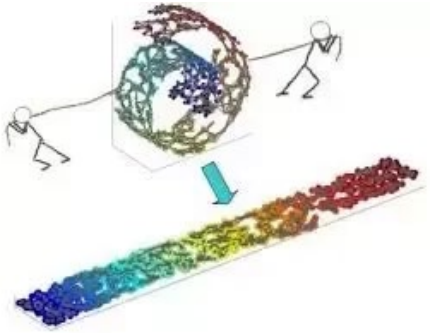
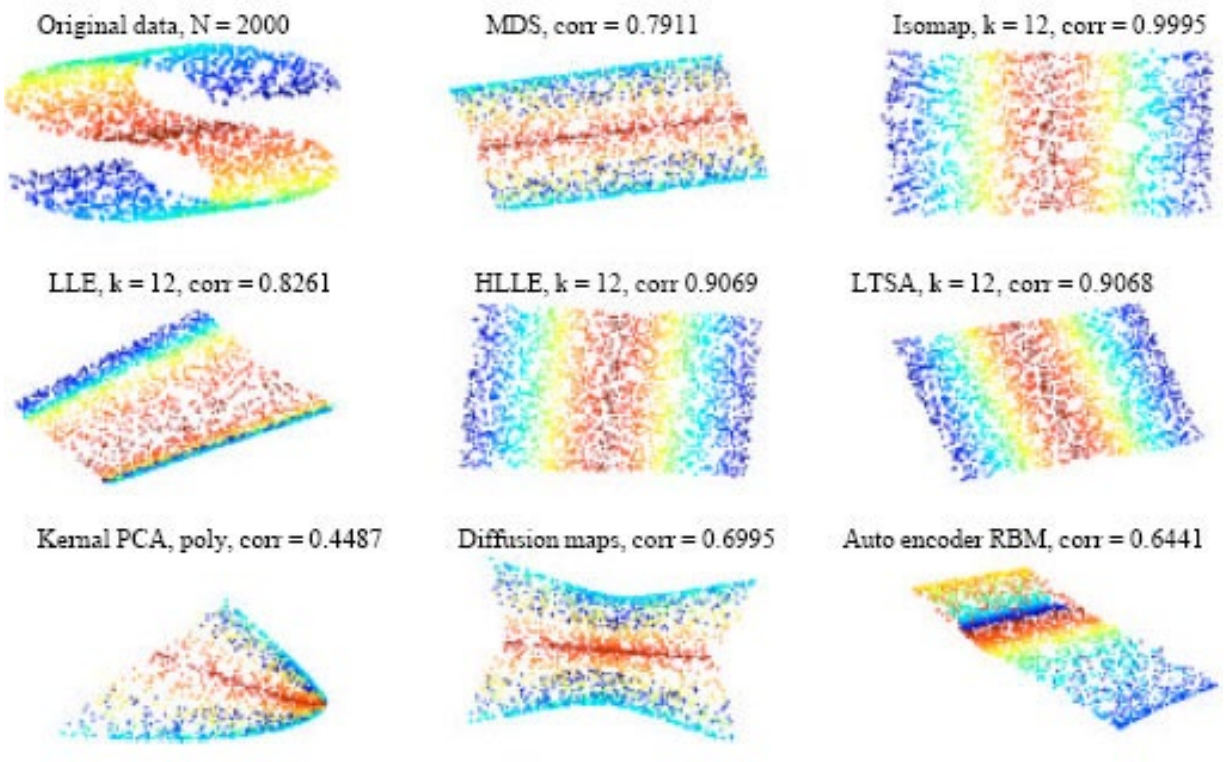




Manifold Learning

PhD(e). Jonnatan Arias Garcia – jonnatan.arias@utp.edu.co – jariasg@uniquindio.edu.co

PhD. David Cardenas peña - dcardenasp@utp.edu.co

PhD. Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

La maldición de la dimensionalidad

Una gran cantidad de datos en ML involucran miles/millones de funciones.

Estas características pueden hacer que el entrenamiento sea muy lento.

Además, hay mucho espacio en dimensiones altas, lo que hace que los conjuntos de datos de alta dimensión sean escasos, ya que es muy probable que la mayoría de las instancias de entrenamiento estén lejos unas de otras.

