Datos Masivos II

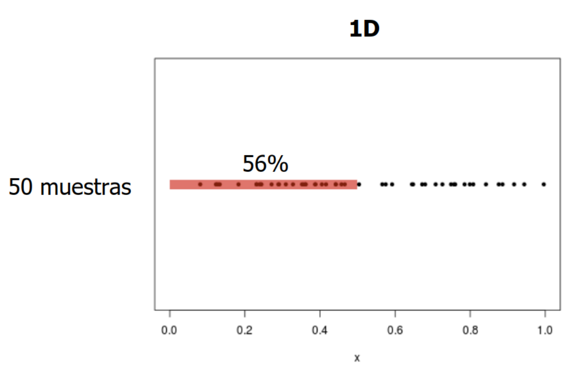
Análisis de componentes principales

Blanca Vázquez-Gómez y Gibran Fuentes-Pineda

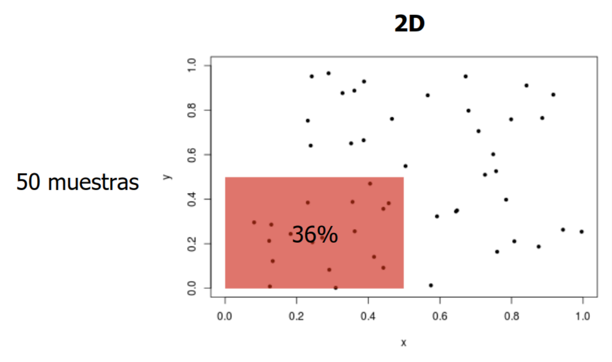
Septiembre 2020

Recordando la maldición de la dimensionalidad

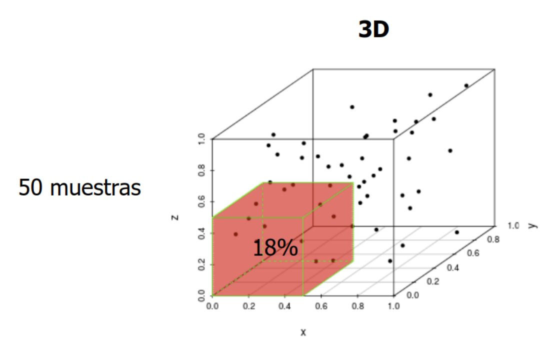
- Objetos cada vez más dispersos conforme aumenta el número de dimensiones



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



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



Recordando la maldición de la dimensionalidad

- Objetos cada vez más dispersos conforme aumenta el número de dimensiones
- **Ejercicio**: ¿Cuántas muestras necesitaría para cubrir un espacio de 1000 dimensiones con una precisión del 56 %?

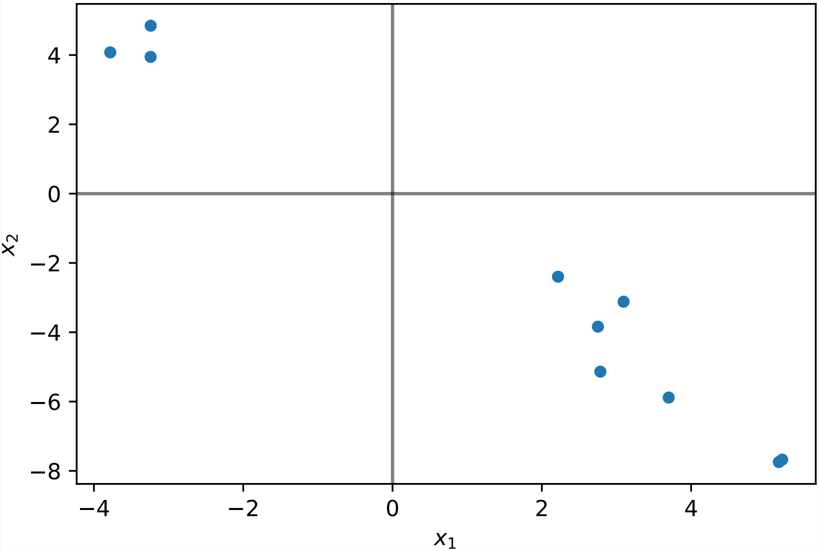
- Ejemplos pueden vivir en una variedad de muchas menores dimensiones que el espacio original

