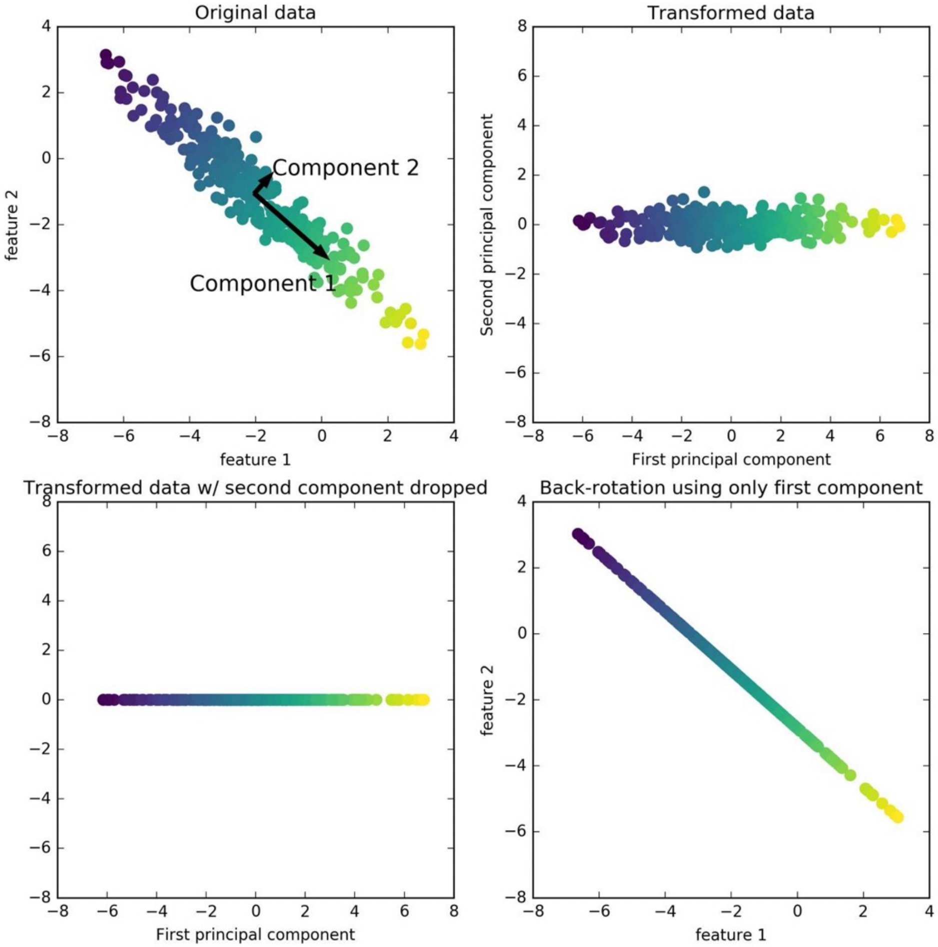
REDUCCIÓN DE DIMENSIONALIDAD

CLASE 25

ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

El análisis de componentes principales, apunta a reemplazar un gran número de variables iniciales, por un **número menor de predictores no correlacionados** que capturen la **mayor fracción de varianza** en los datos.

PCA no es una herramienta para hacer mejores predicciones, sino que ayuda a **resumir y visualizar mejor data multidimensional**, sin perder información.



ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

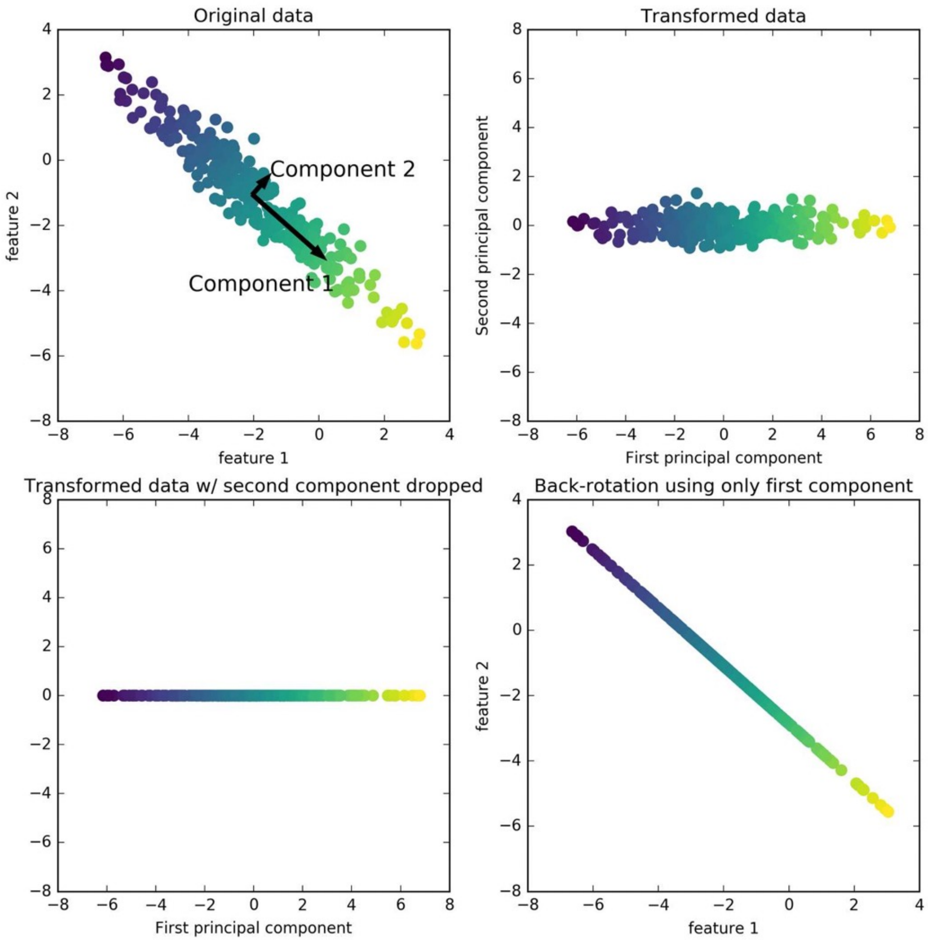
PCA aplica un **enfoque lineal** al problema de la muti-dimensionalidad:

Busca encontrar nuevas variables construidas como una **combinación lineal** de las variables originales, tal que maximizan la varianza explicada.

