Ingeniería de atributos

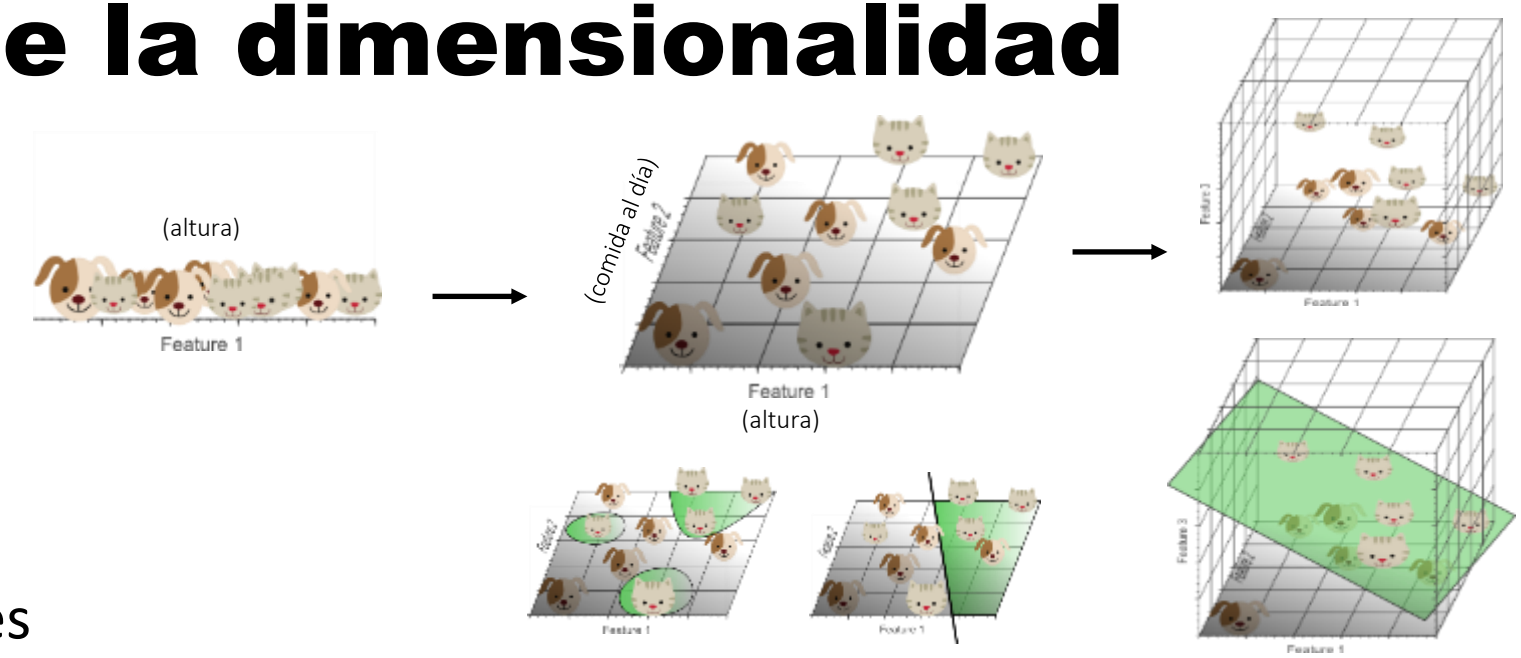
Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

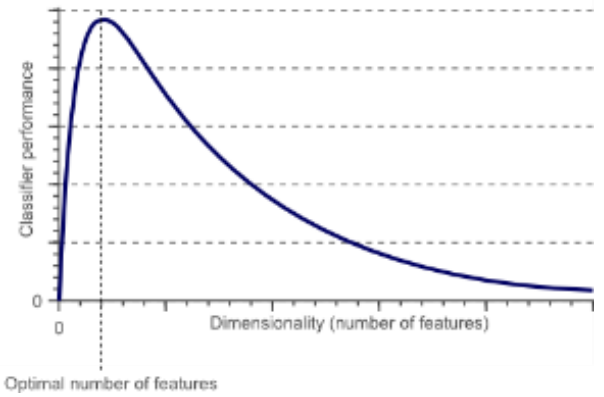
@raimun2

La maldición de la dimensionalidad

La maldición de la dimensionalidad



- Una cantidad correcta de atributos ayudan a crear mejores modelos.
- Los datos de altas dimensiones se vuelven cada vez más dispersos en su espacio.
- Las definiciones de densidad y distancia entre puntos se vuelven menos significativas a mayor numero de atributos.



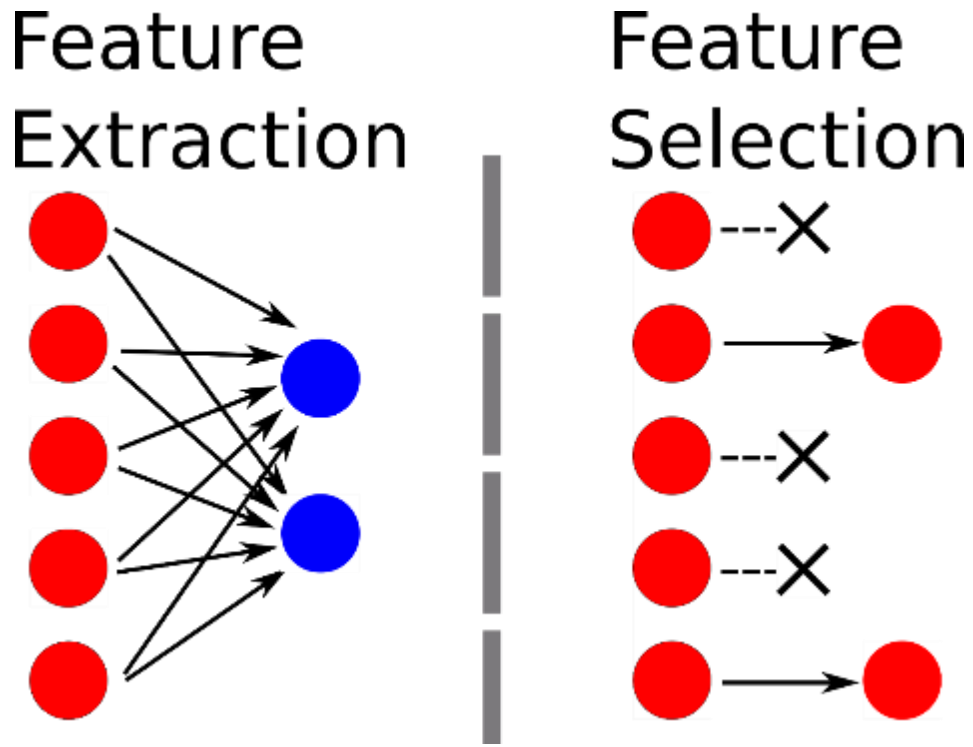
Selección o extracción de atributos

Selección de atributos:

- **Selección de un subconjunto de atributos según algún criterio específico.**

Extracción de atributos:

- **Creación de nuevos atributos a partir de atributos originales**



Pueden hacerse con conocimiento del dominio o algorítmicamente

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} \rightarrow \begin{bmatrix} Y_1 \\ \vdots \\ Y_d \end{bmatrix} = f\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix}\right) \qquad \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} \rightarrow \begin{bmatrix} X_{i_1} \\ \vdots \\ X_{i_d} \end{bmatrix}$$

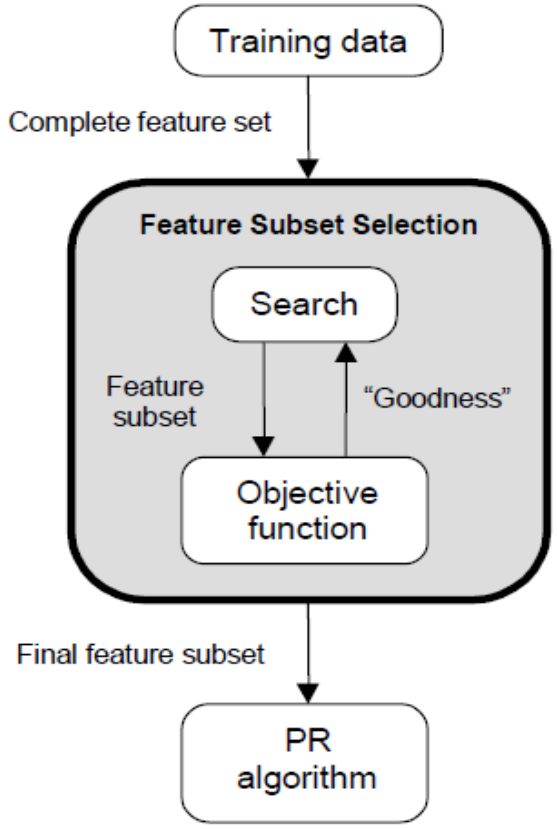
Objetivos

- Mejorar el desempeño de los modelos:
    - Poder predictivo
    - Complejidad
    - Tiempo de ejecución
- Visualizar los datos
- Eliminar el ruido

Selección de atributos
(Feature selection)

Estrategias de selección

Tenemos que seleccionar los atributos con el mejor desempeño



Estrategias de búsqueda:

- Fuerza bruta: búsqueda por todo el espacio.
- Heurísticas: busca en un subconjunto del espacio usando una estrategia.
  - Ranking: seleccione las mejores variables sin tener en cuenta sus interacciones.
  - Generación secuencial hacia adelante / atrás
  - Filtro: selección de variables independientemente del modelo/problema.
  - Wrapper: selección variable basada en el rendimiento del modelo.
  - Embedded: selección de variables incluida en el proceso de entrenamiento del modelo.

Fuerza bruta

Suponiendo **m** atributos, una búsqueda de fuerza bruta implica:

1. Para cada valor de **d** entre 0 y **m**
2. Busca por todas los $\binom{m}{d}$ posibles subconjuntos de tamaño **d**.
3. Evaluar todos los subconjuntos con algún criterio de desempeño
4. Seleccione el subconjunto con mejor rendimiento el criterio definido.

El número de subconjuntos crece exponencialmente.

Ranking

Suponiendo **m** atributos, una búsqueda de fuerza bruta implica:

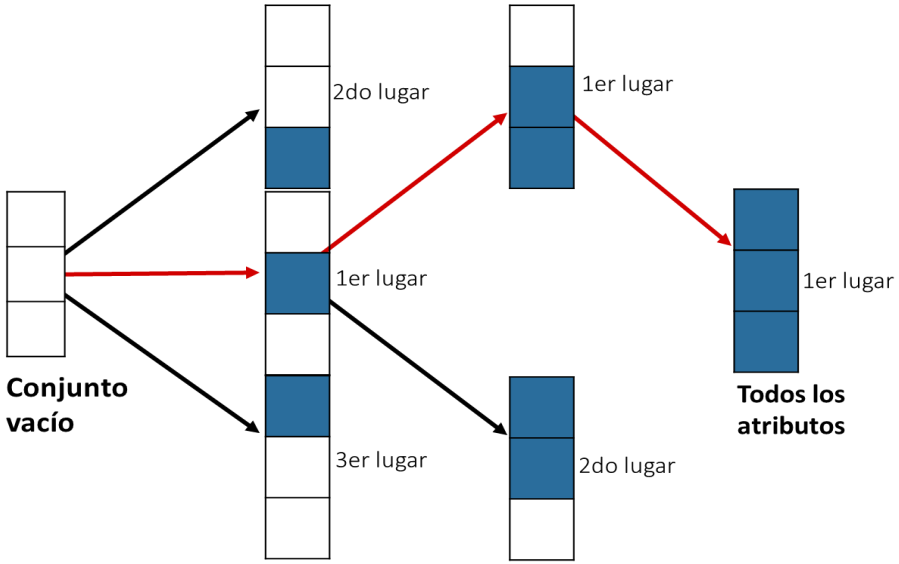
1. Para cada valor de **d** entre 1 y **m**
2. Evaluar cada variable con algún criterio de desempeño
3. Seleccione el subconjunto d de variables con mayores valores según el criterio definido

El número de subconjuntos crece linealmente.
No analiza la interacción entre variables
Se puede aplicar un umbral para seleccionar los atributos