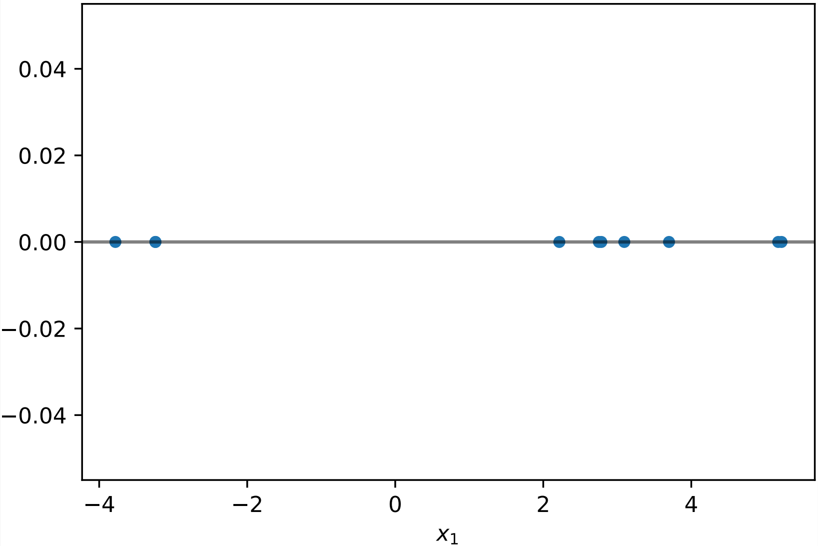
Imagen tomada de Li and Jain, 2005

Análisis de componentes principales (PCA)

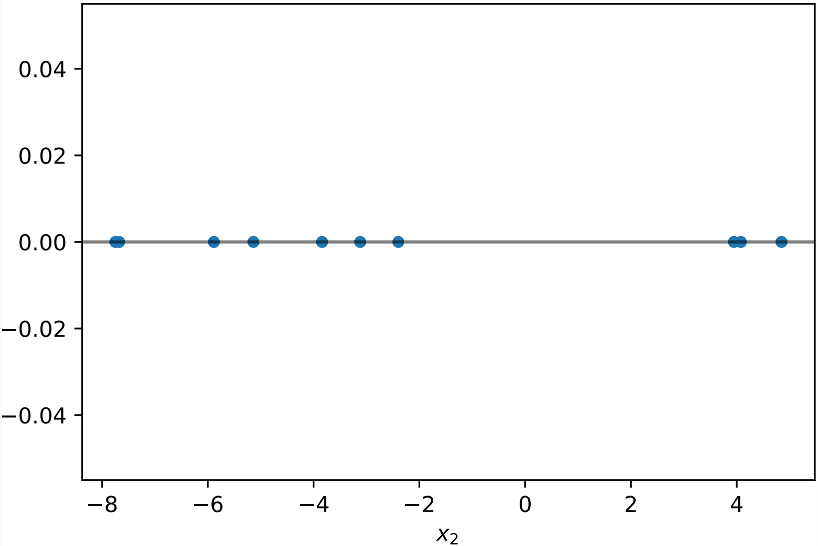
- Proyección ortogonal de un conjunto de vectores
- Genera una nueva vista
- Aplicaciones
  - Visualización
  - Extracción de características
  - Reducción de dimensionalidad
  - Compresión

