

# Reducción de dimensionalidad

Dr. Mauricio Toledo-Acosta  
`mauricio.toledo@unison.mx`

Diplomado Ciencia de Datos con Python

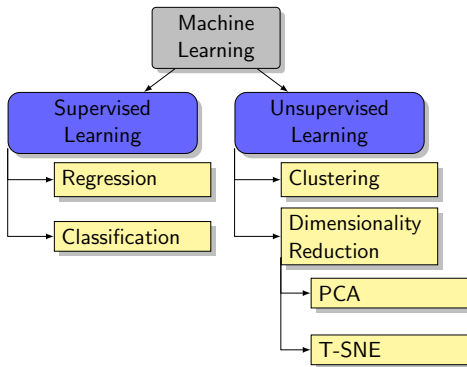
# Table of Contents

1 Introducción

2 PCA

3 t-SNE

# Introducción



# Reducción de dimensionalidad

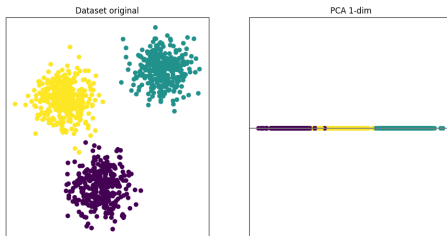
## Reducción de dimensionalidad

La reducción de la dimensionalidad es la transformación de los datos de un espacio de alta dimensión a un espacio de baja dimensión, de manera que la representación de baja dimensión conserve propiedades significativas de los datos originales.

# Reducción de dimensionalidad

## Reducción de dimensionalidad

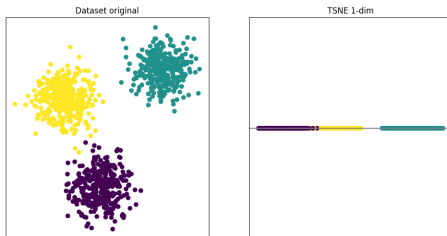
La reducción de la dimensionalidad es la transformación de los datos de un espacio de alta dimensión a un espacio de baja dimensión, de manera que la representación de baja dimensión conserve propiedades significativas de los datos originales.



# Reducción de dimensionalidad

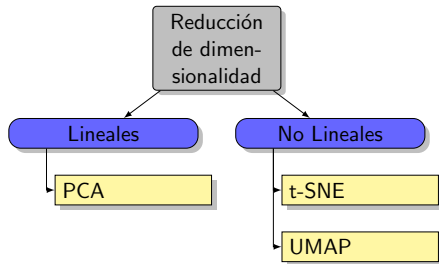
## Reducción de dimensionalidad

La reducción de la dimensionalidad es la transformación de los datos de un espacio de alta dimensión a un espacio de baja dimensión, de manera que la representación de baja dimensión conserve propiedades significativas de los datos originales.



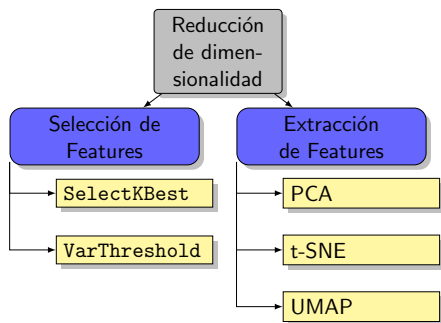
# Clasificación de los Métodos

Desde un punto de vista matemático, se trata de transformar los datos de un espacio de alta dimensión a un espacio de baja dimensión.



# Clasificación de los Métodos

Desde un punto de vista computacional, se trata de reducir el número de features que definen a los datos para hacer más tratables las tareas del Machine Learning.





# Utilidad

## Ventajas

- Puede ser más fácil visualizar los datos.

# Utilidad

## Ventajas

- Puede ser más fácil visualizar los datos.
- Extraer la información más importante de los datos.

# Utilidad

## Ventajas

- Puede ser más fácil visualizar los datos.
- Extraer la información más importante de los datos.
- Obtener features para fines de clasificación.

# Utilidad

## Ventajas

- Puede ser más fácil visualizar los datos.
- Extraer la información más importante de los datos.
- Obtener features para fines de clasificación.
- Menos espacio de almacenamiento.