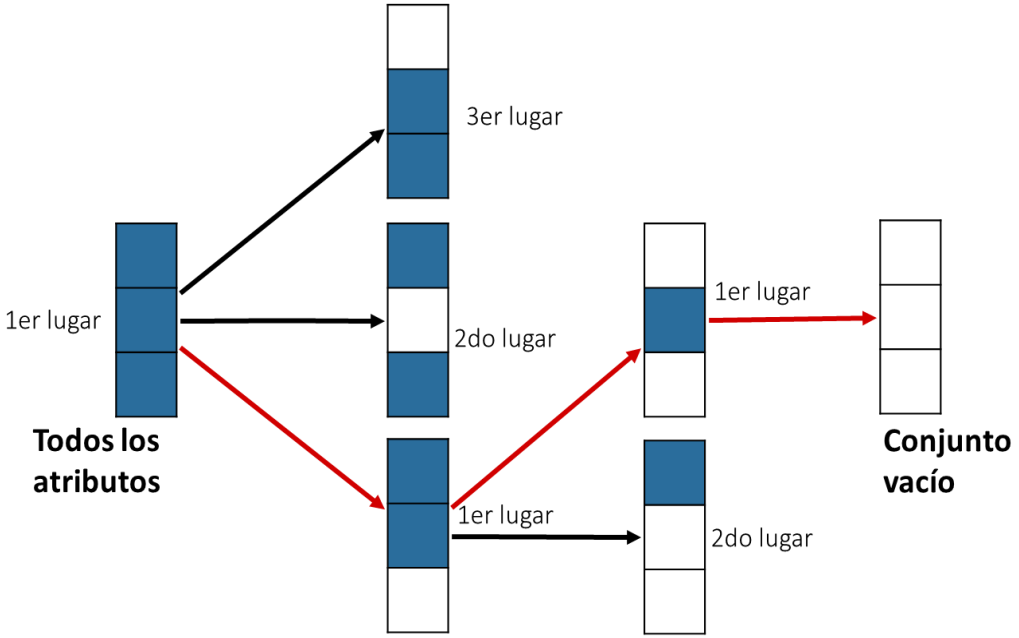
Generación secuencial

- Heurística que agrega / quita variables secuencialmente según variaciones en el desempeño de un criterio definido.
- Puede recibir como parámetro el número de variables que se espera obtener

Generación secuencial hacia adelante



Generación secuencial hacia atrás



Otras Heurísticas

Método de filtro

- La evaluación se basa en la información contenida en el subconjunto de atributos. Se mantienen atributos que pasen el filtro de información.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

Método Wrapper

- La evaluación toma en cuenta el modelo que se aplicará al subconjunto seleccionado de variables/entidades.
- El criterio de selección se se basa en el poder predictivo del modelo con subconjunto de variables.
- De los métodos mas utilizados.

$$CV = \frac{\sigma}{\mu}$$

$$H(X) = -\sum_{i=1}^{N_v} P(x_i)\log_2 P(x_i)$$

Método Embedded

- La evaluación tiene en cuenta el desempeño de los atributos en el modelo.
- Normalmente, la función objetivo penaliza el número de variables

Extracción de atributos
(Feature extraction)

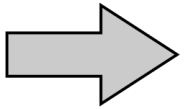
Reducción de dimensionalidad

Objetivos:

- Reducir la cantidad de tiempo y memoria que requieren los algoritmos de minería de datos
- Facilitar visualización.
- Puede ayudar a eliminar funciones irrelevantes o reducir el ruido

Generación de **k** nuevos atributos a partir de **m** atributos originales donde los nuevos atributos son combinaciones de los originales.

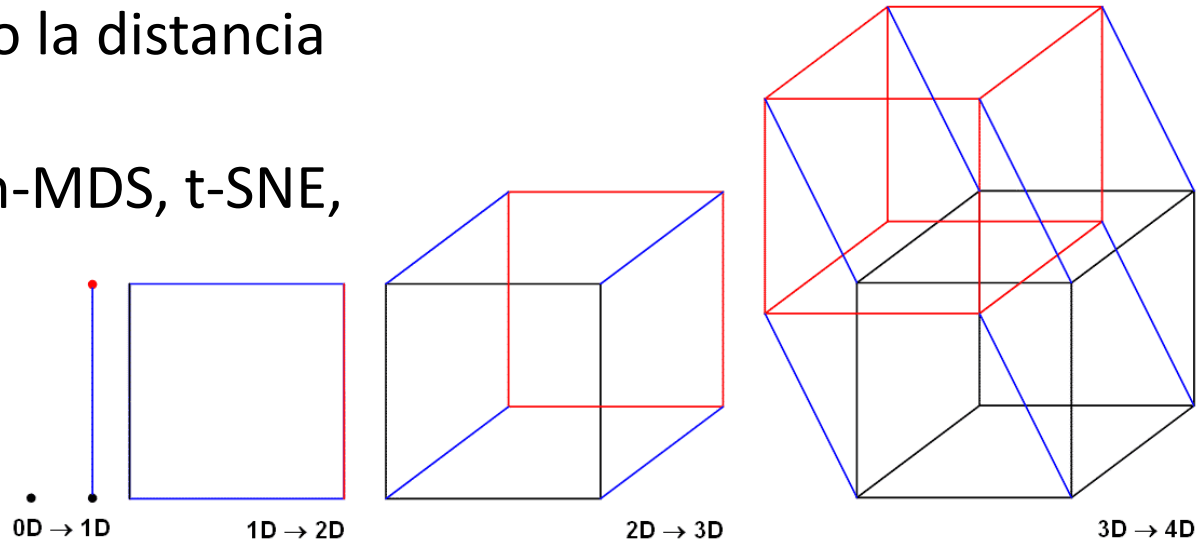
	V_1	V_2	V_3	...	V_{m-1}	V_m
1						
2						
3						
...						
n-1						
n						



	W_1	...	W_k
1			
2			
3			
...			
n-1			
n			

Técnicas de reducción de dimensiones

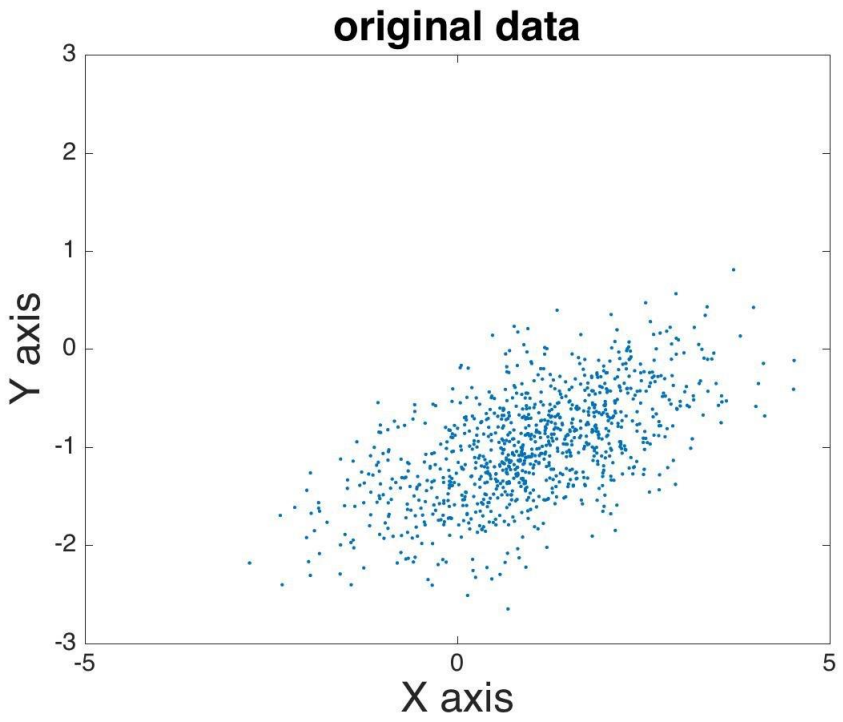
- Análisis de componentes principales (PCA): Transformación lineal, busca minimizar la varianza inexplicable
- Análisis factorial: Combinación lineal de un pequeño número de variables latentes
- Escalamiento multidimensional (MDS): Proyecta datos en subespacio de baja dimensión conservando la distancia entre puntos (puede ser no lineal)
- Otras técnicas supervisadas y no lineales: n-MDS, t-SNE, Kohonen maps