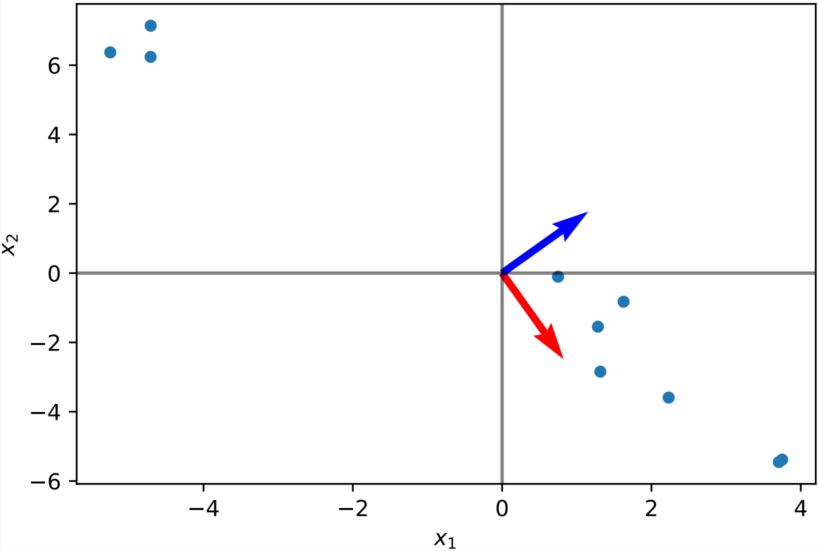
Intuición: datos en 2D



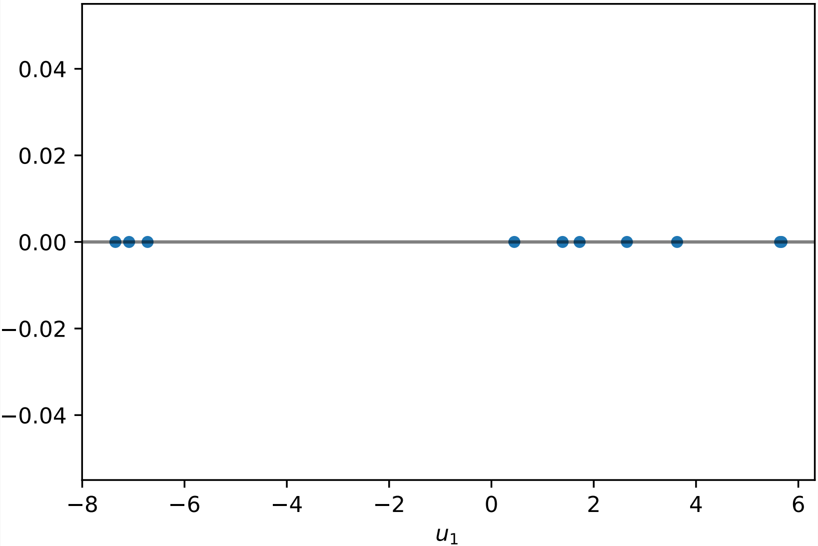

0.04
0.02
0.00
−0.02
−0.04
−4
−2
0
2
4
x_1

Intuición: datos vistos desde el eje y

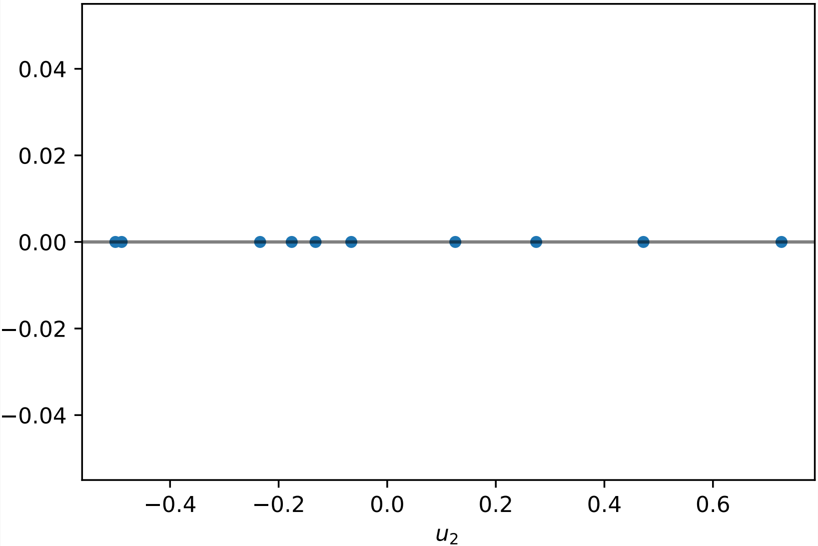



6
4
2
0
−2
−4
−6
x_2
−4
−2
0
2
4
x_1

Intuición: datos proyectados sobre el eje u_1



Intuición: datos proyectados sobre el eje u_2



PCA: formulación de máxima varianza

- Dado un conjunto de vectores $\mathsf{X} = \{\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(n)}\}$ 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

PCA por máxima varianza (1)

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

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

PCA por máxima varianza (1)

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

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

- 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
$$\hat{x}^{(i)} = \mathbf{u}_1^\top \mathbf{x}^{(i)}$$
- 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 $\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{u}_1^\top \mathbf{x}^{(i)} - \mathbf{u}_1^\top \bar{\mathbf{x}}\right] = \mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1$, 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 $\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1$, con la restricción que $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

- Podemos formular esta restricción usando multiplicadores de Lagrange: $\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top \mathbf{u}_1)$

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

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

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

PCA por máxima varianza (3)

- Para obtener el siguiente componente principal, buscamos 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 m componentes principales.
- El conjunto de m componentes principales forman una base ortonormal de funciones.

- Gracias a que forman una base de funciones completa, podemos representar cualquier vector $\mathsf{x}^{(i)}$ como una combinación lineal de los componentes principales

$$\hat{\mathsf{x}}^{(i)} = \sum_{j=1}^{m} \alpha_{i,j} \mathsf{u}_j$$

donde $\alpha_{i,j} = \mathsf{x}^{(i)\top} \mathsf{u}_j$.