Esto aumenta el riesgo de sobreajuste, ya que las predicciones se basarán en extrapolaciones mucho mayores en comparación con las de datos de baja dimensión. Esto se llama *la maldición de la dimensionalidad* .

Tenemos dos enfoques principales para la reducción de dimensionalidad: proyección y manifold learning.

Y PCA?

Se puede utilizar PCAen reducción de dimensionalidad:

Reduce la cantidad de características de un conjunto de datos mientras se mantienen las relaciones esenciales entre los puntos.

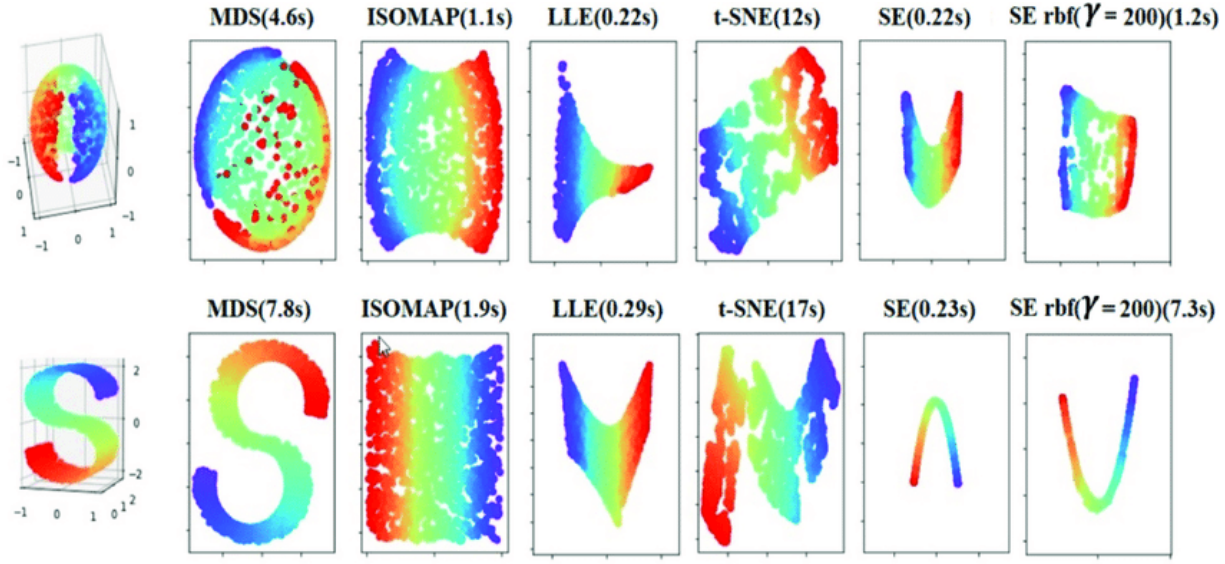
PCA es flexible, rápido y fácilmente interpretable, pero, no funciona tan bien cuando existen relaciones *no lineales* dentro de los datos

Esto lo podemos solucionar con unos estimadores no supervisados denominados manifold learning

Manifold

Es una generalización de conceptos geométricos familiares como líneas, planos y superficies a espacios de dimensiones superiores.

El objetivo del manifold learning es encontrar y representar estas estructuras de bajo dimensionales, permitiendo así la visualización, reducción de dimensionalidad y comprensión de datos complejos.



Manifold Learning

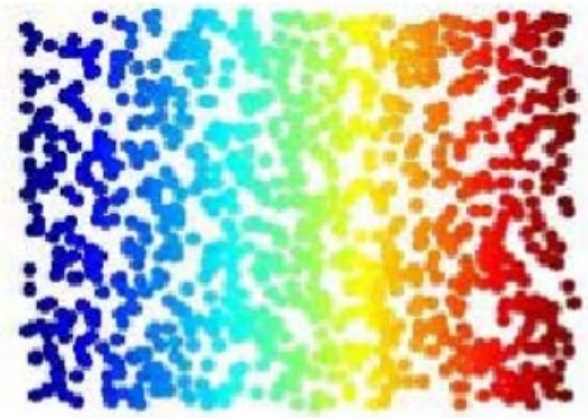
Un manifold bidimensional es cualquier forma 2D que puede encajar en espacio de dimensiones superiores girándola o doblándola.