# Utilidad

## Ventajas

- Puede ser más fácil visualizar los datos.
- Extraer la información más importante de los datos.
- Obtener features para fines de clasificación.
- Menos espacio de almacenamiento.
- Menos features requieren menos tiempo de cálculo (entrenamiento más rápido).

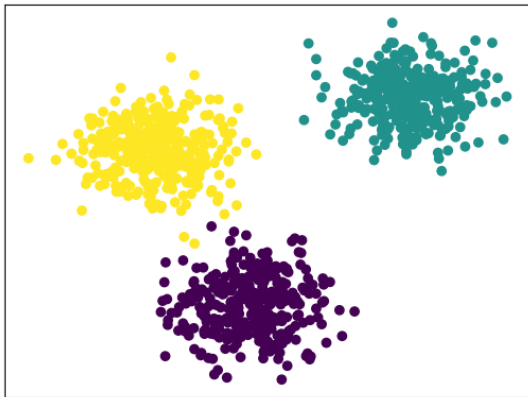
# Utilidad

## Ventajas

- Puede ser más fácil visualizar los datos.
- Extraer la información más importante de los datos.
- Obtener features para fines de clasificación.
- Menos espacio de almacenamiento.
- Menos features requieren menos tiempo de cálculo (entrenamiento más rápido).
- Menos features significan menos complejidad del modelo.

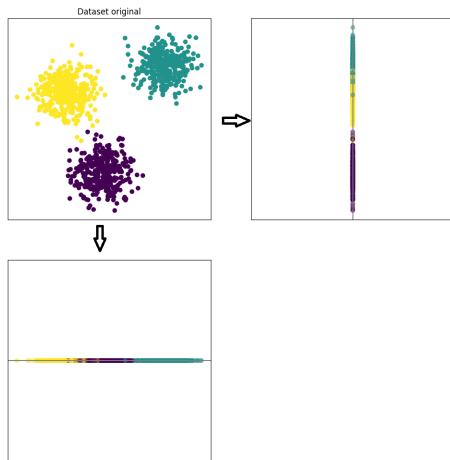
# Feature Selection

Una forma de reducción de dimensionalidad es la selección de features.



# Feature Selection

Una forma de reducción de dimensionalidad es la selección de features.





# Table of Contents

1 Introducción

2 PCA

3 t-SNE

# PCA

Introducido en 1901 por Karl Pearson (1857-1936).

Jonathon Shlens (2014). A tutorial on Principal Component Analysis.

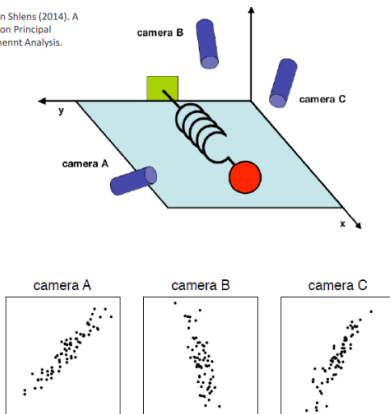


FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

# PCA

PCA puede pensarse como el **ajuste de un elipsoide  $D$ -dimensional al conjunto de datos**, donde cada eje del elipsoide representa una componente principal. Si algún eje del elipsoide es pequeño, entonces la varianza a lo largo de ese eje también es pequeña.

# PCA

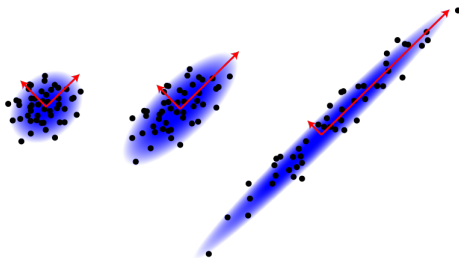
PCA puede pensarse como el **ajuste de un elipsoide  $D$ -dimensional al conjunto de datos**, donde cada eje del elipsoide representa una componente principal. Si algún eje del elipsoide es pequeño, entonces la varianza a lo largo de ese eje también es pequeña.

