



Reducción de Dimensionalidad

Machine Learning



Agenda

1. Motivación
2. Reducción de Dimensionalidad
3. Análisis de Componentes Principales
4. Implementación Práctica
5. Aplicación de PCA



Motivación

Motivación

En esta clase vamos a tratar un tema particular del aprendizaje no supervisado que es la **reducción de dimensionalidad**.



¿Qué modelo elegir?

Después de mucho pensar, llegan a las siguientes opciones:

- Tener más datos para el entrenamiento.
- Considerar menos características.
- Obtener más características.
- Considerar combinaciones de características más complejas.
- Aumentar o reducir el valor de λ .



¿Qué modelo elegir?

Considerando las opciones iniciales, tenemos que:


- Tener más datos para el entrenamiento. → *resuelve* varianza alta
- Considerar menos características. → *resuelve* varianza alta
- Obtener más características. → *resuelve* sesgo alto
- Considerar combinaciones de características más complejas. → *resuelve* sesgo alto
- Aumentar el valor de λ . → *resuelve* varianza alta
- Reducir el valor de λ . → *resuelve* sesgo alto



Reducción de Dimensionalidad

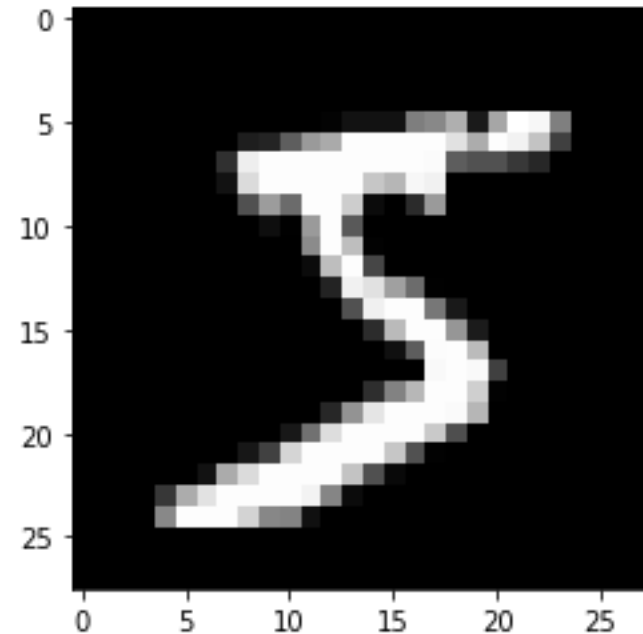


Reducción de Dimensionalidad

- Muchos problemas de ML requieren el uso de **miles o millones de características** para entrenar los modelos.
 - Esto ocasiona que los entrenamientos sean **lentos** y que no lleguen a una **buena solución**.
 - Esto se conoce como la **maldición de la dimensionalidad**.
- 


Reducción de Dimensionalidad

- En práctica, es posible reducir el número de características.
- Consideren el problema de clasificación de dígitos que presenta el conjunto de datos MNIST.
- ¿Cuáles son las características del modelo?
- ¿Cómo reducen el número de características?





Reducción de Dimensionalidad

- Al contar con una mayor dimensión (los datos), existe un alto riesgo de que estos sean dispersos (se encuentren lejos uno del otro).
 - Una nueva instancia puede estar lejos de las otras, lo que dificulta el proceso de inferencia.
 - Entre mayor número de dimensiones, mayor es el riesgo de sobreajuste.
- 

Reducción de Dimensionalidad

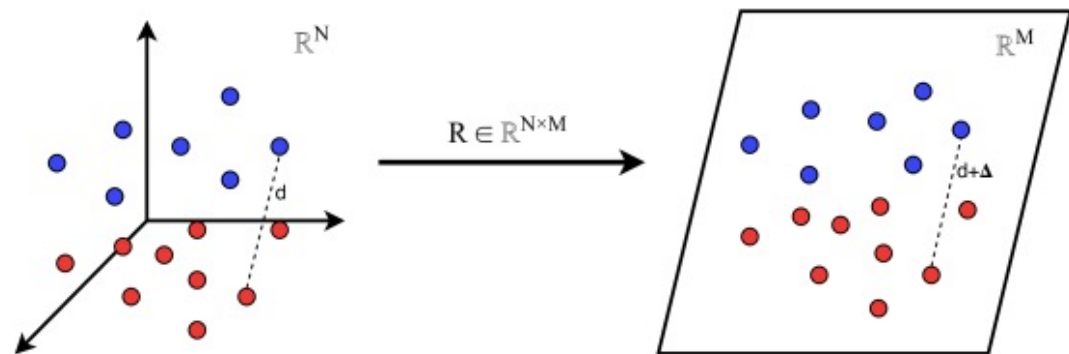
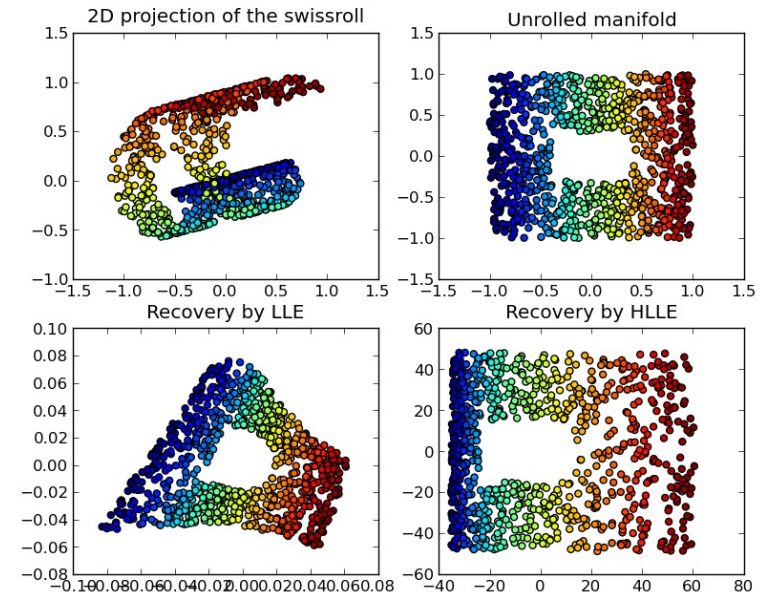
- Para evitar esto, se pueden considerar más datos.
 - No siempre es posible
 - Requiere muchos datos conforme se incrementa el número de dimensiones.



Reducción de Dimensionalidad

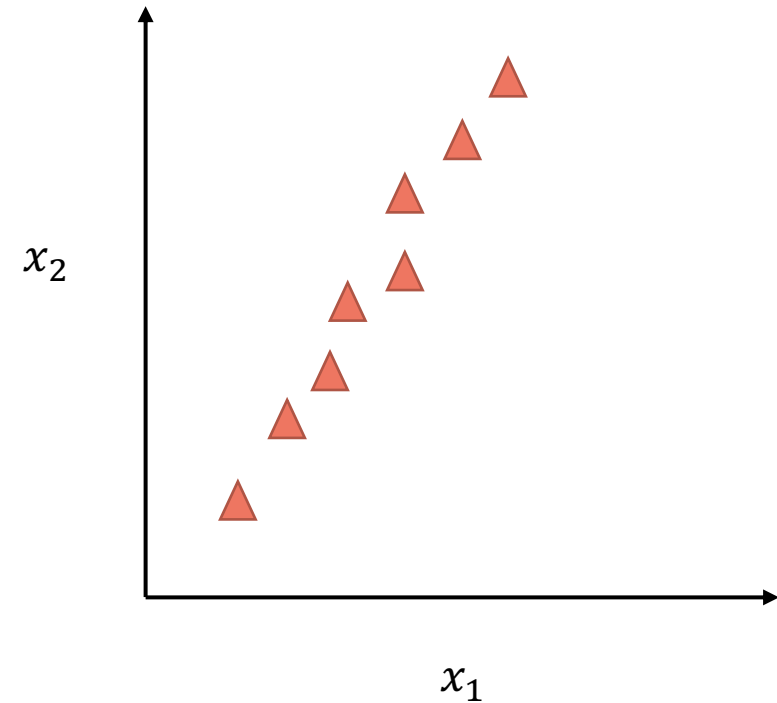
Existen dos líneas para los algoritmos de reducción de dimensionalidad:

- Proyección
- Manifolds



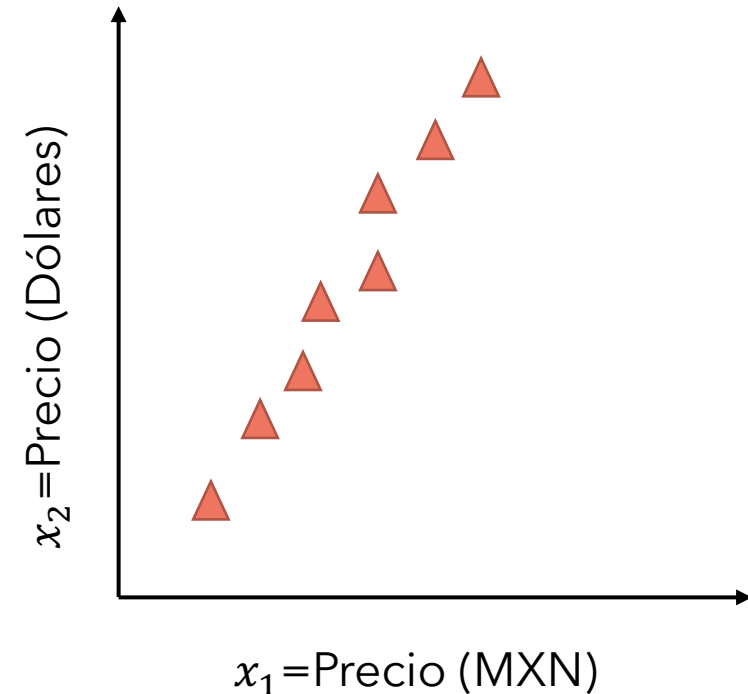
Reducción de Dimensionalidad

- Rara vez las características de un modelo son las adecuadas.
- Puede necesitar más o menos.
- ¿Cómo determinar aquellas que *sobran*?



Reducción de Dimensionalidad

- Rara vez las características de un modelo son las adecuadas.
- Puede necesitar más o menos.
- ¿Cómo determinar aquellas que *sobran*?



Reducción de Dimensionalidad

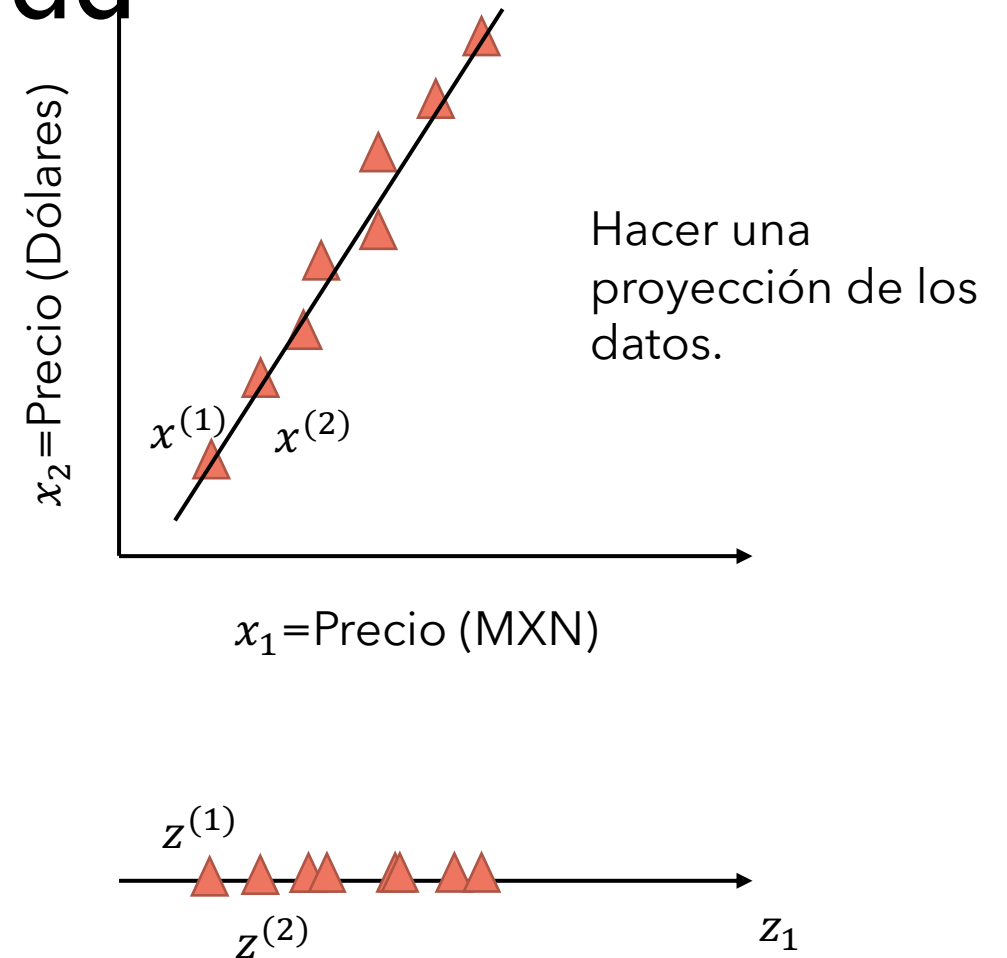
Reducir la dimensionalidad de la información

$$f: \mathbb{R}^2 \rightarrow \mathbb{R}$$

$$x^{(1)} \in \mathbb{R}^2 \rightarrow z^{(1)} \in \mathbb{R}$$

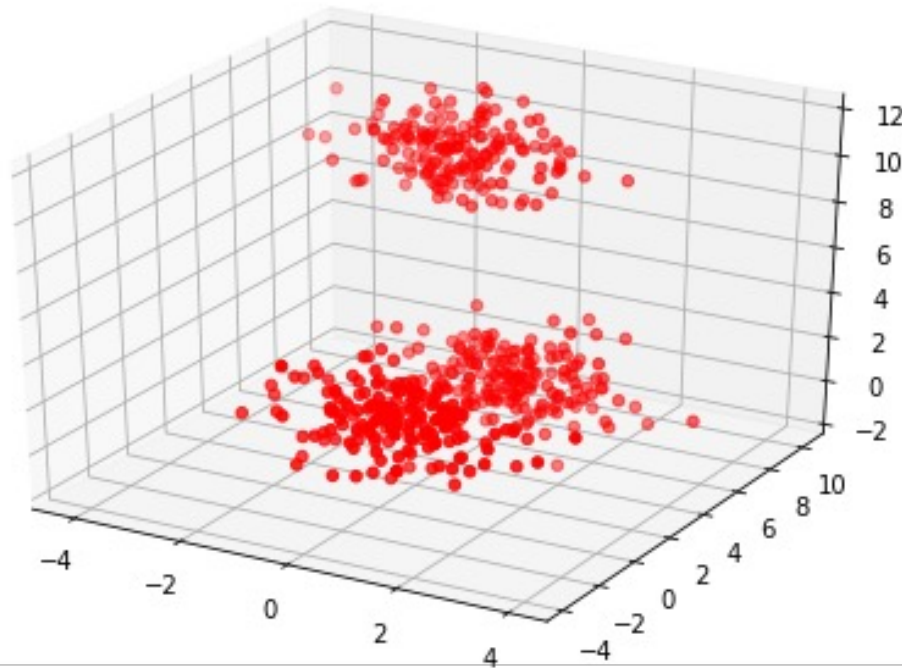
$$x^{(2)} \in \mathbb{R}^2 \rightarrow z^{(2)} \in \mathbb{R}$$