Hipotesis del Manifold Learning

“Los datos de alta dimensión del mundo real se encuentra en variedades de baja dimensión, incrustados dentro de espacios de alta dimensión”

En términos simples, los datos de dimensiones superiores se pueden encuentra enmascarados en una variedad de dimensiones inferiores mas cercanas. El proceso de modelar el manifold en estas instancias, se llama **manifold learning**

(a) Swiss roll



(b)Original manifold

Métodos Base en Manifold Learning

- Escalamiento Multidimensional (MDS)
- Mapeo Isométrico (IsoMap)
- Locally Linear Embedding (LLE)

Distancia, disimilitud y similitud

Son definidas por un par de objetos en un espacio. En matemáticas una función de distancia es también una metrica que satisface:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0$ si y solo si $x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

Dado un conjunto de disimilitudes, uno puede preguntarse si estos valores de distancias , se pueden interpretar como distancias euclidianas

Distancia Euclidiana o No-Euclidiana

Dada una matriz de disimilitud (distancia) $D = (d_{ij})$, MDS busca encontrar$\boldsymbol{x_1}, \ldots, \boldsymbol{x_n} \in \mathbb{R}^{\boldsymbol{p}}$ tal que

$$d_{ij} \approx \left|\left|\boldsymbol{x_i} - \boldsymbol{x_j}\right|\right|_{\mathbf{2}}$$ tan cercano como sea posible.

A menudo, para algunas p grandes, existe una configuración $x_1, \ldots, x_n$ con una distancia exacta que coincide con $d_{ij} \equiv \left|\left|x_i - x_j\right|\right|_2$

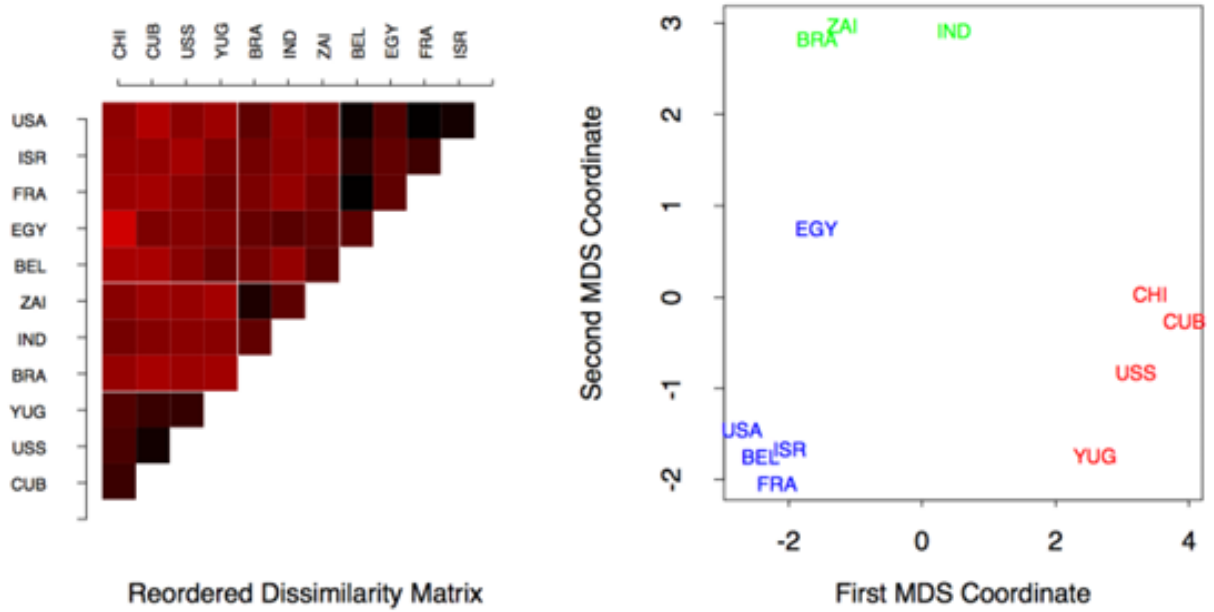
En tal caso, la distancia d involucrada se llama distancia euclidiana. Sin embargo, hay casos en los que la disimilitud es la distancia, pero no existe ninguna configuración en ningún p con coincidencia perfecta.

$$d_{ij} \neq \left|\left|\boldsymbol{x_i} - \boldsymbol{x_j}\right|\right|_{\mathbf{2}}$$ para algún i, j

Esta distancia se llama distancia no euclidiana.