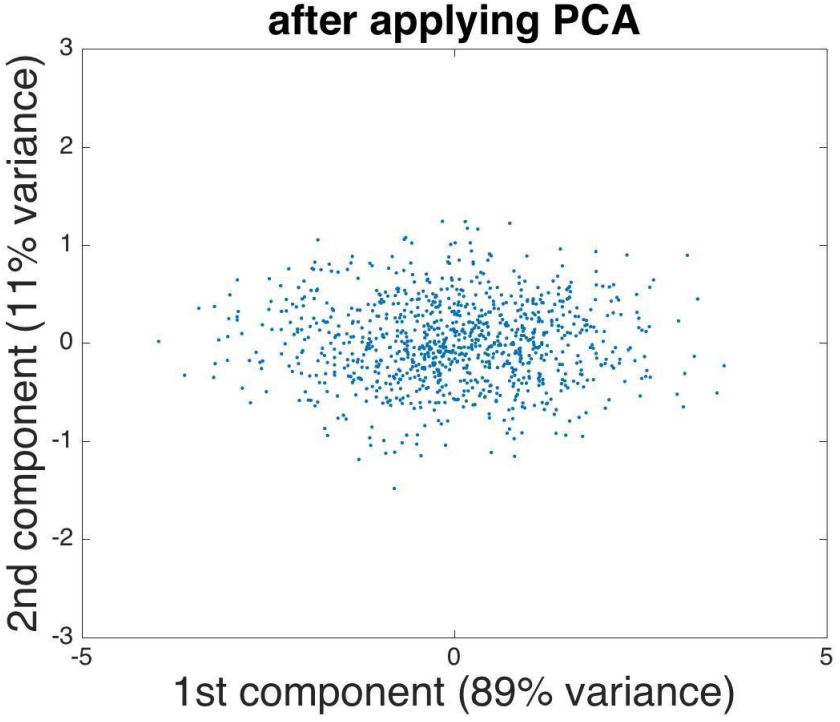


Análisis de componentes principales (PCA)

Definición

PCA es una transformación lineal, donde las nuevas variables son ortogonales (no están relacionadas y la covarianza entre cualquier par de variables es cero) y se ordenan en función del porcentaje de la variabilidad de datos explicada.

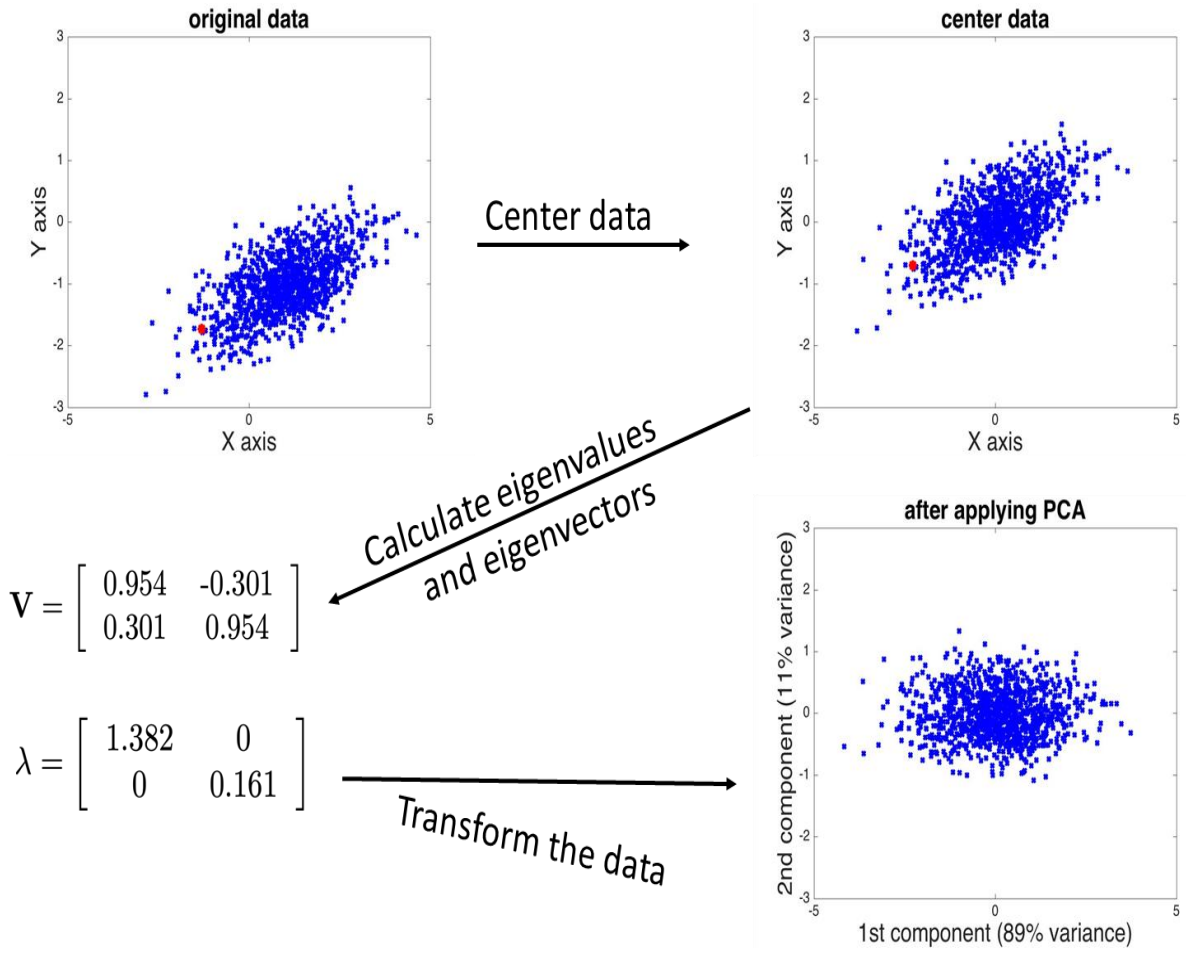




Algoritmo PCA

Explicación de alto nivel de abstracción:

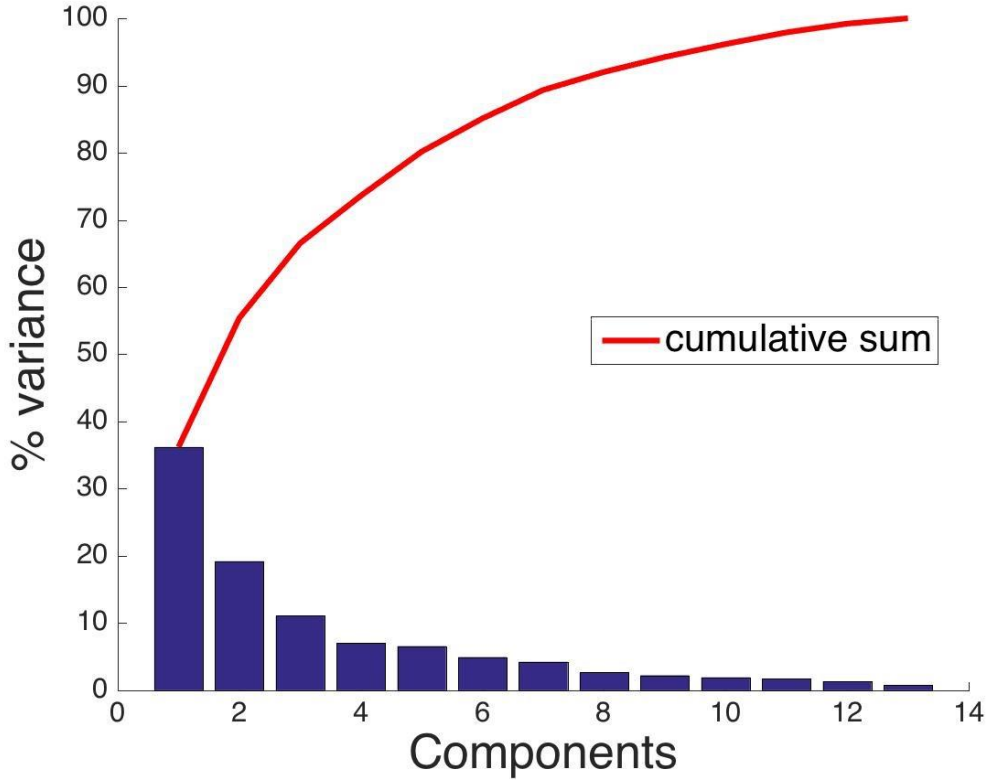
1. Genera combinación lineal de variables normalizadas que maximicen la varianza de la variable dependiente (primer componente principal).
2. Genera combinación lineal de variables normalizadas que maximicen la varianza de la variable dependiente, sujeto a que variable dependiente sea ortogonal a componente principal anterior
3. Repetir hasta que no quede varianza por ser explicada (siempre será inferior al numero de atributos originales)



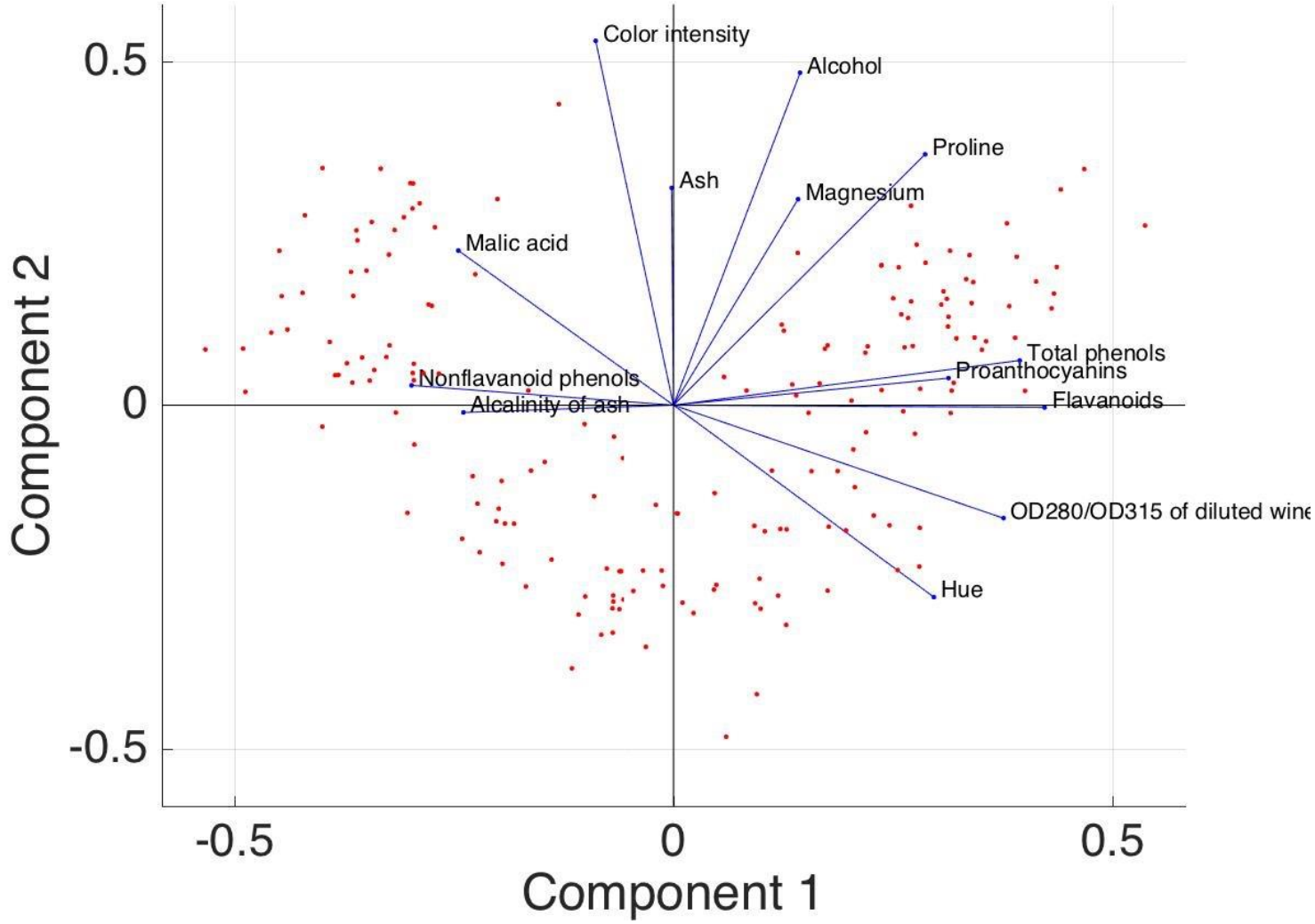
Número de componentes principales

Los valores propios estandarizados representan el porcentaje de varianza explicado por cada componente.