PCA **transforma los datos a un nuevo sistema de coordenadas** de tal manera que la mayor varianza se sitúa en la primera coordenada, la segunda mayor varianza en la segunda coordenada, y así sucesivamente.

# PCA

PCA puede pensarse como el **ajuste de un elipsoide  $D$ -dimensional al conjunto de datos**, donde cada eje del elipsoide representa una componente principal. Si algún eje del elipsoide es pequeño, entonces la varianza a lo largo de ese eje también es pequeña.

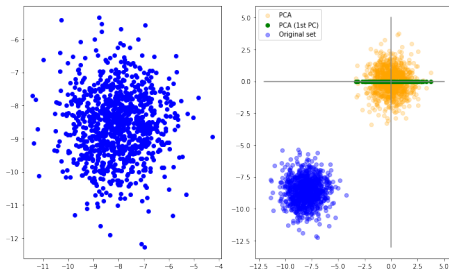
PCA transforma los datos a un nuevo sistema de coordenadas de tal manera que la mayor varianza se sitúa en la primera coordenada, la segunda mayor varianza en la segunda coordenada, y así sucesivamente.



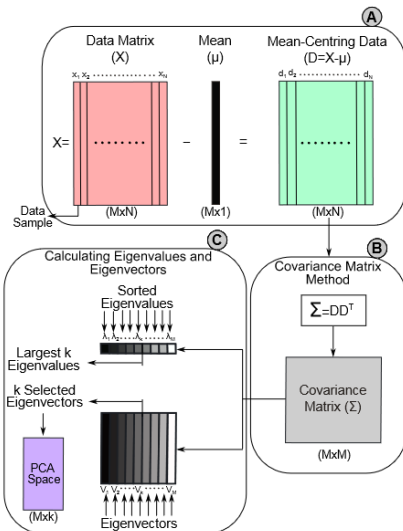
# PCA

PCA puede pensarse como el ajuste de un elipsoide  $D$ -dimensional al conjunto de datos, donde cada eje del elipsoide representa una componente principal. Si algún eje del elipsoide es pequeño, entonces la varianza a lo largo de ese eje también es pequeña.

PCA transforma los datos a un nuevo sistema de coordenadas de tal manera que la mayor varianza se sitúa en la primera coordenada, la segunda mayor varianza en la segunda coordenada, y así sucesivamente.



# El método



Cambio de base

Matriz de covarianza

Matriz de covarianza - Ejemplo

Eigenvectores y eigenvalores

Diagonalizar la matriz de covarianza

La proyección es

$$X_{pca} = X \cdot P$$

$P$  es la matriz de los  $k$  eigenvectores columna con norma más grande.

# Ventajas y Desventajas

## Ventajas

- Permite eliminar variables correlacionadas.
- Permite la visualización de datos.
- Puede ayudar a reducir el overfitting.



# Ventajas y Desventajas

## Ventajas

- Permite eliminar variables correlacionadas.
- Permite la visualización de datos.
- Puede ayudar a reducir el overfitting.

## Desventajas

- Suele requiere de un escalamiento de datos antes.
- Se pierde información.
- Perdemos la interpretabilidad de las variables de entrada.

# Table of Contents

1 Introducción

2 PCA

3 t-SNE

# T-SNE

Desarrollado en 2008 por Laurens van der Maaten (FAIR) y Geoffery Hinton (University of Toronto). Significa **t-distributed Stochastic Neighbor Embedding**.

# T-SNE

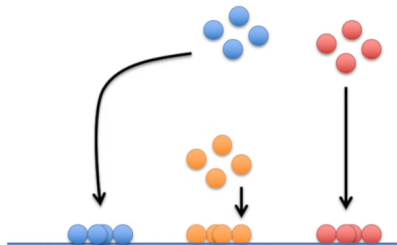
Desarrollado en 2008 por Laurens van der Maaten (FAIR) y Geoffery Hinton (University of Toronto). Significa **t-distributed Stochastic Neighbor Embedding**.

T-SNE modela cada objeto de alta dimensión mediante un punto bidimensional o tridimensional de tal forma que los objetos similares se modelan mediante puntos cercanos y los objetos no similares se modelan mediante puntos distantes con una alta probabilidad.