Escalamiento Multidimensional MDS

- Dado disimilitudes entre puntos, reconstruir un mapa que preserve las distancias
  - Cualquier disimilitud (No es necesario que sea una métrica)
  - El mapa reconstruido tiene coordenadas $\boldsymbol{x_i} = (x_{i1}, x_{i2})$ y la distancias $\left(\left|\left|\boldsymbol{x_i} - \boldsymbol{x_j}\right|\right|_2\right.$



Escalamiento Multidimensional MDS

- MDS es una familia de algoritmos, cada uno diseñado para llegar a una configuración optima low-dimensional

  - Classical MDS
  - Metric MDS
  - Non-metric MDS

MDS Clásico

Se asume que las distancias originales entre los objetos son métricas que se pueden representar de manera exacta en baja dimensión.

Por ende busca preservar las relaciones de distancias exactas.

Esto se hace con el uso de SVD (descomposición de valores singulares para calcular las coordenadas de los objetos en baja dimensión.

Es computacionalmente eficiente, y adecuado cuando las distancias originales son confiables y se deben preservar de manera precisa.

MDS Clásico

Dada una matriz de proximidad D de tamaño *nxn*, donde n es el numero de objetos, la representación del espacio de baja dimensión *X* se calcula:

$$X = U_k \Lambda_k^{1/2}$$

Donde:

* U_k es una matriz *nxk* de vectores singulares izquierdos correspondientes a los valores k singulares mas grandes de *D*

* Λ_k es una matriz diagonal *kxk* de los *k* valores singulares mas grandes de D

MDS Métrico

A comparación del clásico, se relajan las restricciones de preservación exacta y se permiten cierto margen. Usa técnicas de optimización para minimizar una función de costo que mide las distancias originales y las del espacio de baja dimensión.

Suese ser mas flexible que el clásico y manejar situaciones donde las distancias originales no son confiables o ruidosas.

MDS Métrico

El espacio de baja dimensión *X* se calcula:

$$\min_{x} \sum_{i<j} \left(d_{ij} - \left|\left| x_i - x_j \right|\right| \right)^2$$

Donde $d_{i,j}$es la distancia original entre los objetos i, j y $\left|\left| x_i - x_j \right|\right|$es la distancia euclidiana entre las representaciones de los objetos i, jen el espacio *X*

MDS No-métrico

Permite que las relaciones de orden entre las distancias sean aun mas relajadas. Suele utilizar técnicas de optimización.

Sirve cuando la matriz de proximidad solo proporciona información cualitativa sobre las relaciones de los objetos, como rankings, juicios de preferencia..

MDS No-métrico

El espacio de baja dimensión *X* se calcula:

$$\min_{x} \sum_{i<j} \left(D_{ij} - \left|\left| x_i - x_j \right|\right| \right)^2$$

Donde $D_{i,j}$ son las medidas de similitud/desimilitud originales entre los objetos i, j y $\left|\left| x_i - x_j \right|\right|$es la distancia euclidiana entre las representaciones de los objetos i, jen el espacio *X*

MDS consideraciones

Nuestra discusión ha considerado embeddings *lineales* , que esencialmente consisten en rotaciones, traslaciones y escalamientos en espacios de dimensiones superiores.

Cuando MDS falla es cuando la incorporación no es lineal, es decir, cuando va más allá de este simple conjunto de operaciones.

Locally Linear Embedding LLE

Si los puntos en un conjunto de alta dimensión se encuentran en una manifold de baja dimensión incrustada en el espacio de alta dimensión, entonces las relaciones de vecindad local entre los puntos se mantendrán incluso después de la reducción de dimensionalidad.

- Puede manejar conjuntos de datos con muchas dimensiones.
- Útil para descubrir estructuras de manifold no lineales y puede capturar formas complejas en los datos.
- Es computacionalmente eficiente, especialmente en comparación con técnicas más intensivas en cálculos, como el t-SNE.