\vdots

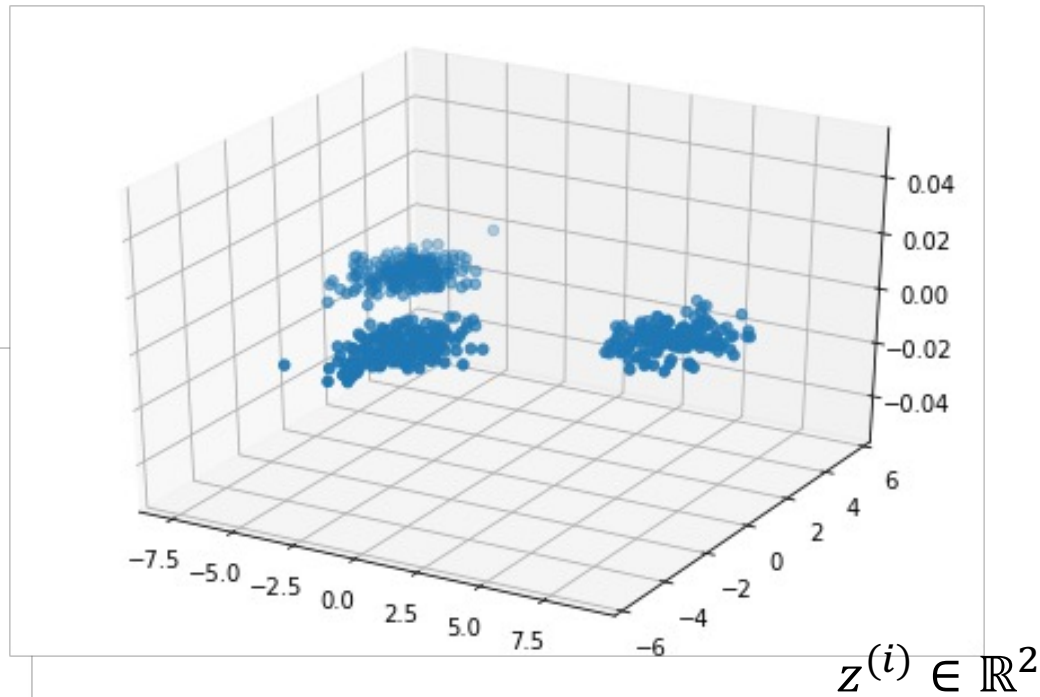


z_1 sería la posición en la recta nueva.