- Esto es

$$\hat{\mathsf{x}}^{(i)} = \sum_{j=1}^{m} (\mathsf{x}^{(i)\top} \mathsf{u}_j) \mathsf{u}_j$$

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

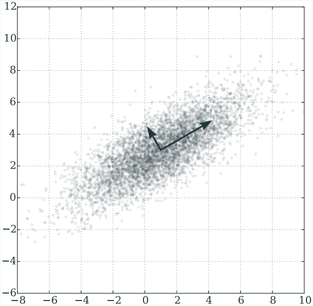
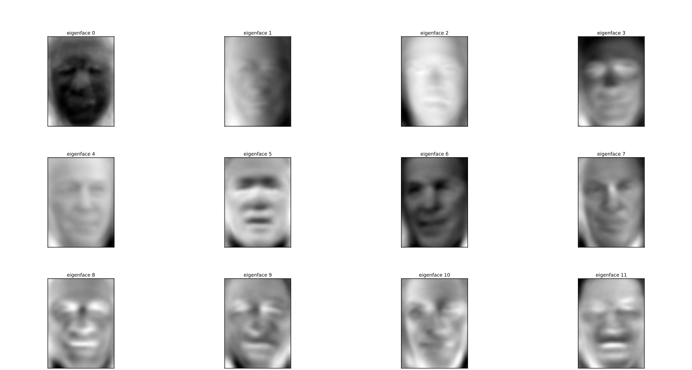


Figura tomada de Wikipedia (*Principal Component Analysis*)

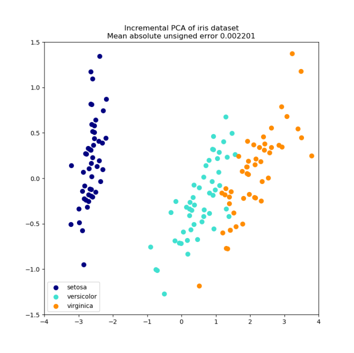
PCA aplicado a imágenes de rostros

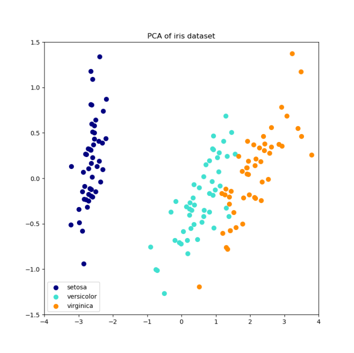
- Componentes principales se toman como base (**eigenfaces**)
- Nuevos rostros se proyectan en subespacio encontrado para ser comparados



Ejemplo de *`http://scikit-learn.org`*

PCA incremental





Ejemplo de *`http://scikit-learn.org`*

- Variables latentes continuas $\mathbf{z} \in \mathbb{R}^K$, con a priori gaussiana

$$\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

- Variables observadas continuas $\mathbf{x} \in \mathbb{R}^d$ con[1]

$$\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{U}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

- Distribución sobre $\mathbf{x}$ está dada por

$$P(\mathbf{x}) = \int P(\mathbf{x}|\mathbf{z})P(\mathbf{z})d\mathbf{z} = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$$

donde $\mathbf{C} = \mathbf{U}\mathbf{U}^\top + \sigma^2\mathbf{I}$

[1] Cuando $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$, $\boldsymbol{\mu}_0 = \mathbf{0}$ y $\boldsymbol{\Sigma}_0 = \mathbf{I}$, se conoce como *análisis de componentes principales probabilista* (PPCA).

Proceso generativo de PPCA

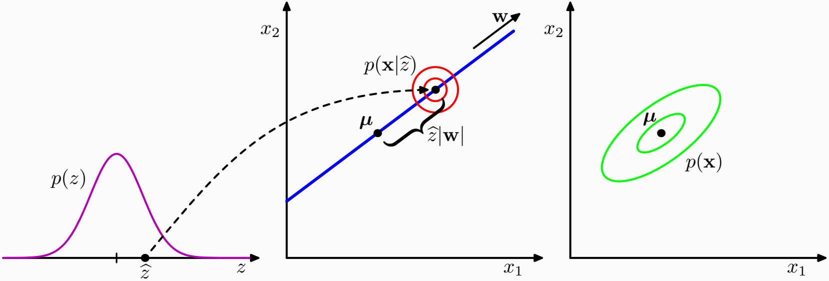


Imagen tomada de Bishop, PRML 2007

EM para PCA

- Presuponiendo $\sigma^2 = 0$, se pueden encontrar parámetros de PCA por máxima verosimilitud usando el algoritmo EM
  1. Paso E: $\tilde{\mathsf{Z}} = (\mathsf{U}^\top \mathsf{U})^{-1} \mathsf{U}^\top \tilde{\mathsf{X}}$
  2. Paso M: $\mathsf{U} = \tilde{\mathsf{X}}^\top \tilde{\mathsf{Z}}^\top (\tilde{\mathsf{Z}} \tilde{\mathsf{Z}}^\top)^{-1}$

  donde $\tilde{\mathsf{X}} = \mathsf{X}^\top$