Locally Linear Embedding LLE

El proceso de LLE se puede dividir en los siguientes pasos:

1. **Vecindad local:** Para cada punto en alta dimensión, se identifican sus vecinos más cercanos. La vecindad puede ser definida por una cantidad fija de vecinos o por un radio de vecindad definido.

Teniendo un conjunto de datos de alta dimensión representado por una matriz X de tamaño $D \times N$, donde D es la dimensión de los datos originales y N es el número de muestras.

El punto 1 implica encontrar los k vecinos más cercanos para cada punto en el espacio de alta dimensión.

Locally Linear Embedding LLE

2. Reconstrucción local: Para cada punto, se encuentra una representación lineal de él mismo basada en sus vecinos. Esto implica encontrar los pesos (coeficientes) que mejor reconstruyen el punto como una combinación lineal de sus vecinos.

$$\min_{W_i} \left\| x_i - \sum_{j \in N(i)} w_{ij} s_j \right\|^2$$

Donde x_i es el punto de interés

$N(i)$ es el conjunto de vecinos mas cercanos de i

w_{ij} son los pesos que se asignan a cada vecino para reconstruir el punto i

Locally Linear Embedding LLE

3. Embedding global: Se encuentra una representación de baja dimensión de los datos que conserva las relaciones de vecindad local. Esto implica encontrar las coordenadas de los puntos en el espacio de baja dimensión de tal manera que las relaciones lineales entre los puntos en el espacio original se conserven lo mejor posible.

$$\min_{Y} \sum_{i=1}^{N} ||y_i - \sum_{j \in N(i)} w_{ij} y_j||^2$$

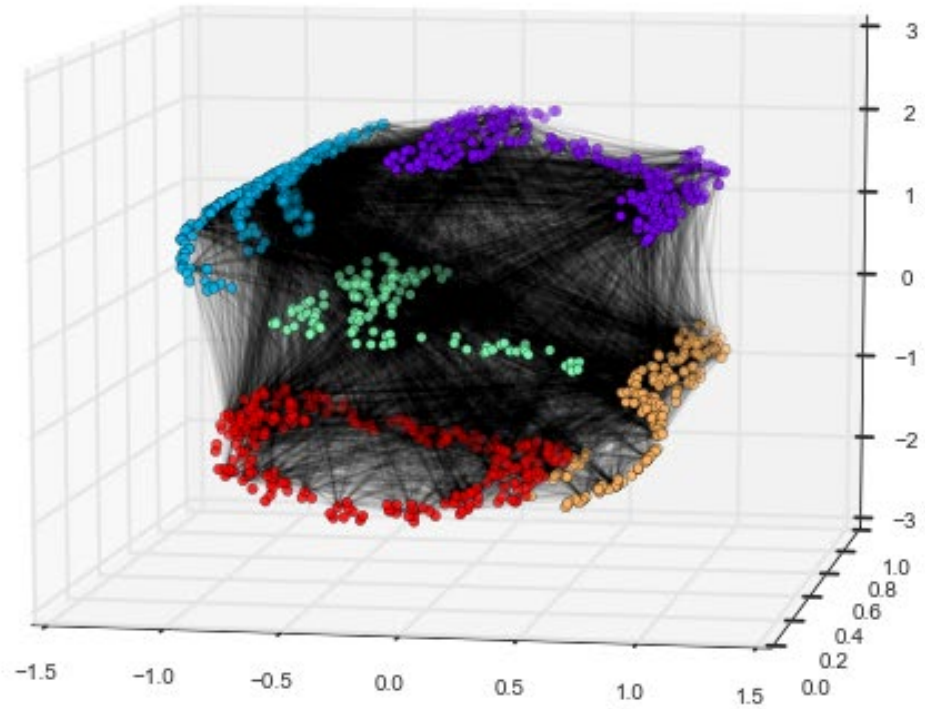
Y matriz de tamaño *dxN* que representa coords. De los puntos de baja dim. (d es la dimensión deseada)

y_i es la coord. De x_i en baja dim.