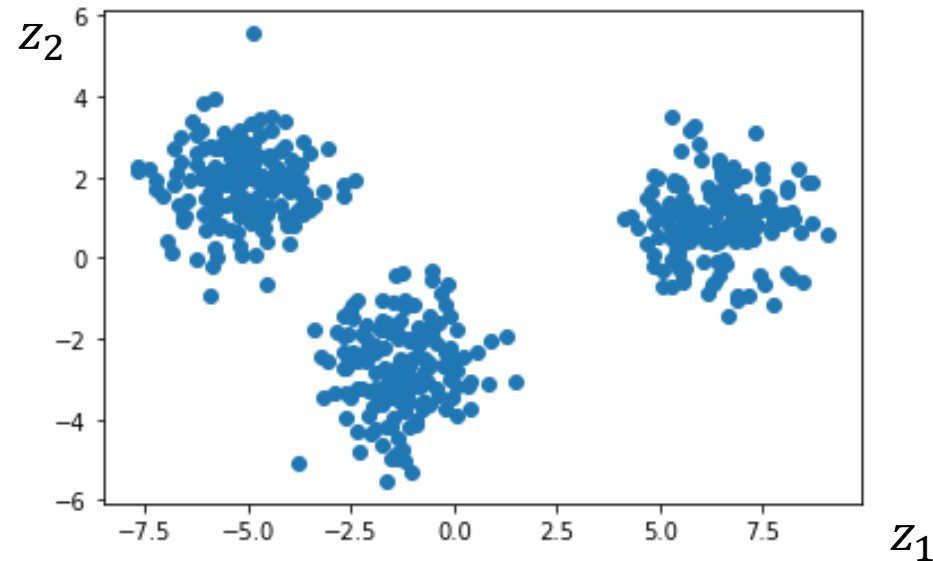
Reducción de Dimensionalidad



$$x^{(i)} \in \mathbb{R}^3$$



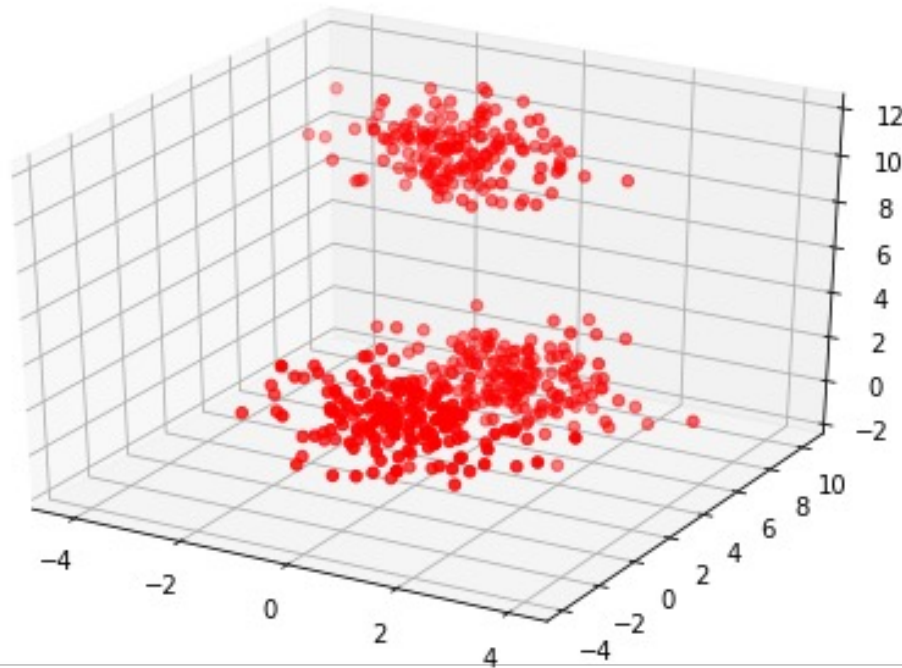
$$z^{(i)} \in \mathbb{R}^2$$



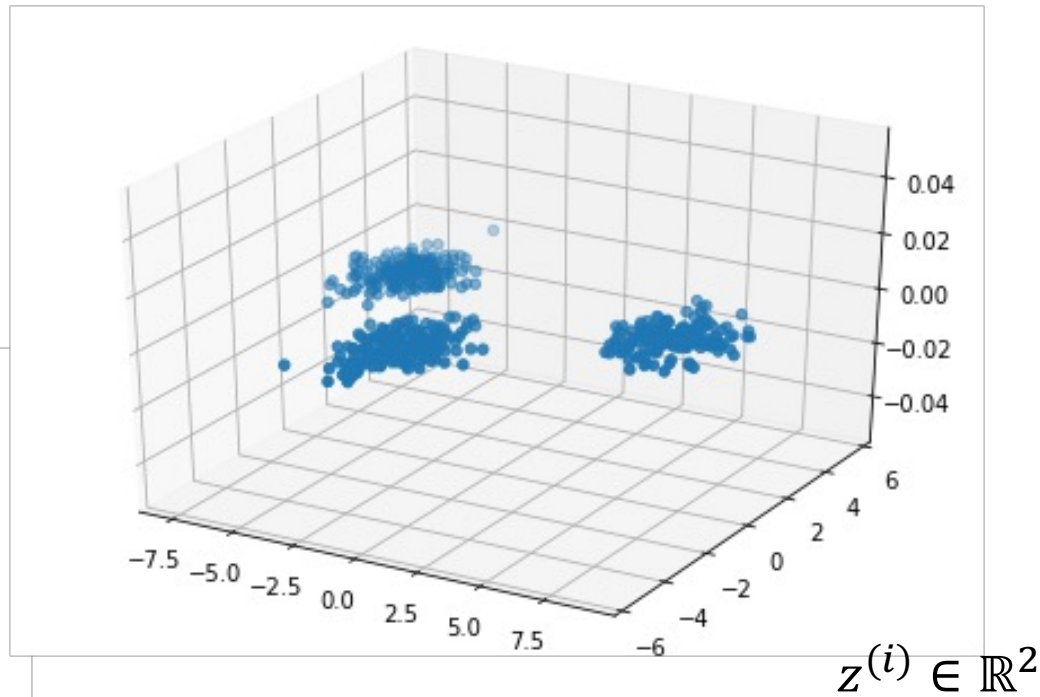
$$z^{(i)} = (z_1^{(i)}, z_2^{(i)})$$

Disminuir de
 \mathbb{R}^3 a \mathbb{R}^2

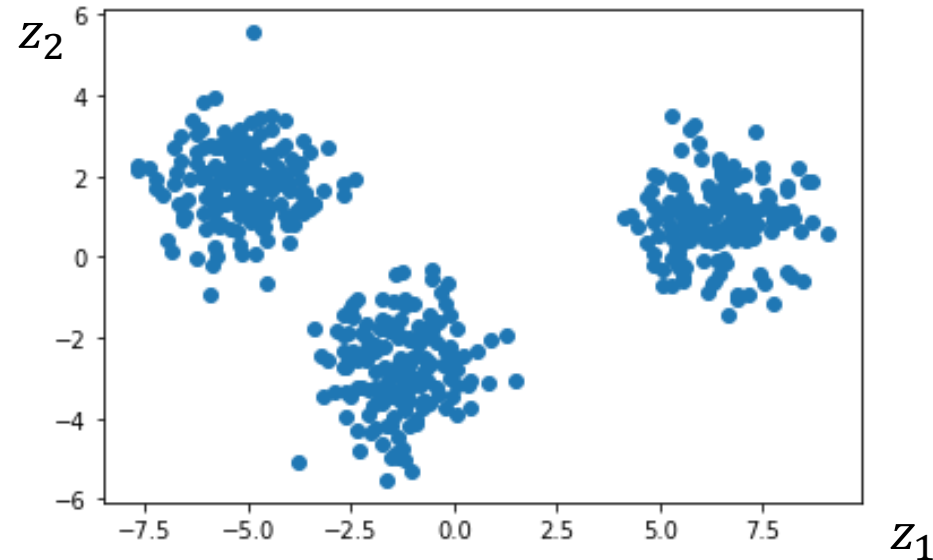
Reducción de Dimensionalidad



$$x^{(i)} \in \mathbb{R}^3$$




$$z^{(i)} \in \mathbb{R}^2$$



$$z^{(i)} = (z_1^{(i)}, z_2^{(i)})$$


Disminuir de
 \mathbb{R}^3 a \mathbb{R}^2



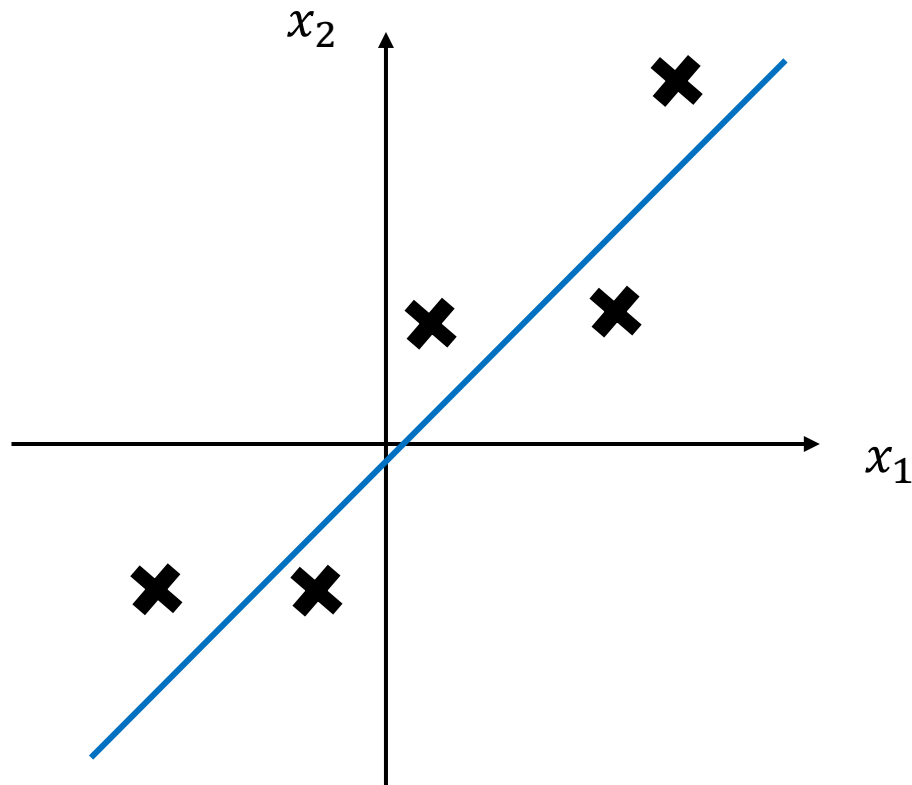
Análisis de Componentes Principales



Análisis de Componentes Principales

- Es una técnica que permite comprimir información.
 - Permite encontrar un nuevo conjunto de variables, de dimensión menor a las originales, que retienen la mayor parte de la información de los datos.
 - Por información se refiere a la variación de los datos.
- 

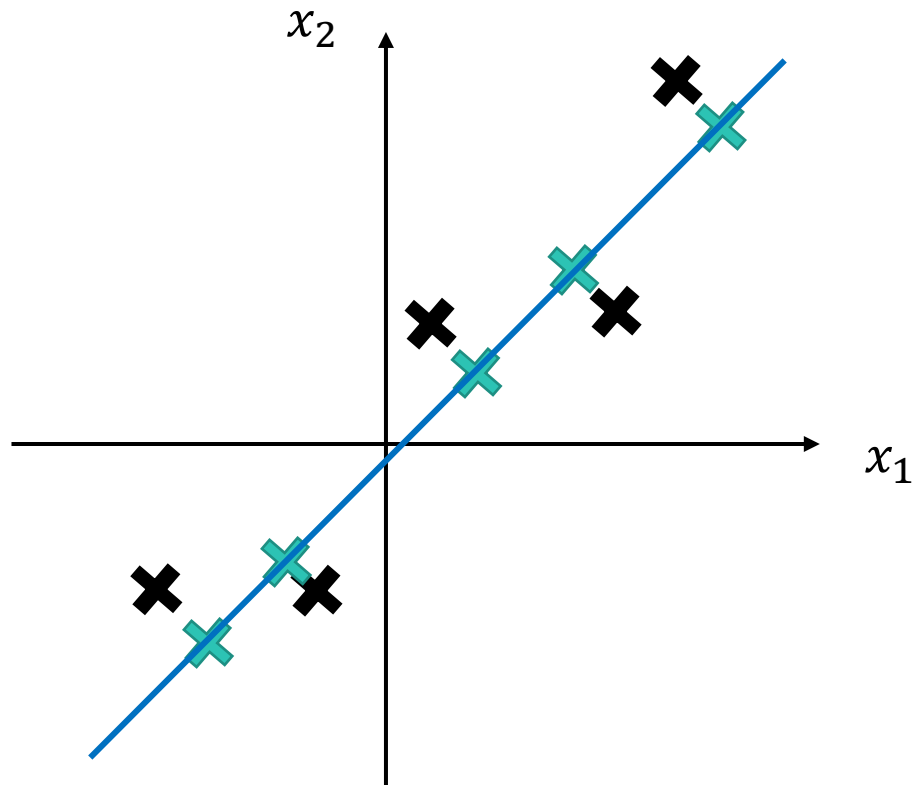
Análisis de Componentes Principales



Tenemos datos en \mathbb{R}^2 y deseamos reducir la dimensionalidad a \mathbb{R} .

Idea principal: determinar la recta en la cual vamos a proyectar los datos.

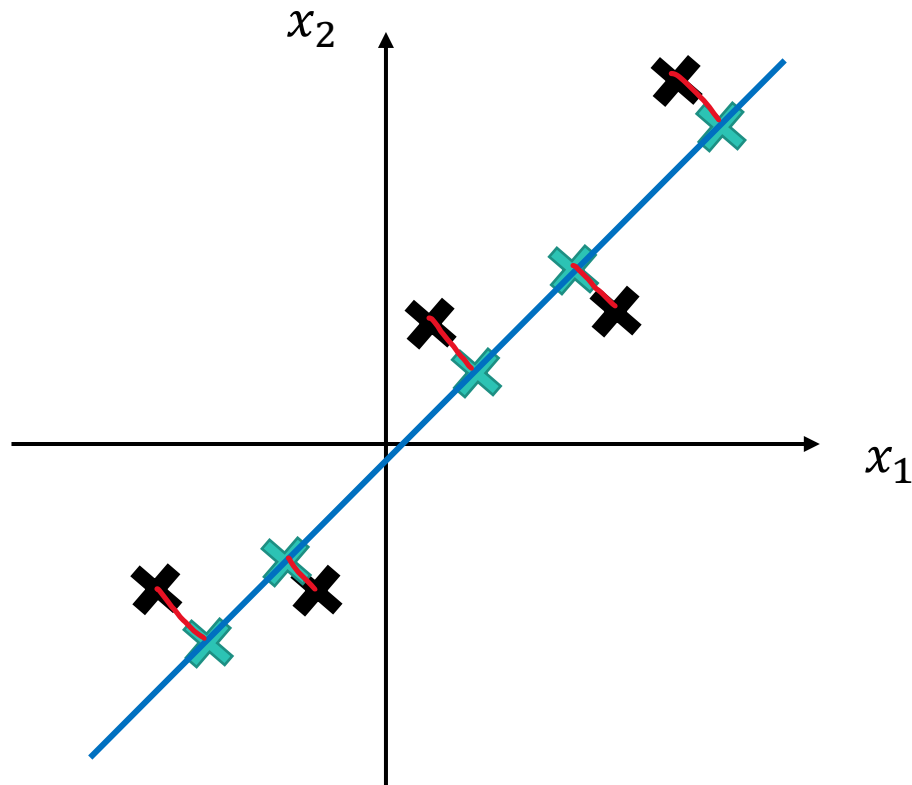
Análisis de Componentes Principales



Tenemos datos en \mathbb{R}^2 y deseamos reducir la dimensionalidad a \mathbb{R} .