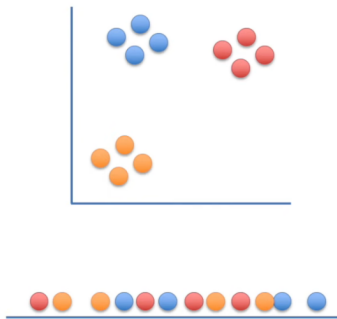
# T-SNE: El proceso

- 1 Construir una distribución de probabilidad sobre pares de objetos en alta dimensión de tal manera que a los objetos similares se les asigna una probabilidad más alta mientras que a los puntos disímiles se les asigna una probabilidad más baja.
- 2 Definir una distribución de probabilidad similar sobre los puntos del mapa de baja dimensión, y minimiza la divergencia de Kullback-Leibler entre las dos distribuciones con respecto a las ubicaciones de los puntos en el mapa.

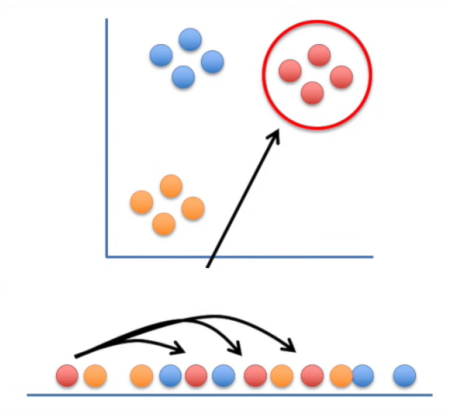
# El proceso: Versión 1



# El proceso: Versión 1

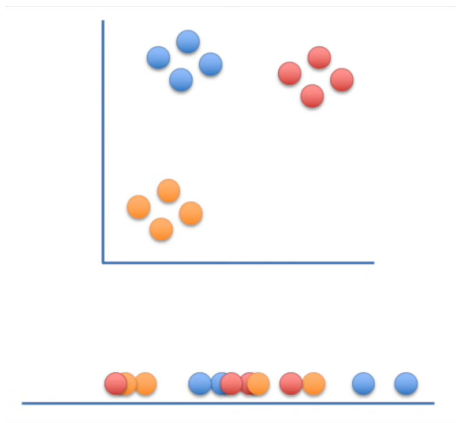


# El proceso: Versión 1

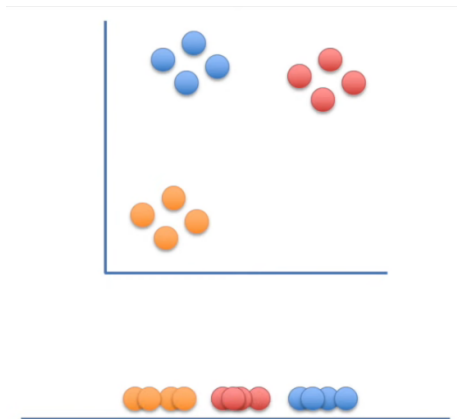




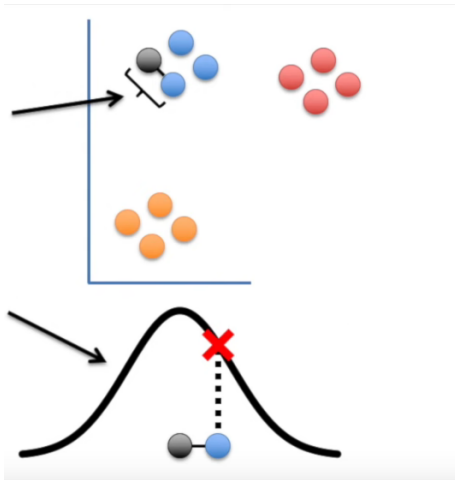
# El proceso: Versión 1



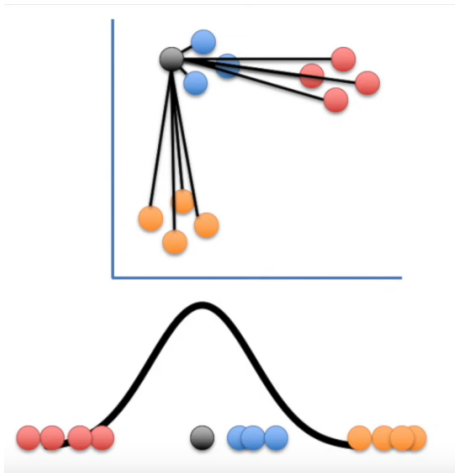
# El proceso: Versión 1



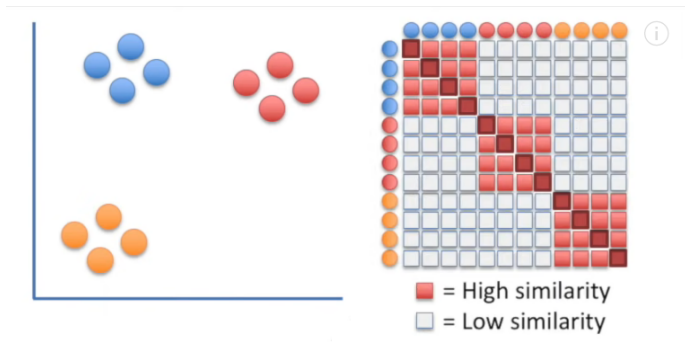
# El proceso: Versión 2



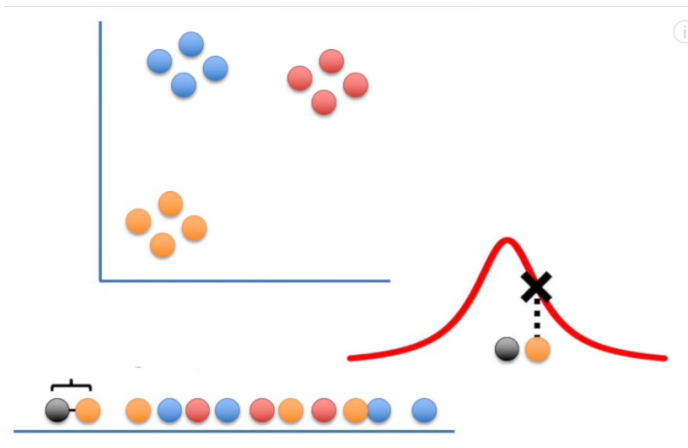