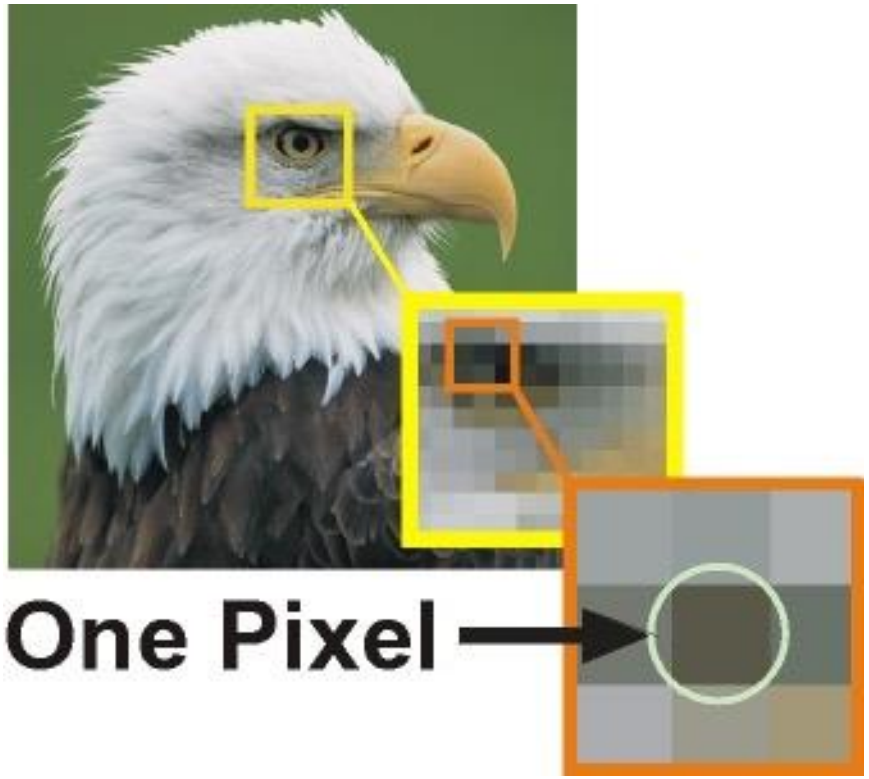
Visualizando la varianza explicada por cada componente podemos determinar el número de componentes que deberíamos utilizar.



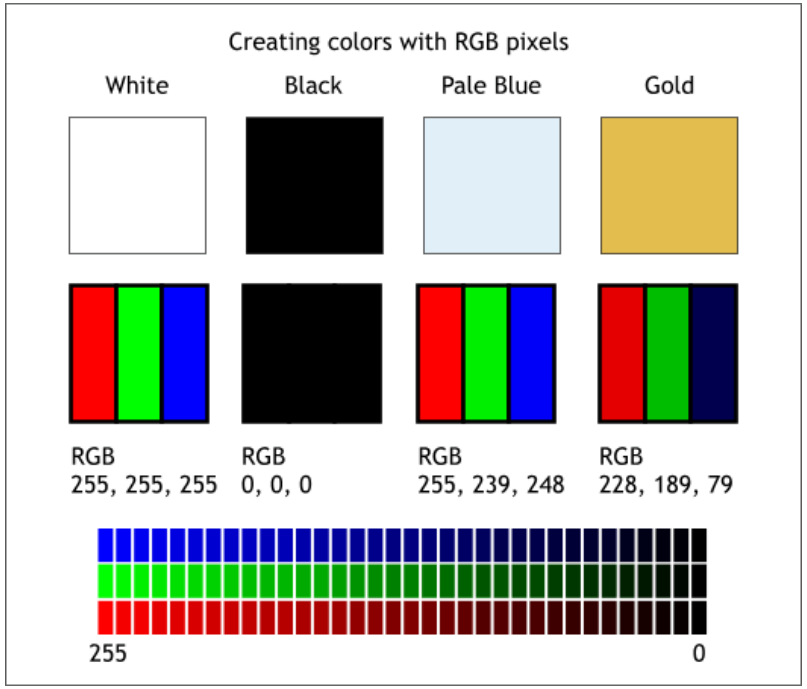
Diagramas de carga



PCA en imágenes

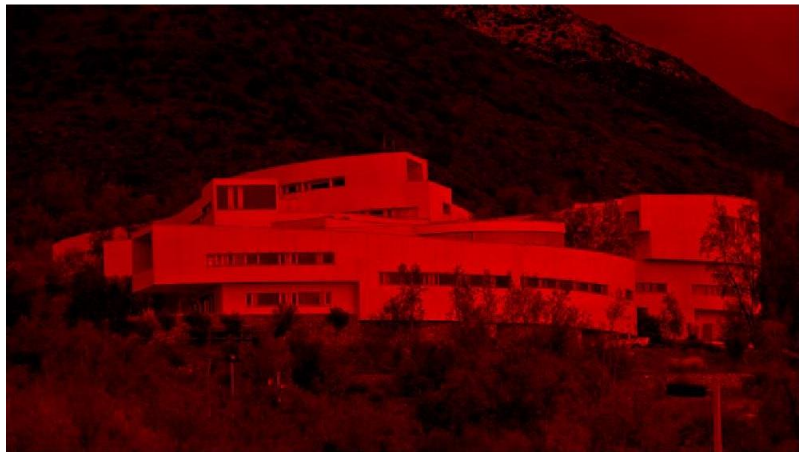


- La imagen digital contiene un número fijo de filas y columnas de píxeles.
- Los píxeles RGB son el elemento individual más pequeño de una imagen, con valores que representan el brillo de un color determinado en cualquier punto específico, entre 0 y 255.