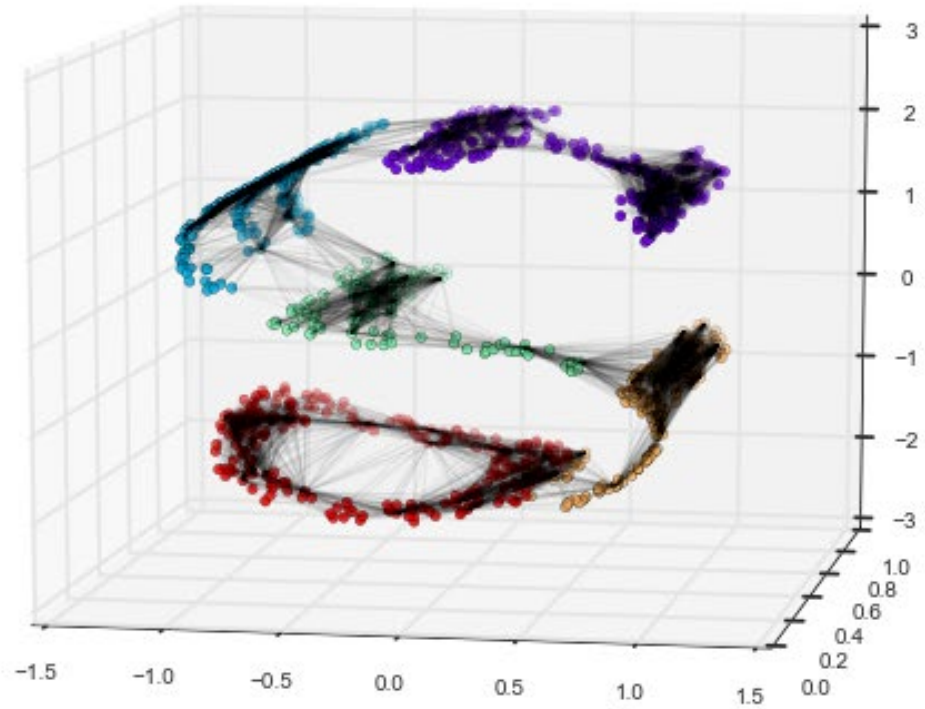
w_{ij} son los pesos obtenidos en la reconstrucción local