# El proceso: Versión 2



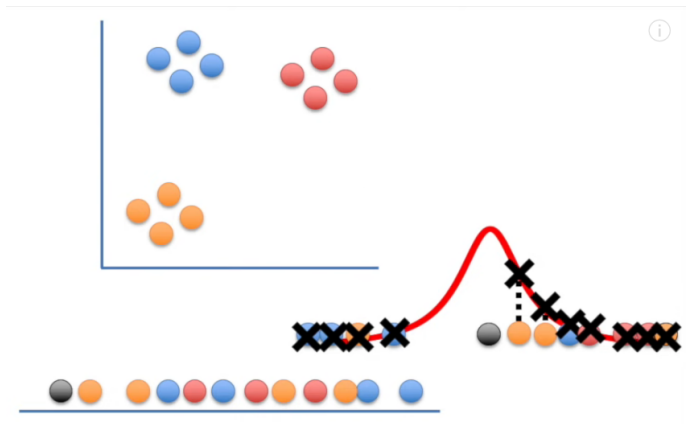
# El proceso: Versión 2



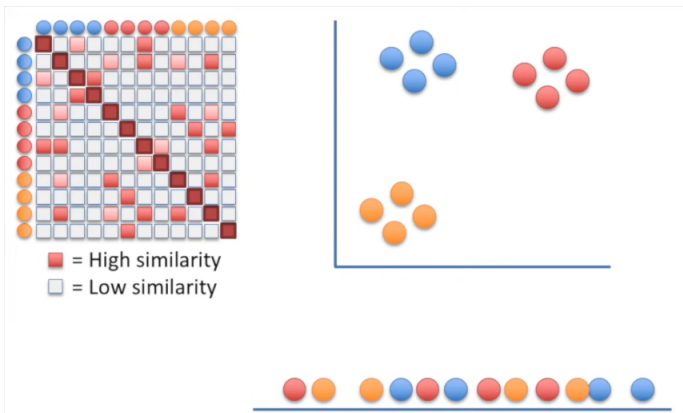
# El proceso: Versión 2



# El proceso: Versión 2

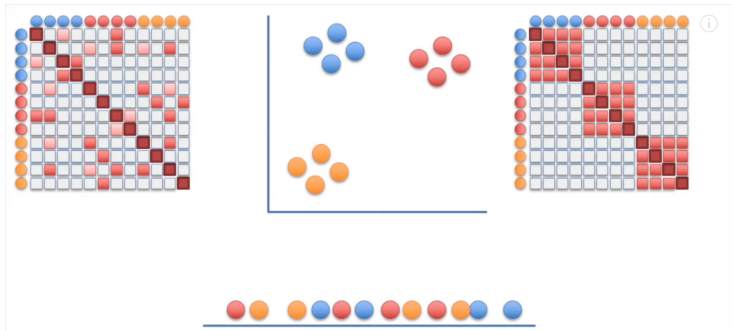


# El proceso: Versión 2

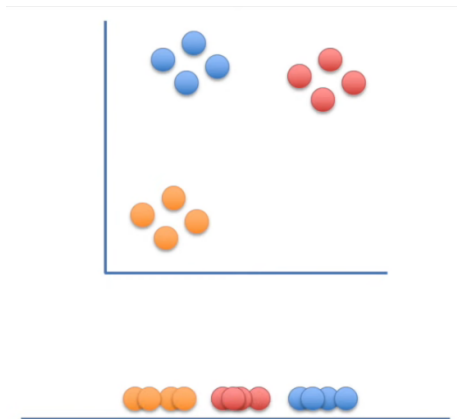




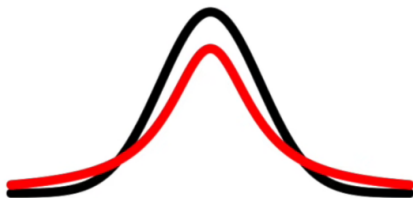
# El proceso: Versión 2



# El proceso: Versión 2



## El proceso: Versión 2



La distribución  $t$  de Student tiene como función de densidad

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

donde

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

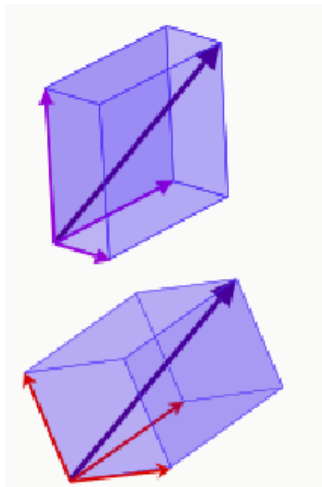
# Consideraciones sobre t-SNE

- t-SNE es computacionalmente caro y relativamente lento (por ejemplo, hay que calcular las distancias entre parejas de puntos).
- Tiene la ventaja sobre PCA de funcionar bien en datos que no están relacionados linealmente.
- En general, PCA preserva la estructura global de los datos, mientras que t-SNE preserva las estructuras locales.
- Suele ser una buena opción para visualizar datos.

# Table of Contents

## 4 Appendix

# Cambios de Base



- Las coordenadas de un punto en son los coeficientes de los vectores canónicos unitarios

$$e_1 = (1, 0, \dots, 0),$$

...

$$e_D = (0, 0, \dots, 1).$$

- Todo punto puede ser expresado en una infinidad de bases.
- Para cambiar de base hay que multiplicar el vector por la matriz

$$\begin{pmatrix} v_1^{(1)} & \dots & v_1^{(D)} \\ \vdots & \ddots & \vdots \\ v_D^{(1)} & \dots & v_D^{(D)} \end{pmatrix}$$

# Matriz de Covarianza

¿Cómo obtenemos la nueva base de coordenadas para PCA? Esta base de vectores deben ser las direcciones de máxima varianza, estas son las **componentes principales**.

PCA

# Matriz de Covarianza

¿Cómo obtenemos la nueva base de coordenadas para PCA? Esta base de vectores deben ser las direcciones de máxima varianza, estas son las **componentes principales**.

