

ANÁLISIS DE DATOS MASIVOS

ANÁLISIS DE COMPONENTES PRINCIPALES

Blanca Vázquez

23 de enero de 2025

- Objetos cada vez más dispersos conforme aumenta el número de dimensiones

LA HIPÓTESIS DE LA VARIEDAD

- Ejemplos pueden vivir en una variedad de muchas menores dimensiones que el espacio original

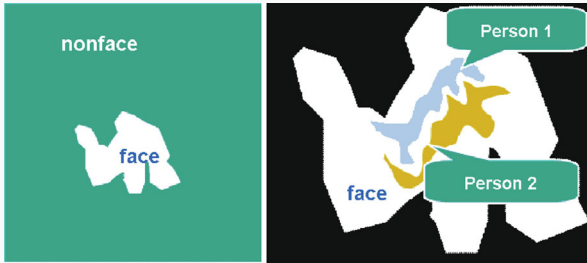
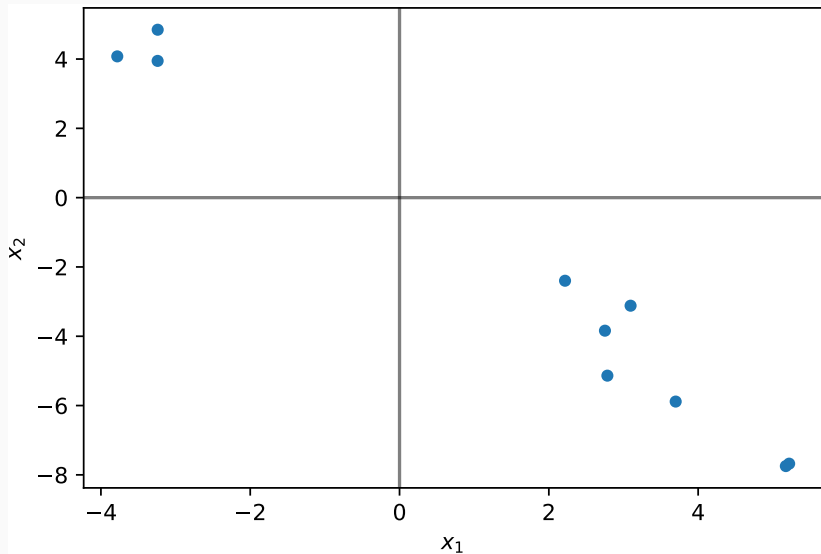


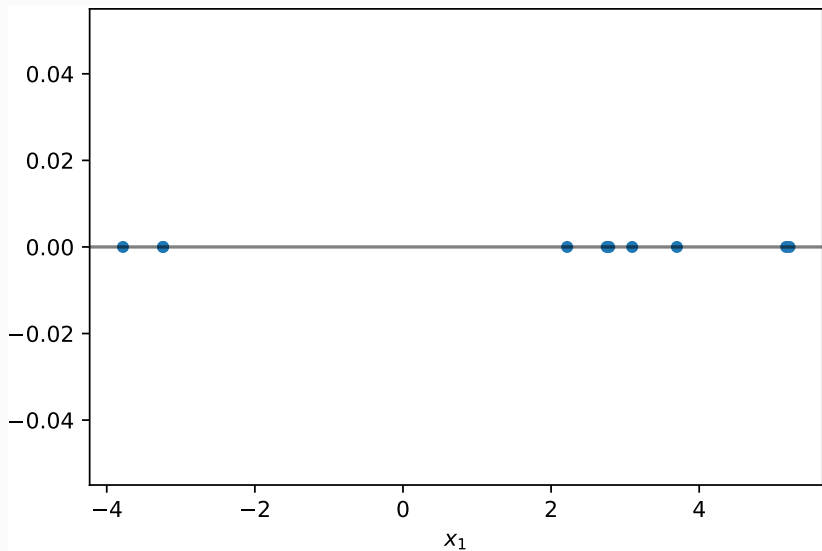
Imagen tomada de Li and Jain, 2005

- Proyección ortogonal de un conjunto de vectores
- Genera una nueva vista
- Aplicaciones
 - Visualización
 - Extracción de características
 - Reducción de dimensionalidad
 - Compresión

