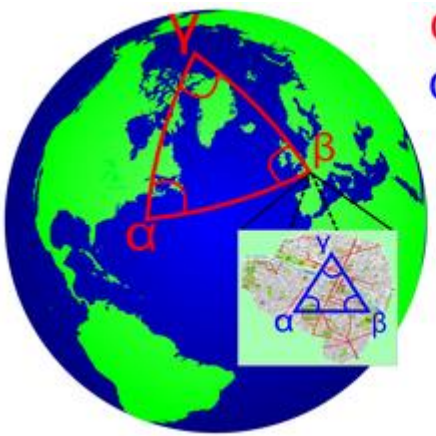
ISOMAP

Ing. Juan M. Rodríguez

ISOMAP

¿Qué pasa cuando tenemos una **variedad,** usando la definición matemática, sobre la cual están distribuidos los datos?

Una variedad es una porción del espacio, de dimensión **n** que se parece a $\mathbb{R}^n$, en un espació de dimensión mayor. (hablando mal y pronto)



$\alpha + \beta + \gamma > 180°$

$\alpha + \beta + \gamma = 180°$

La porción del espacio $\mathbb{R}^3$ delimitada por alfa, beta y gama es aproximadamente un triángulo en $\mathbb{R}^2$

ISOMAP

Hipótesis de variedades

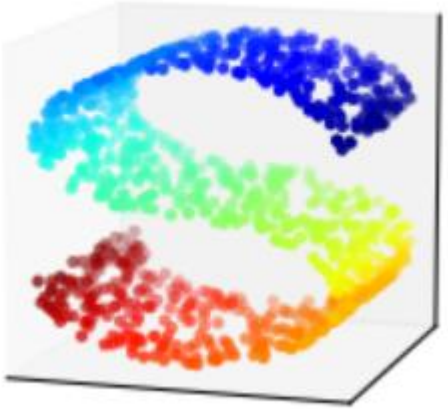
Esta hipótesis sostiene que la mayoría de los conjuntos de datos de alta dimensión del mundo real quedan cerca de una variedad con muchas menos dimensiones.

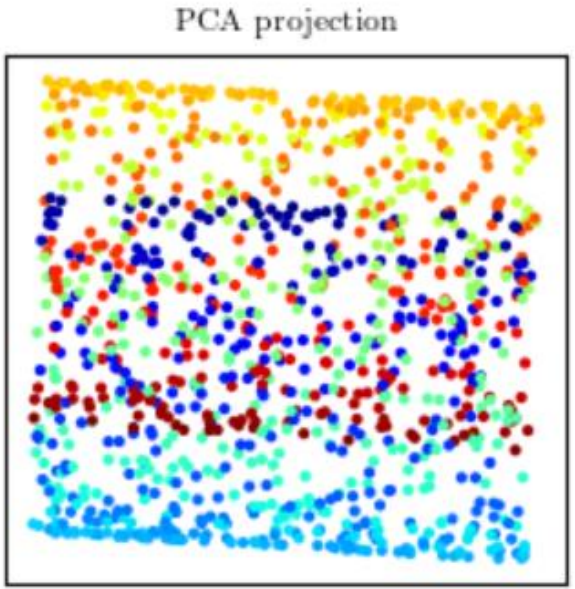
Este supuesto se observa a menudo de forma empírica.

Ejemplo: **MNIST** tiene datos de entrada en un espacio de 784 dimensiones (28x28 píxeles). Si generamos imágenes aleatorias, permitiendo total libertad en cada dimensión (cada dimensión puede ir de 0 a 255), sólo una fracción increíblemente pequeña podrían ser catalogados de números manuscritos (cercana a cero). Al observarlas vemos que tienen similitudes: líneas conectadas, centradas, contraste, etc. A la fuerza deben ocupar un subespacio mucho menor que $\mathbb{R}^{784}$

ISOMAP

Supongamos que en $\mathbb{R}^3$ los datos se distribuyen como en una S. El color me muestra la variable de salida



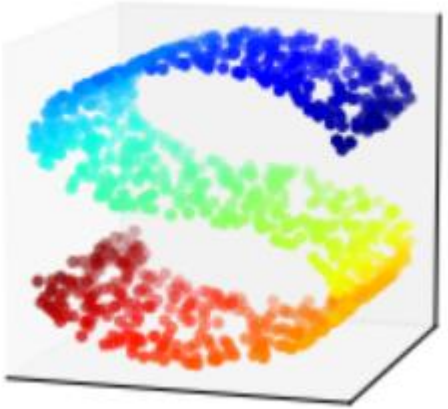


Si aplico PCA para reducir la dimensionalidad y ver los datos obtengo esto.
Si bien puedo verificar que la distancia euclidiana entre los puntos azules y los rojos quizá no es tanta, si sigo el camino siguiendo la topología de los datos (la “S”), la distancia entre estos es máxima (mayor que entre cualquier otro par de puntos de colores)