Consideremos la matriz de covarianza

$$C_X = \frac{1}{N} X \cdot X^T.$$

PCA



# Matriz de Covarianza

¿Cómo obtenemos la nueva base de coordenadas para PCA? Esta base de vectores deben ser las direcciones de máxima varianza, estas son las **componentes principales**.

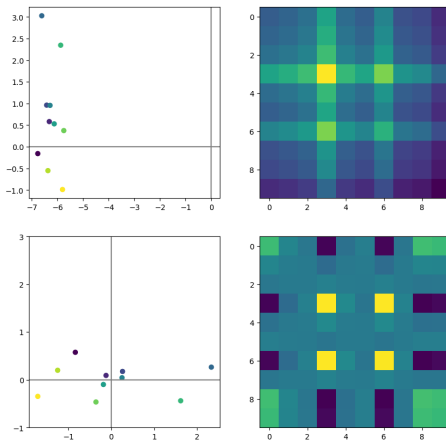
Consideremos la matriz de covarianza

$$C_X = \frac{1}{N} X \cdot X^T.$$

- Los términos altos en la diagonal corresponden a una varianza alta.
- Los términos altos fuera de la diagonal corresponden a una redundancia alta.

PCA

# Matriz de covarianza



PCA

# Eigenvalores y Eigenvectores

Cada matriz  $M$  de tamaño  $n \times n$  se puede ver como una transformación del espacio  $\mathbb{R}^n$  en el espacio  $\mathbb{R}^n$ . Es decir, toma un punto  $p \in \mathbb{R}^n$  y devuelve otro vector  $M \cdot p \in \mathbb{R}^n$ . ¿Cómo es este punto respecto al inicial?

PCA

# Eigenvalores y Eigenvectores

Cada matriz  $M$  de tamaño  $n \times n$  se puede ver como una transformación del espacio  $\mathbb{R}^n$  en el espacio  $\mathbb{R}^n$ . Es decir, toma un punto  $p \in \mathbb{R}^n$  y devuelve otro vector  $M \cdot p \in \mathbb{R}^n$ . ¿Cómo es este punto respecto al inicial?

Hay vectores especiales, dependiendo de  $M$ , que son transformados en un múltiplo de ellos mismos.

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$
$$A \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1 \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$
$$A \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 3 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

# Eigenvalores y Eigenvectores

Cada matriz  $M$  de tamaño  $n \times n$  se puede ver como una transformación del espacio  $\mathbb{R}^n$  en el espacio  $\mathbb{R}^n$ . Es decir, toma un punto  $p \in \mathbb{R}^n$  y devuelve otro vector  $M \cdot p \in \mathbb{R}^n$ . ¿Cómo es este punto respecto al inicial?

Hay vectores especiales, dependiendo de  $M$ , que son transformados en un múltiplo de ellos mismos.

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$
$$A \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1 \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$
$$A \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 3 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

# Diagonalizar la matriz de covarianza

De acuerdo con lo anterior, para nuestra matriz de covarianza  $C_X$  deseamos:

- Minimizar la redundancia, medida por la magnitud de la covarianza.
- Maximizar la varianza.

PCA

# Diagonalizar la matriz de covarianza

De acuerdo con lo anterior, para nuestra matriz de covarianza  $C_X$  deseamos:

- Minimizar la redundancia, medida por la magnitud de la covarianza.
- Maximizar la varianza.

Esto, en términos de matrices, quiere decir **Diagonalizar**. Es decir, encontrar matrices  $P$  y  $D$  tales que

$$D = P \cdot C_X \cdot P^T$$

PCA

# Diagonalizar la matriz de covarianza

De acuerdo con lo anterior, para nuestra matriz de covarianza  $C_X$  deseamos:

- Minimizar la redundancia, medida por la magnitud de la covarianza.
- Maximizar la varianza.

Esto, en términos de matrices, quiere decir **Diagonalizar**. Es decir, encontrar matrices  $P$  y  $D$  tales que

$$D = P \cdot C_X \cdot P^T$$

Esto se hace encontrando los eigenvector y eigenvalores de  $C_X$ .

PCA