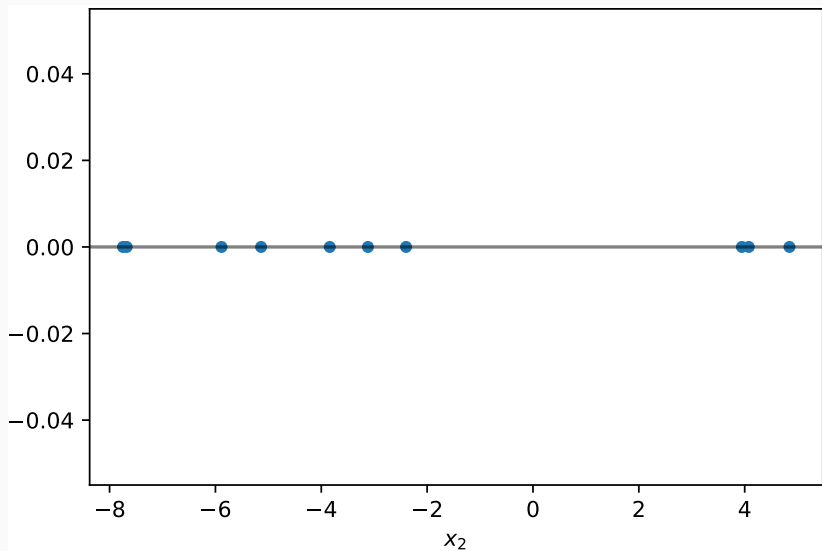
INTUICIÓN: DATOS EN 2D



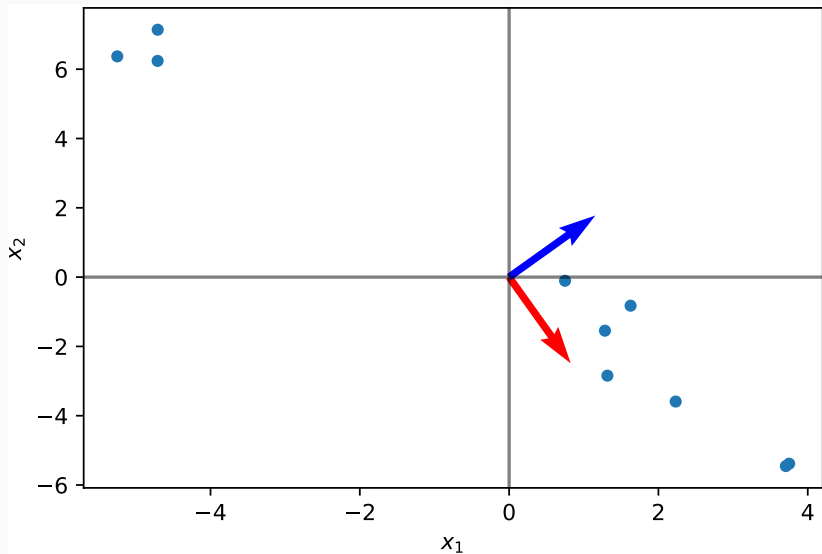
INTUICIÓN: DATOS VISTOS DESDE EL EJE x



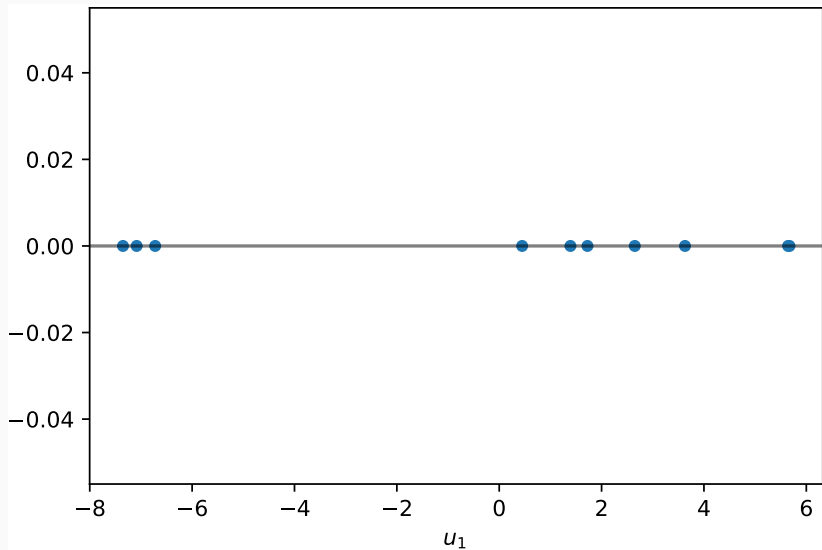
INTUICIÓN: DATOS VISTOS DESDE EL EJE y



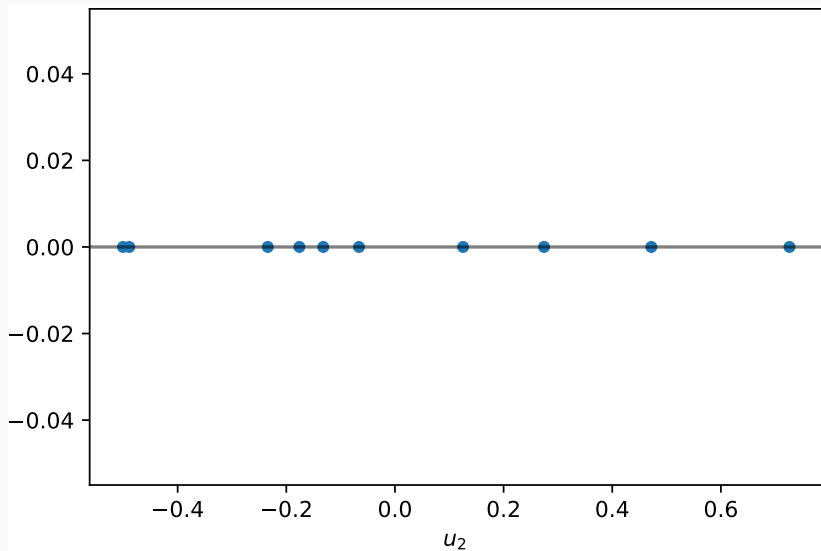
INTUICIÓN: NUEVOS EJES u_1 Y u_2



INTUICIÓN: DATOS PROYECTADOS SOBRE EL EJE u_1



INTUICIÓN: DATOS PROYECTADOS SOBRE EL EJE u_2



- Dado un conjunto de vectores $\mathbf{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ de d dimensiones, el primer componente principal es el vector \mathbf{u}_1 que maximice la varianza de los datos proyectados, donde \mathbf{u}_1 es un vector de d dimensiones

- La proyección de un vector sobre el componente principal está dado por

$$\hat{\mathbf{x}}^{(i)} = \mathbf{u}_1^\top \mathbf{x}^{(i)}$$

PCA POR MÁXIMA VARIANZA (1)

- La proyección de un vector sobre el componente principal está dado por

$$\hat{\mathbf{x}}^{(i)} = \mathbf{u}_1^\top \mathbf{x}^{(i)}$$

- La media de los datos proyectados es $\mathbf{u}_1^\top \bar{\mathbf{x}}$, donde

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}$$

PCA POR MÁXIMA VARIANZA (1)

- La proyección de un vector sobre el componente principal está dado por

$$\hat{\mathbf{x}}^{(i)} = \mathbf{u}_1^\top \mathbf{x}^{(i)}$$

- La media de los datos proyectados es $\mathbf{u}_1^\top \bar{\mathbf{x}}$, donde