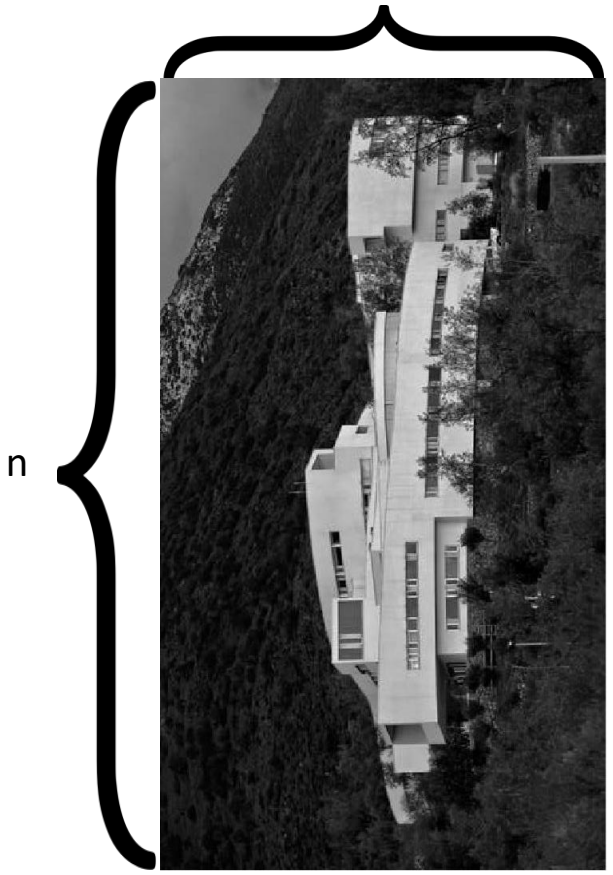
Ejemplo imágenes

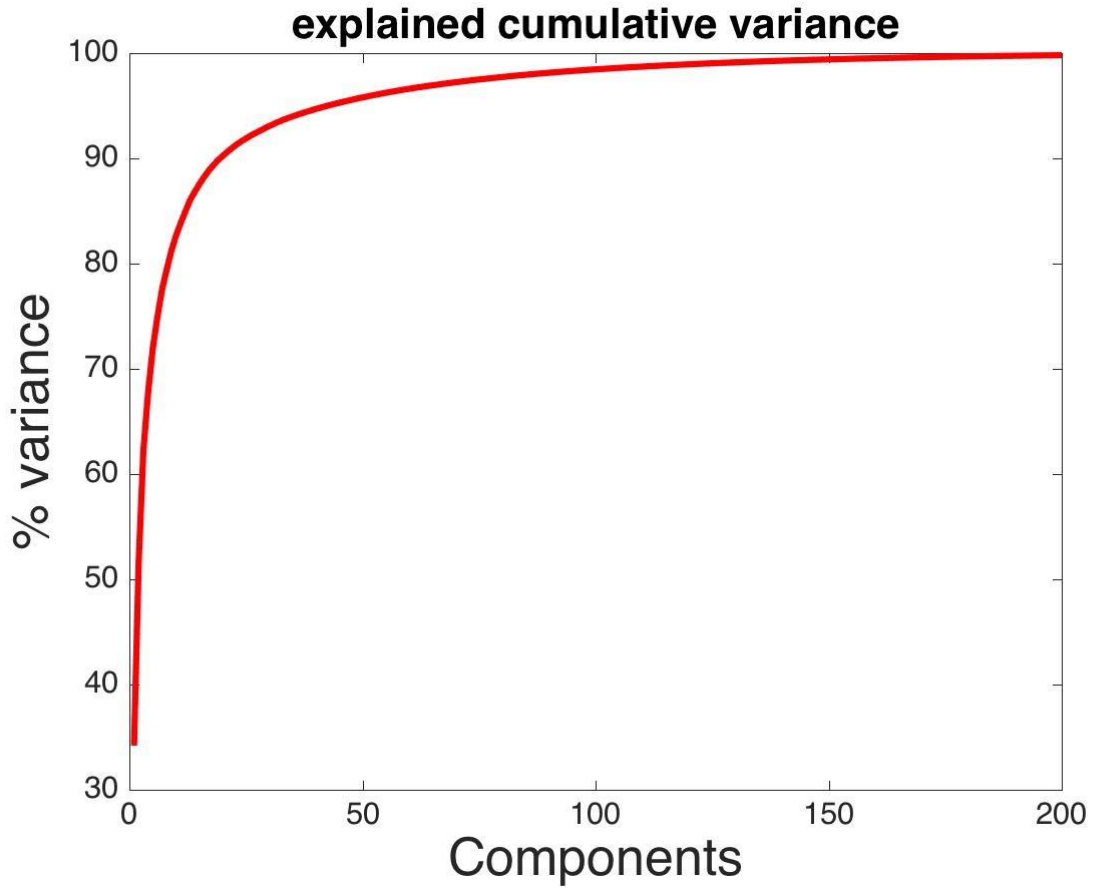
Por ejemplo, esta imagen UAI se puede separar en sus canales RGB



Ejemplo imágenes

En la compresión de datos, la dimensión de datos se reduce mediante PCA y se preserva un subconjunto de componentes, con la correspondiente pérdida de información.

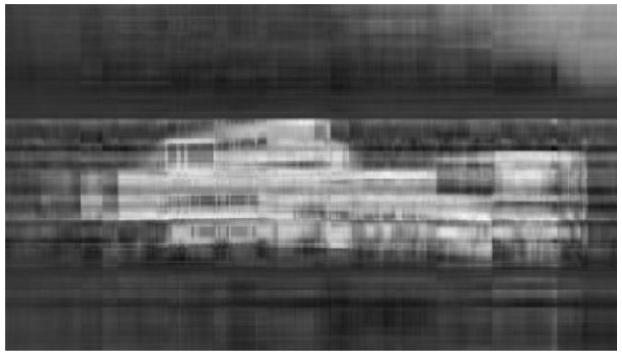




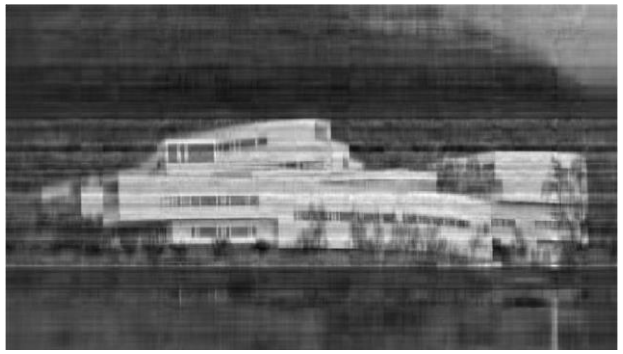
Ejemplo imágenes

En la compresión de datos, la dimensión de datos se reduce mediante PCA y se preserva un subconjunto de componentes, con la correspondiente pérdida de información.