W se calcula diferentes para LLE estándar, LLE regularizado y demás variantes

MDS Linkages

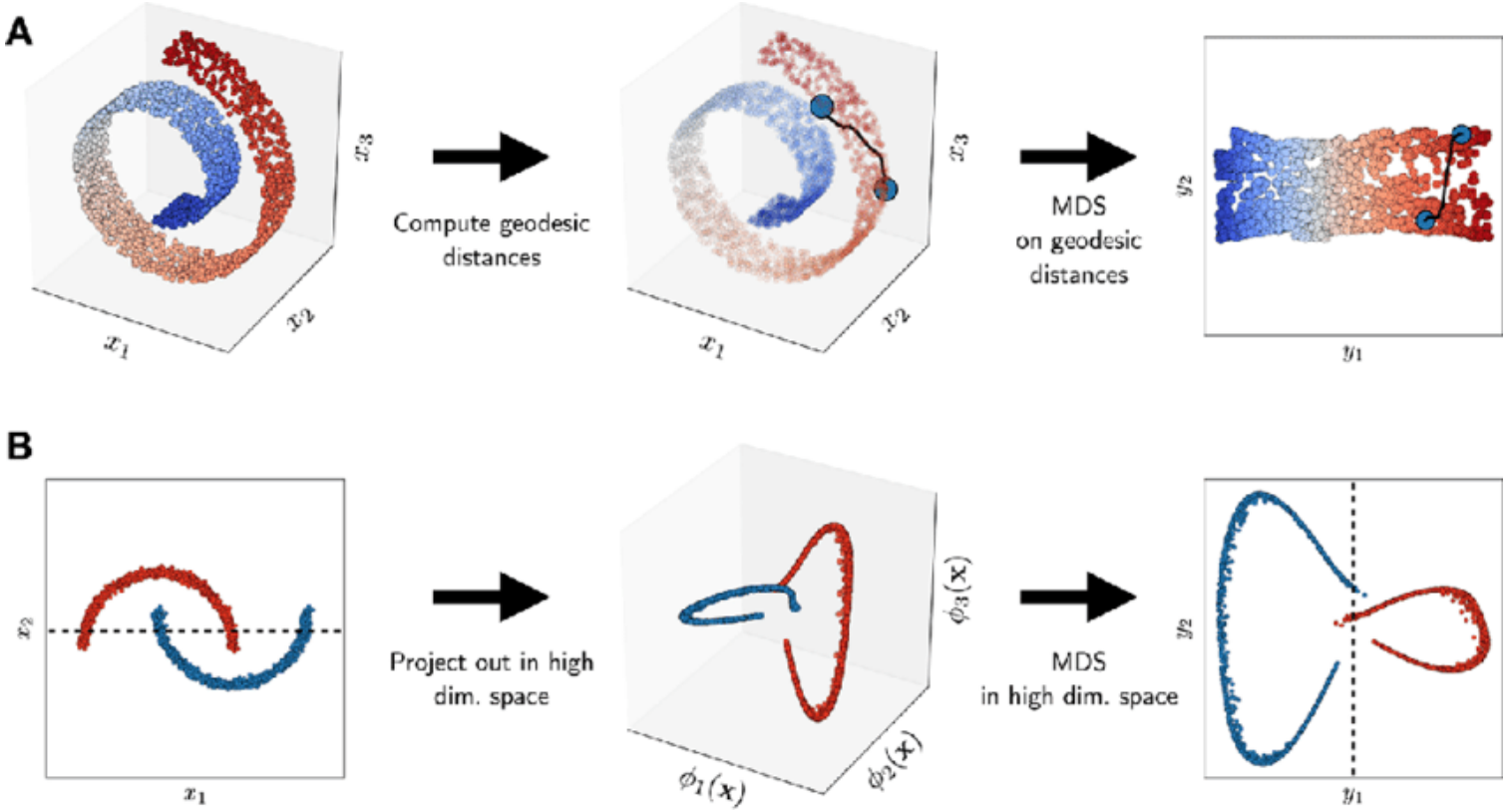


LLE Linkages (100 NN)



Mapeo Isométrico (IsoMap)

Isomap busca capturar esta estructura no lineal al calcular las distancias geodésicas entre los puntos en el espacio de alta dimensión, y luego proyectar estos puntos en un espacio de dimensiones bajas mientras se conservan estas distancias geodésicas lo más fielmente posible.



Mapeo Isométrico (IsoMap)

El algoritmo de Isomap consta de los siguientes pasos:

1. **Construcción del grafo de vecindad:** Se calculan las distancias euclidianas entre todos los pares de puntos en el conjunto de datos de alta dimensión. Luego, se utiliza un método de vecinos más cercanos para construir un grafo de vecindad, donde los puntos se conectan si son vecinos cercanos entre sí.
2. **Cálculo de las distancias geodésicas:** Se calculan las distancias geodésicas entre todos los pares de puntos en el grafo de vecindad utilizando un algoritmo como Dijkstra o Floyd-Warshall. Estas distancias geodésicas representan las distancias a lo largo de la variedad subyacente y capturan la estructura de los datos de manera más efectiva que las distancias euclidianas en espacios de alta dimensión.

Mapeo Isométrico (IsoMap)

3. Embedding en un espacio de baja dimensión: Finalmente, se utiliza un algoritmo de reducción de dimensionalidad, como MDS (Multidimensional Scaling), para proyectar los puntos en un espacio de dimensiones bajas mientras se conservan las distancias geodésicas tanto como sea posible.

Esto proporciona una representación de los datos en un espacio de baja dimensión que captura la estructura subyacente de los datos.

Mapeo Isométrico (IsoMap)

función de costo:

$$\min_Y \sum_{i=1}^{N} \sum_{j \in N(i)} (d_{ij} - \|y_i - y_j\|^2)^2$$

Donde:

- Y es una matriz de tamaño $d \times N$ que representa las coordenadas de los puntos en el espacio de baja dimensión, donde d es la nueva dimensión deseada.
- d_{ij} son las distancias geodésicas calculadas entre los puntos i y j.
- y_i y y_j son las coordenadas de los puntos i y j en el espacio de baja dimensión, respectivamente.