Idea principal: determinar la recta en la cual vamos a proyectar los datos.

Análisis de Componentes Principales

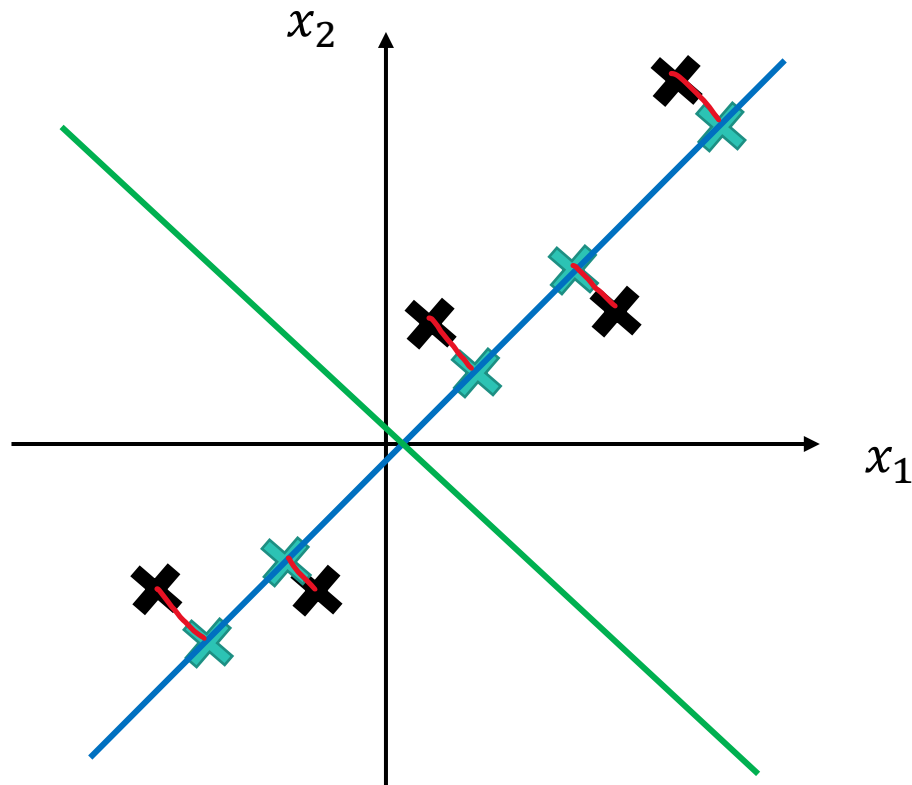


Tenemos datos en \mathbb{R}^2 y deseamos reducir la dimensionalidad a \mathbb{R} .

Idea principal: determinar la recta en la cual vamos a proyectar los datos.

Minimizando la suma cuadrática de las distancias (en rojo) de las proyecciones.

Análisis de Componentes Principales

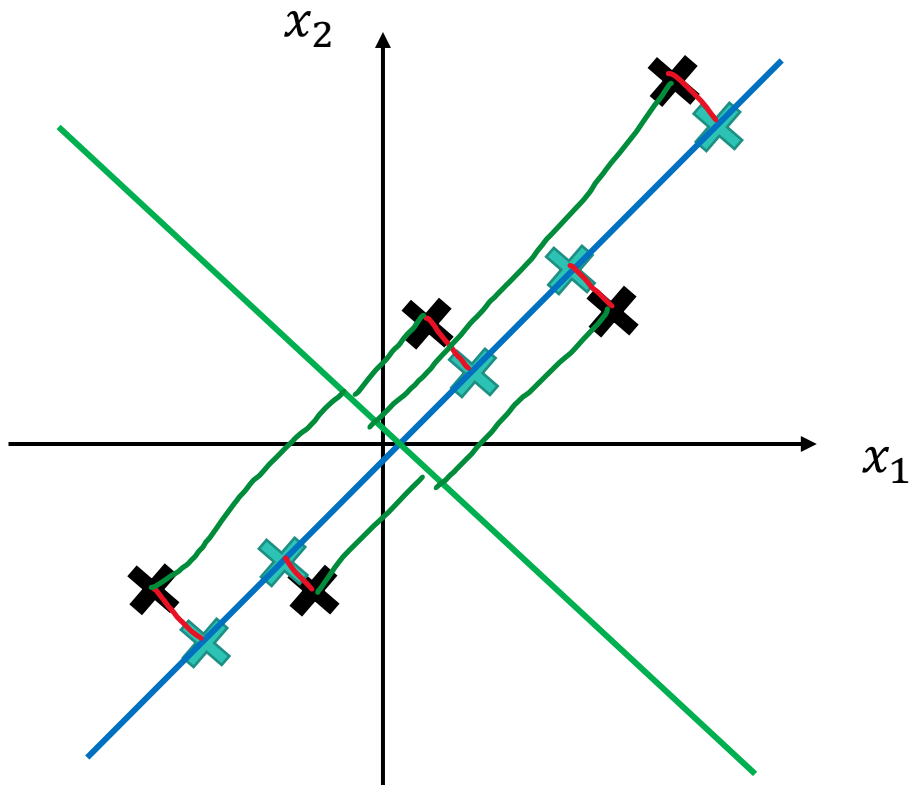


Tenemos datos en \mathbb{R}^2 y deseamos reducir la dimensionalidad a \mathbb{R} .

Idea principal: determinar la recta en la cual vamos a proyectar los datos.

Minimizando la suma cuadrática de las distancias (en rojo) de las proyecciones.

Análisis de Componentes Principales

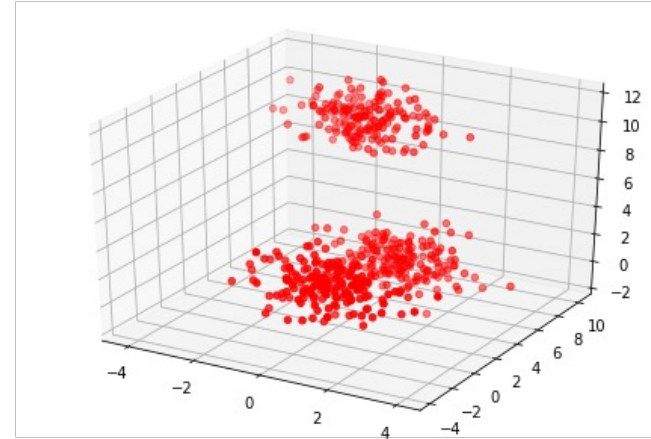
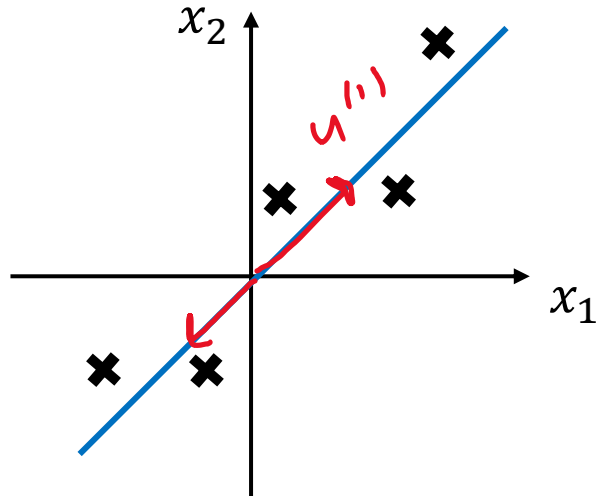


Tenemos datos en \mathbb{R}^2 y deseamos reducir la dimensionalidad a \mathbb{R} .

Idea principal: determinar la recta en la cual vamos a proyectar los datos.