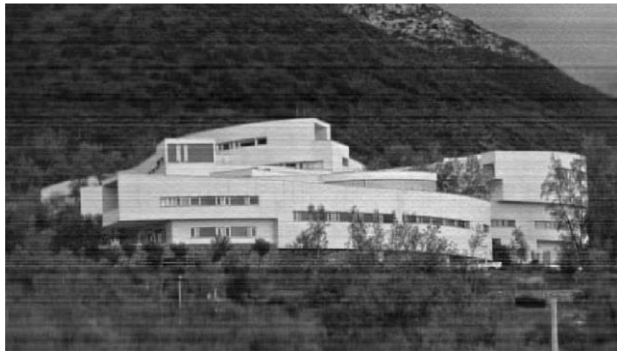
n



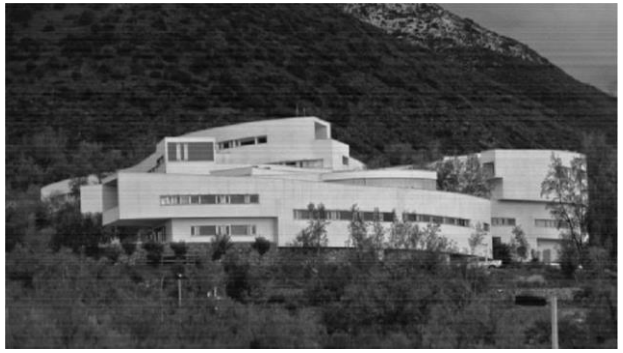
NC=10=>83%



NC=20=>90%



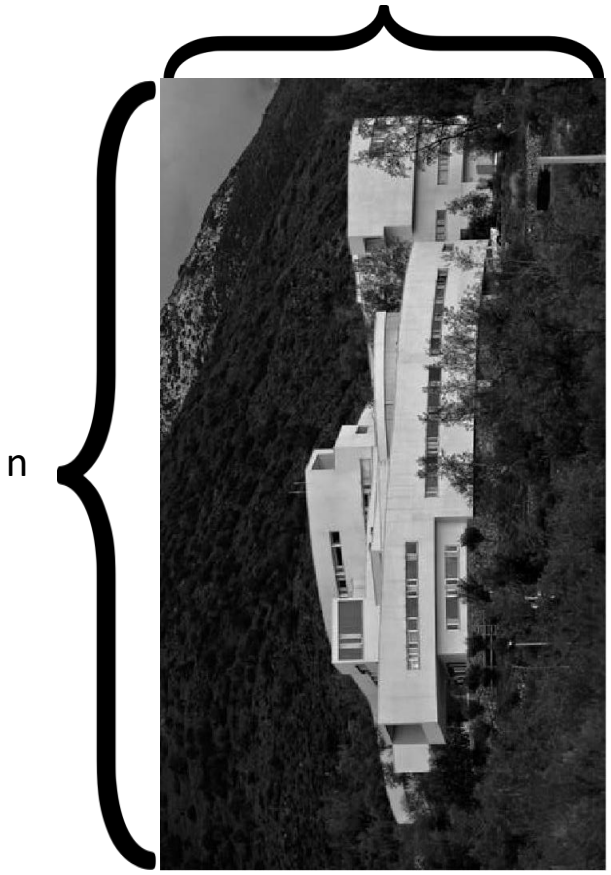
NC=80=>97.7%

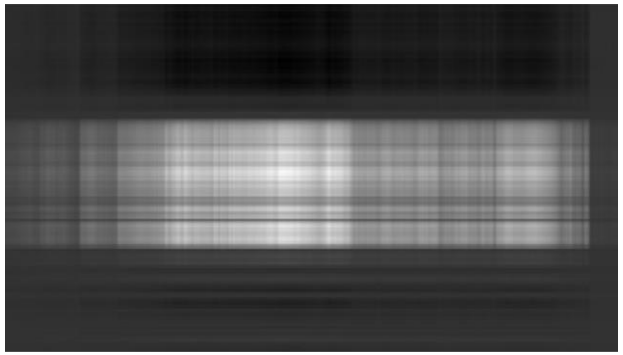


NC=160=>99.5%

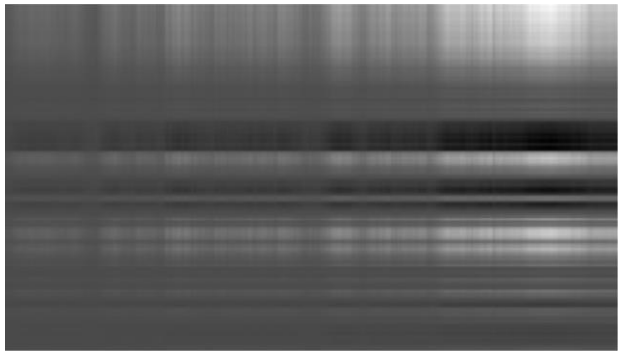
Ejemplo imágenes

También podemos ver los datos recuperados por cada componente

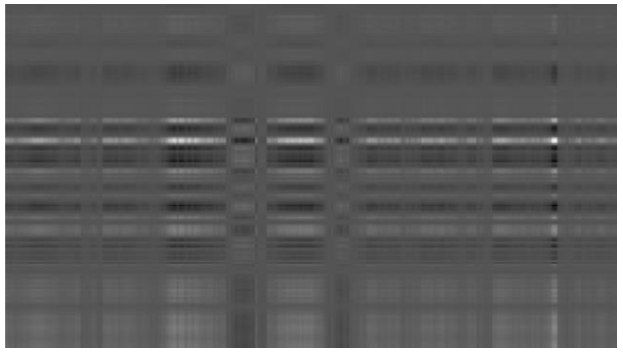




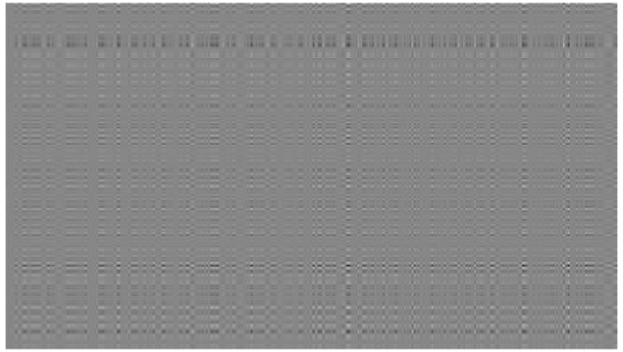
First component



Second component



Tenth component



100th component

Ingeniería de atributos

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2