Sean:

$$z_u \Rightarrow \text{componente principal}$$
$$x_h \Rightarrow \text{variables originales}$$

$$\boxed{z_u = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k}$$



ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

Sean: $z_u \Rightarrow$ componente principal,
$x_h \Rightarrow$ variables originales

$$\boxed{z_u = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k}$$

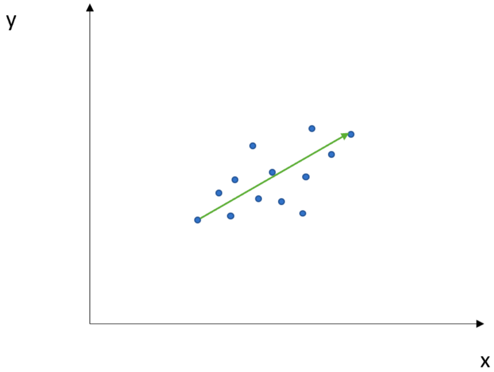
El problema matemático es **encontrar los coeficientes a_h** tal que z_u maximice la varianza explicada por las variables originales.

La solución está basada en los **autovalores y autovectores de la matriz de covarianza** entre las observaciones.

$$\sum = \begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(y,x) & Var(y) \end{bmatrix}$$

Matriz de covarianza

$$\begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(y,x) & Var(y) \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \end{bmatrix}' = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \end{bmatrix}$$



Autovalor ($\boldsymbol{v_1}$) y autovector (λ_1) de la matriz de covarianza

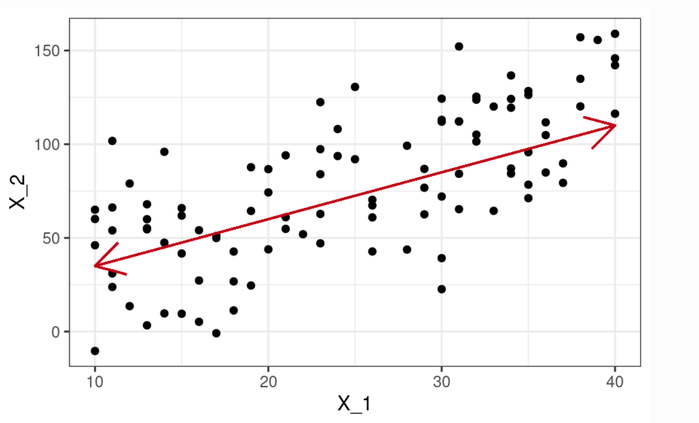
23:02

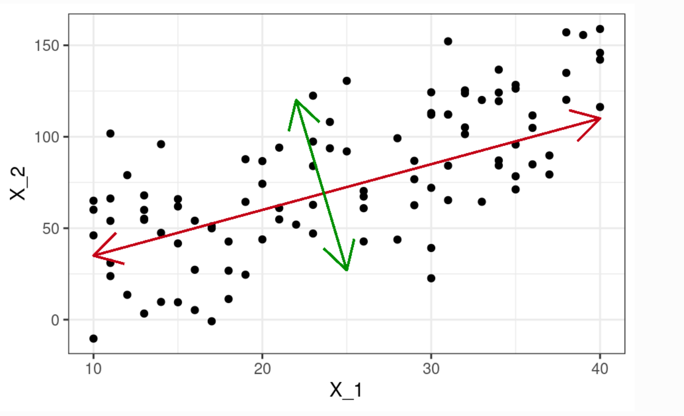
ANÁLISIS DE COMPONENTES PRINCIPALES

Ejemplo: conjunto de observaciones 2-D

- Ordenando los autovalores en orden descendente, el primer autovector representa la dirección de máxima dispersión en los datos, y sucesivamente…

- La raíz cuadrada del autovector equivale a la desviación estándar explicada por dicho componente.





23:02

ANÁLISIS DE COMPONENTES PRINCIPALES: EJEMPLO

Dataset de imágenes “Faces Dataset, Cambridge.

Objetivo: entrenar un modelo que permita identificar a qué persona pertenece cada imagen (es decir, reconocer su cara)

ANÁLISIS DE COMPONENTES PRINCIPALES: EJEMPLO

Dataset de imágenes “Faces Dataset, Cambridge.

Primera componente principal (autovector) ➔ captura el 80% de la varianza de las imágenes.

La reconstrucción de las imágenes mejora a medida que se consideran más componentes principales.

ANÁLISIS DE COMPONENTES PRINCIPALES: EJEMPLO

Figure 3-11. Reconstructing three face images using increasing numbers of principal components

Dataset de imágenes “Faces Dataset, Cambridge.

Primera componente principal (autovector) ➔ captura el 80% de la varianza de las imágenes.

La reconstrucción de las imágenes mejora a medida que se consideran más componentes principales.