Minimizando la suma cuadrática de las distancias (en rojo) de las proyecciones.

Análisis de Componentes Principales



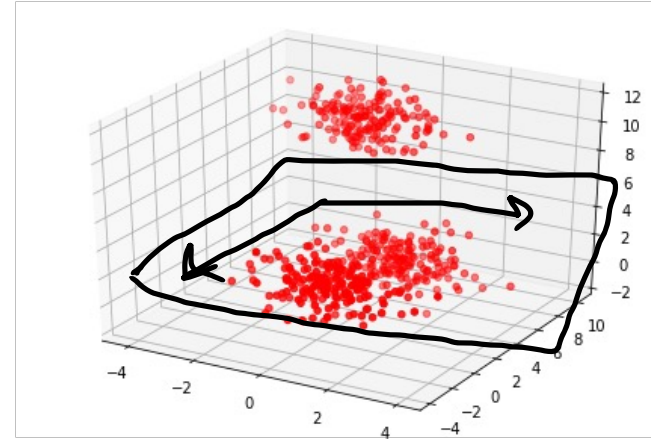
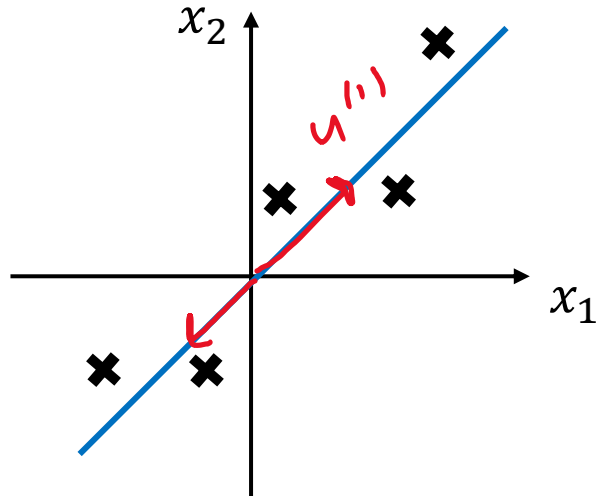
Reducir de \mathbb{R}^2 a \mathbb{R} :

Encontrar un vector $u^{(1)} \in \mathbb{R}^n$ en el cual proyectar los datos de tal manera que se minimice el error de proyección.

Reducir de \mathbb{R}^m a \mathbb{R}^k :

Encontrar k vectores $u^{(1)}, u^{(2)}, \dots, u^{(k)} \in \mathbb{R}^n$ en el cual proyectar los datos de tal manera que se minimice el error de proyección.

Análisis de Componentes Principales



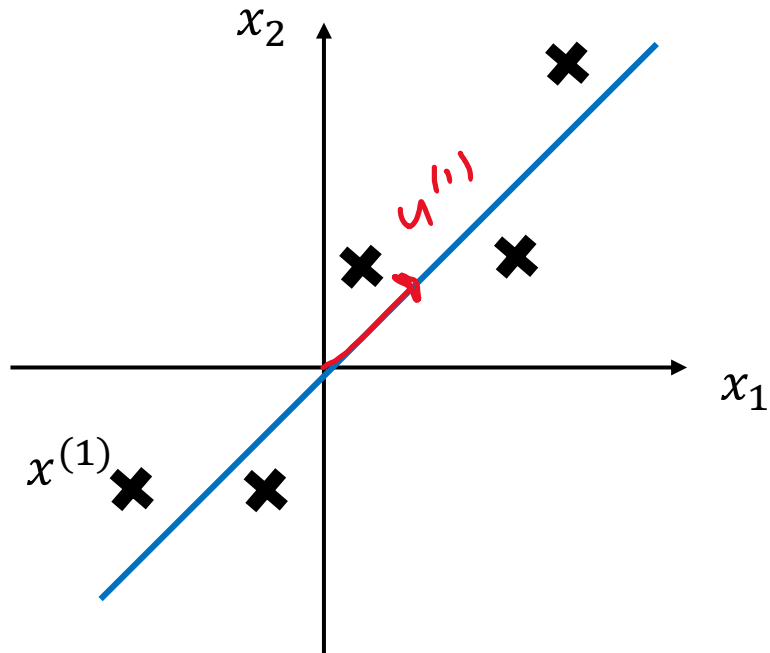
Reducir de \mathbb{R}^2 a \mathbb{R} :

Encontrar un vector $u^{(1)} \in \mathbb{R}^n$ en el cual proyectar los datos de tal manera que se minimice el error de proyección.

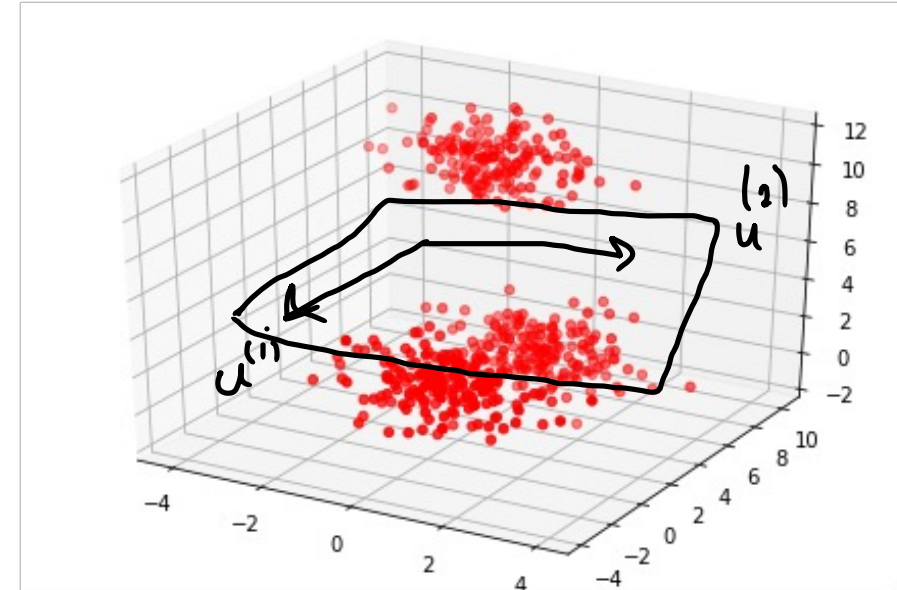
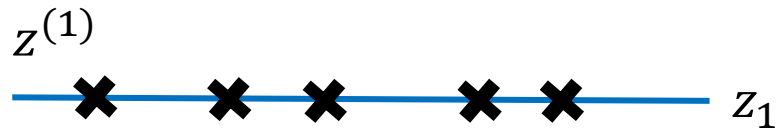
Reducir de \mathbb{R}^m a \mathbb{R}^k :

Encontrar k vectores $u^{(1)}, u^{(2)}, \dots, u^{(k)} \in \mathbb{R}^n$ en el cual proyectar los datos de tal manera que se minimice el error de proyección.