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}$$

- La varianza es $\frac{1}{n} \sum_{i=1}^n [\mathbf{u}_1^\top \mathbf{x}^{(i)} - \mathbf{u}_1^\top \bar{\mathbf{x}}]^2 = \mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1$, donde

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^{(i)} - \bar{\mathbf{x}})(\mathbf{x}^{(i)} - \bar{\mathbf{x}})^\top$$

PCA POR MÁXIMA VARIANZA (2)

- Queremos encontrar el vector \mathbf{u}_1 que maximice la varianza de los datos proyectados $\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1$, con la restricción que $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

PCA POR MÁXIMA VARIANZA (2)

- Queremos encontrar el vector \mathbf{u}_1 que maximice la varianza de los datos proyectados $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$, con la restricción que $\mathbf{u}_1^T \mathbf{u}_1 = 1$
- Podemos formular esta restricción usando multiplicadores de Lagrange: $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T \mathbf{u}_1)$

PCA POR MÁXIMA VARIANZA (2)

- Queremos encontrar el vector \mathbf{u}_1 que maximice la varianza de los datos proyectados $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$, con la restricción que $\mathbf{u}_1^T \mathbf{u}_1 = 1$
- Podemos formular esta restricción usando multiplicadores de Lagrange: $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T \mathbf{u}_1)$
- Derivando e igualando a cero, tenemos

$$\mathbf{S} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

PCA POR MÁXIMA VARIANZA (2)

- Queremos encontrar el vector \mathbf{u}_1 que maximice la varianza de los datos proyectados $\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1$, con la restricción que $\mathbf{u}_1^\top \mathbf{u}_1 = 1$
- Podemos formular esta restricción usando multiplicadores de Lagrange: $\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top \mathbf{u}_1)$

- Derivando e igualando a cero, tenemos

$$\mathbf{S} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

- Esto es, \mathbf{u}_1 es un vector propio de \mathbf{S} , donde $\lambda_1 = \mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1$ es su valor propio que se corresponde con la varianza de los datos proyectados

- El siguiente componente principal es el vector propio que maximice la varianza de los datos proyectados entre el conjunto de vectores ortogonales a los que ya han sido elegidos.

Este proceso se realiza de forma incremental hasta obtener los K componentes principales.

- El conjunto de K componentes principales forman una base ortonormal de funciones.

- Para proyectar un vector $\mathbf{x}^{(i)}$ sobre los componentes principales

$$\mathbf{z}^{(i)} = \mathbf{U}^\top [\mathbf{x}^{(i)} - \bar{\mathbf{x}}]$$

donde $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K]$ es una matriz cuyas columnas se corresponden con los K componentes principales.

- La reconstrucción está dada por

$$\hat{\mathbf{x}}^{(i)} = \mathbf{U}\mathbf{z}^{(i)} + \bar{\mathbf{x}}$$

PCA POR VECTORES Y VALORES PROPIOS

- Busca subespacio de K dimensiones que maximiza varianza (o minimiza error) de los ejemplos
 - Definido por eigenvectores $\mathbf{u}_1, \dots, \mathbf{u}_K$ con eigenvalores más grandes $\lambda_1, \dots, \lambda_K$ de la matriz de covarianza

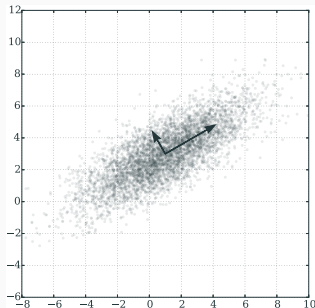


Figura tomada de Wikipedia (Principal Component Analysis)

PCA APLICADO A IMÁGENES DE ROSTROS

- Componentes principales se toman como base (**eigenfaces**)
- Nuevos rostros se proyectan en subespacio encontrado para ser comparados