Análisis de Componentes Principales



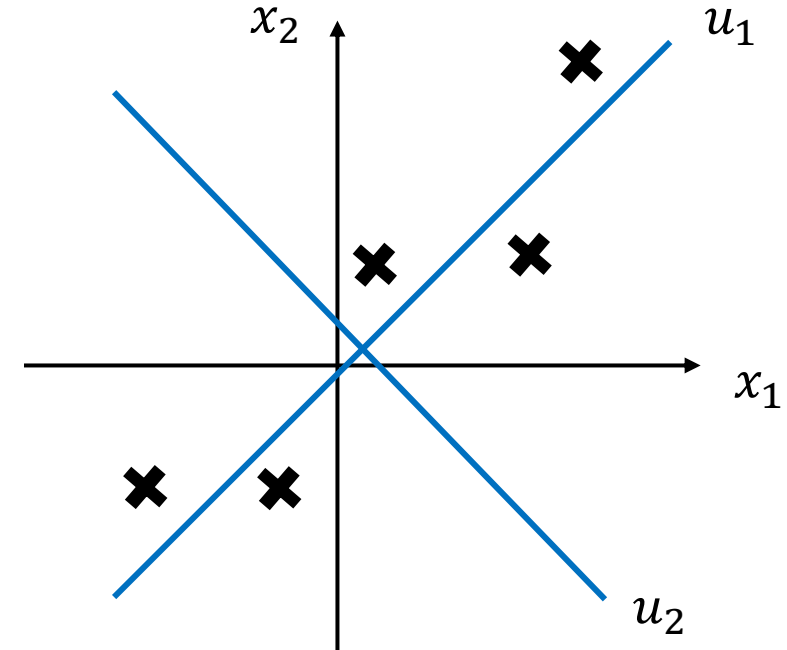
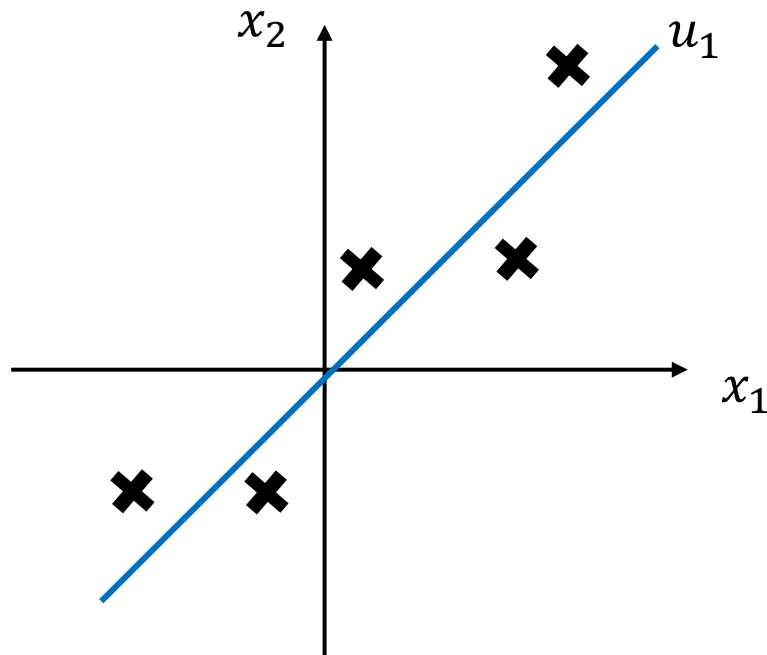
$$x^{(i)} \in \mathbb{R}^2 \rightarrow z^{(i)} \in \mathbb{R}$$



$$x^{(i)} \in \mathbb{R}^3 \rightarrow z^{(i)} \in \mathbb{R}^2$$

Debemos determinar los vectores $u^{(i)}$ y los valores $z^{(i)}$.

Análisis de Componentes Principales



- El **primer componente principal** u_1 es el hiperplano que minimice la distancia cuadrática de los datos.
- El **segundo componente principal** u_2 es el hiperplano perpendicular a el primer componente que minimice la distancia cuadrática de los datos.
- En general, se ajusta un hiperplano que minimice la distancia cuadrático con los datos, cada uno **ortogonal a todos los anteriores**.

Análisis de Componentes Principales

Dado un conjunto de datos $X = (x^{(1)}, \dots, x^{(n)})$, donde $x^{(i)} \in \mathbb{R}^p$, se define **la primer componente principal** como:

$$u_1 \equiv \mathbf{a}_1^T \mathbf{x} = \sum_{i=1}^p \mathbf{a}_{i1} x_i$$

donde el vector $\mathbf{a}_1 = (a_{11}, a_{21}, \dots, a_{p1})$ se elige tal que $\text{var}(u_1)$ **es máxima**.

Análisis de Componentes Principales

El k -ésimo componente principal se define como

donde el vector se elige tal que sujeto a
y, además,

$$u_k \equiv \mathbf{a}_k^T \mathbf{x} \quad k = 1, \dots, p$$

$$\mathbf{a}_k = (a_{1k}, a_{2k}, \dots, a_{pk})$$

$\text{var}(u_1)$ es máxima

$$\text{cov}(u_k, u_l) = 0 \quad \text{para } k > l \geq 1$$

$$\mathbf{a}_k^T \mathbf{a}_k = 1$$

Análisis de Componentes Principales

Para encontrar \mathbf{a}_1 , recordemos que

$$\text{var}(u_1) = E[u_1^2] - E[u_1]^2$$

ya que u_1 es un vector que es una combinación lineal de \mathbf{a}_1 ,

$$E[\mathbf{a}_1^T x] = \mathbf{a}_1^T E[x]$$

y

$$\begin{aligned}\text{var}[u_1] &= E[\mathbf{a}_1^T x x^T \mathbf{a}_1] - E[\mathbf{a}_1^T x] E[\mathbf{a}_1^T x]^T \\ &= \mathbf{a}_1^T E[x x^T] \mathbf{a}_1 - \mathbf{a}_1^T E[x] E[x]^T \mathbf{a}_1 \\ &= \mathbf{a}_1^T (E[x x^T] - E[x] E[x]^T) \mathbf{a}_1 \\ &= \mathbf{a}_1^T \text{Cov}(x) \mathbf{a}_1\end{aligned}$$

Donde $\text{Cov}(x)$ es la matriz de covarianza de los datos $x^{(i)} \in \mathbb{R}^p$.

Análisis de Componentes Principales

Para encontrar \mathbf{a}_1 , se maximiza $\text{var}[u_1]$ sujeto a $\mathbf{a}_1^T \mathbf{a}_1 = 1$. Vamos a resolverlo con multiplicadores de Lagrange. La ecuación que resulta es

$$\mathbf{a}_1^T S \mathbf{a}_1 - \lambda(\mathbf{a}_1^T \mathbf{a}_1 - 1)$$

Actividad: deriven con respecto a \mathbf{a}_1 e igualen a 0.

Análisis de Componentes Principales

Para encontrar \mathbf{a}_1 , se maximiza $\text{var}[u_1]$ sujeto a $\mathbf{a}_1^T \mathbf{a}_1 = 1$. Vamos a resolverlo con multiplicadores de Lagrange. La ecuación que resulta es

$$\mathbf{a}_1^T \mathbf{S} \mathbf{a}_1 - \lambda (\mathbf{a}_1^T \mathbf{a}_1 - 1)$$

Derivando e igualando a cero, se tiene que

$$\mathbf{S} \mathbf{a}_1 - \lambda \mathbf{a}_1 = 0$$

o bien

$$(\mathbf{S} - \lambda \mathbf{I}) \mathbf{a}_1 = 0$$

Es decir, \mathbf{a}_1 es un **eigenvector** de \mathbf{S} correspondiente al eigenvalor $\lambda \equiv \lambda_1$.

Análisis de Componentes Principales

Al maximizar $\text{var}[u_1]$ encontramos que

$$\text{var}[u_1] = \mathbf{a}_1^T \mathbf{S} \mathbf{a}_1 = \mathbf{a}_1^T \lambda_1 \mathbf{a}_1 = \lambda_1$$

Es decir, λ_1 es el mayor eigenvalor de \mathbf{S} .

Análisis de Componentes Principales

Para encontrar el siguiente vector de coeficientes \mathbf{a}_2 , se plantea el siguiente problema:

$$\begin{aligned} &\text{Maximizar } \text{var}[u_2] \\ &\text{sujeto a } \text{cov}[u_2, u_1] = 0 \\ &\quad \mathbf{a}_2^T \mathbf{a}_2 = 1 \end{aligned}$$

Notando que:

$$\text{cov}[u_2, u_1] = \mathbf{a}_1^T \mathbf{S} \mathbf{a}_2 = \lambda_1 \mathbf{a}_1^T \mathbf{a}_2$$

Si λ y φ son multiplicadores de Lagrange, se desea maximizar

$$\mathbf{a}_2^T \mathbf{S} \mathbf{a}_2 - \lambda(\mathbf{a}_2^T \mathbf{a}_2 - 1) - \varphi \mathbf{a}_2^T \mathbf{a}_1$$

Análisis de Componentes Principales

Al resolver el problema de optimización se encuentra que \mathbf{a}_2 también es un eigenvector de \mathbf{S} cuyo eigenvalor $\lambda \equiv \lambda_2$ es el segundo más grande.

En general

$$\text{var}[u_k] = \mathbf{a}_k^T \mathbf{S} \mathbf{a}_k = \lambda_k.$$

- El k -ésimo más grande eigenvalor de \mathbf{S} es la varianza de la k -ésima componente principal.
- La k -ésima componente principal u_k retiene la k -ésima fracción de la variación de los datos.
- El eigenvector asociado permite determinar las proyecciones de los datos.

Análisis de Componentes Principales

Dado una colección de datos $X = (x^{(1)}, \dots, x^{(n)})$, donde $x^{(i)} \in \mathbb{R}^p$ se define un vector de p componentes principales

$$\mathbf{u} = (u_1, u_2, \dots, u_p)$$

de acuerdo a $\mathbf{z} = \mathbf{A}^T \mathbf{x}$, donde $\mathbf{A}^{p \times p}$ es una matriz ortogonal cuya k -ésima columna es el k -ésimo eigenvector \mathbf{a}_k de \mathbf{S} .

Análisis de Componentes Principales

Antes de empezar el procedimiento, siempre hay que **centrar los datos**. Esto es, si la media para la j -ésima característica del vector de datos $x_j^{(i)}$ esta dada por

$$\mu_j = \frac{1}{n} \sum_{i=1}^n x_j^{(i)},$$

se reemplaza $x_j^{(i)}$ con $x_j^{(i)} - \mu_j$.

Análisis de Componentes Principales

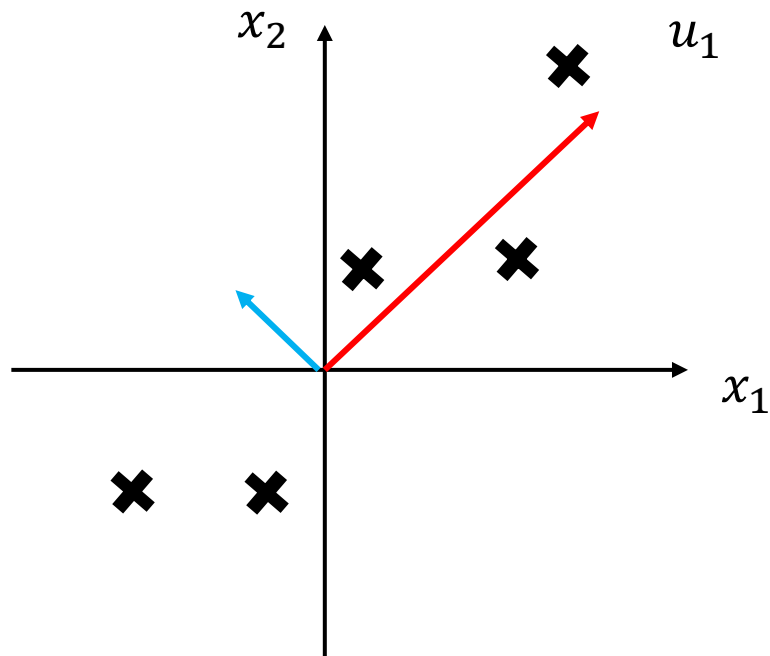
Para el cómputo práctico de PCA:

1. Determinar la matriz de covarianza

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}^T \mathbf{X}$$

2. Se calculan los eigenvectores y eigenvalores de la matriz de covarianza.
3. Se construye la matriz $\mathbf{A}^{n \times k}$. *(Alternativamente, se eligen los k eigenvectores cuyo eigenvalor sea el más grande).*
4. Se calcula la proyección $\mathbf{Z} = \mathbf{XA}$

Análisis de Componentes Principales



$$S = \begin{bmatrix} \text{Var}[X] & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}[Y] \end{bmatrix}$$

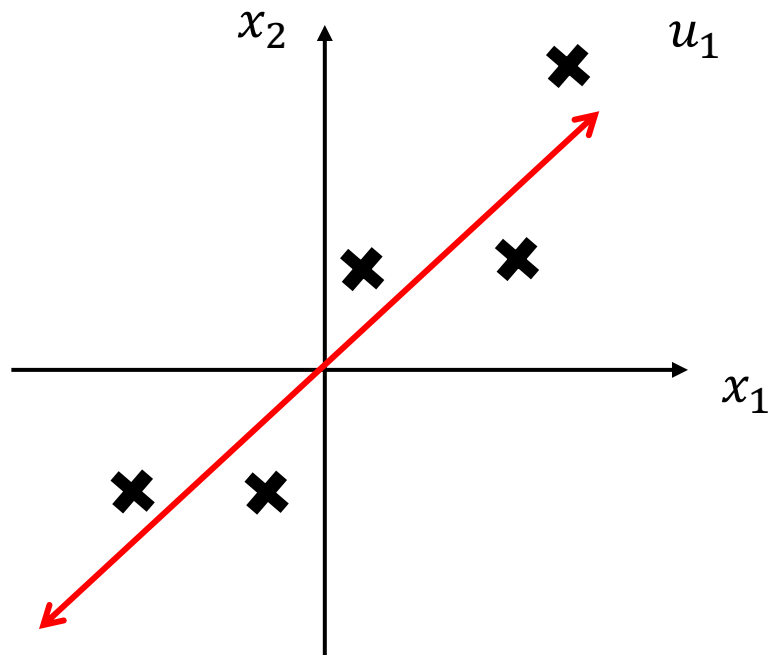
$$S = \begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$$

$$u_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad u_2 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

$$\lambda_1 = 11 \quad \lambda_2 = 1$$

- Los eigenvectores son ortogonales ya que la matriz de covarianza es simétrica.
- Los eigenvectores indican la tendencia de los datos y su dispersión.

Análisis de Componentes Principales



$$S = \begin{bmatrix} \text{Var}[X] & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}[Y] \end{bmatrix}$$

$$S = \begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$$

$$u_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\lambda_1 = 11$$

$$A = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$Z = X \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

- Se elige el eigenvector de mayor magnitud (i.e., el que tenga mayor eigenvalor).
- Se hace la proyección en ese vector.



Implementación Práctica de PCA

¿Cómo se elige k ?

Error de proyección cuadrático medio: $\frac{1}{n} \sum_{i=1}^n \|x^{(i)} - x_{approx}^{(i)}\|^2$

Variación total de la información: $\frac{1}{n} \sum_{i=1}^n \|x^{(i)}\|^2$

Lo más común es elegir k lo más pequeño tal que

$$\frac{\frac{1}{n} \sum_{i=1}^n \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{n} \sum_{i=1}^n \|x^{(i)}\|^2} \leq \alpha$$

Lo cual nos dice que $(1 - \alpha)\%$ de la varianza de los datos se retiene.

¿Cómo se elige k ?

El algoritmo se vería así:

1. Intentar PCA con $k = 1$.
2. Determinar $A, u_1, u_2, \dots, u_k, x_{approx}^{(1)}, \dots, x_{approx}^{(n)}$.
3. Determinar si

$$\frac{\frac{1}{n} \sum_{i=1}^n \|x^{(i)} - x_{approx}\|^2}{\frac{1}{n} \sum_{i=1}^n \|x^{(i)}\|^2} \leq \alpha$$



Aplicando PCA

Uso Primordial – Reducción de Dimensionalidad

- Si tenemos un conjunto de datos $D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$, donde $x^{(i)} \in \mathbb{R}^p$ para p grande...
- se puede aplicar PCA para determinar nuevos $z^{(1)}, \dots, z^{(n)} \in \mathbb{R}^k$, con $k < p$.
- De tal manera que se tiene un conjunto $D' = \{(z^{(1)}, y^{(1)}), \dots, (z^{(n)}, y^{(n)})\}$.
- Después, se puede aplicar cualquier algoritmo de aprendizaje.

Uso Primordial – Reducción de Dimensionalidad

- El fin de aplicar esto es:
 - **Acelerar** los algoritmos de entrenamiento.
 - Es **posible** prevenir *overfitting* (es mejor usar regularización).
- Lo que hace PCA es determinar un mapeo $x^{(i)} \rightarrow z^{(i)}$ (matriz A) que se aplica únicamente con los datos del **conjunto de entrenamiento**.
- Una vez determinado dicho mapeo, es **deseable** aplicar el mismo a los datos del conjunto de prueba y/o validación.

PCA no es un paso obligatorio

Diseño de un sistema de ML:

1. Obtener datos $D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$.
2. Aplicar PCA.
3. Entrenar un modelo con $D' = \{(z^{(1)}, y^{(1)}), \dots, (z^{(n)}, y^{(n)})\}$.
4. Evaluar en el conjunto de prueba o con validación cruzada.
5. Aceptar o modificar algunos de los pasos anteriores.

Es buena idea probar con y sin PCA. Si no jala el sistema como se planea, puede ser buena idea recurrir a PCA.



¡Gracias!

Luis Zúñiga

[Correo: luis.zuniga@correo.uia.mx](mailto:luis.zuniga@correo.uia.mx)

Sitio web: <https://lzun.github.io>