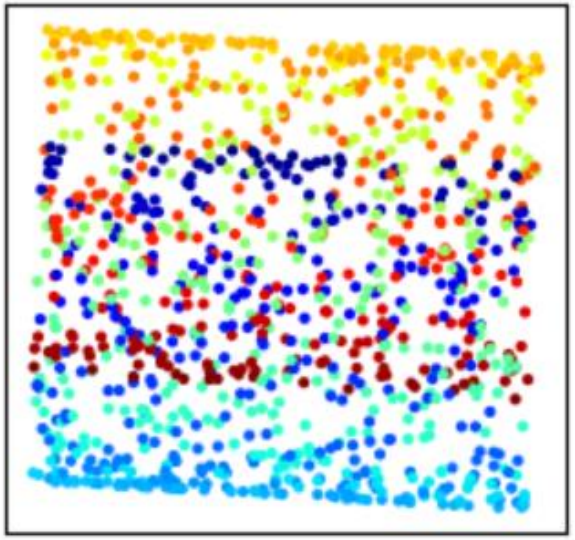
ISOMAP

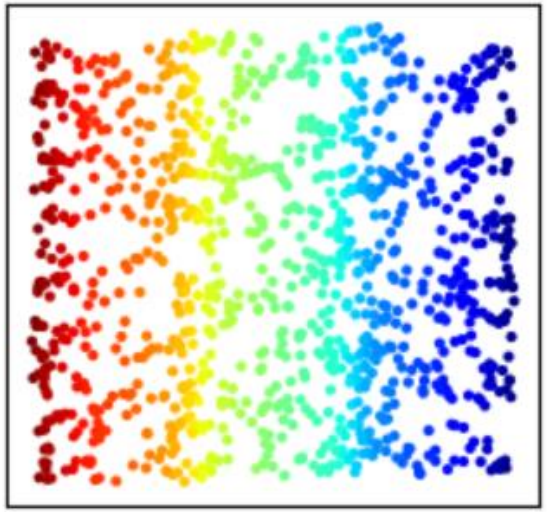
Supongamos que en $\mathbb{R}^3$ los datos se distribuyen como en una S. El color me muestra la variable de salida

¡Este es el tipo de proyección que quisiera obtener!



PCA projection



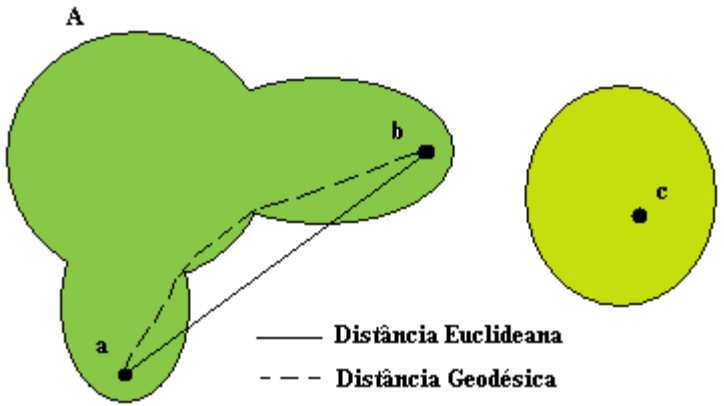


ISOMAP

¿Cómo puedo obtener una proyección así?

Utilizando la distancia **Geodésica** en lugar de la distancia **Euclidiana**!!

En geometría, **la línea geodésica** se define como la línea de mínima longitud que une dos puntos en una superficie dada, y está c**ontenida en esta superficie**.



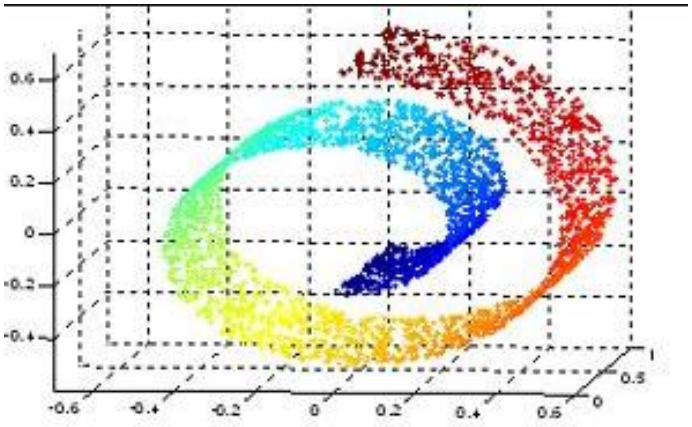
Pero hay un problema. No sé a *priori* si los datos siguen una forma particular o no. Y muchos menos sé qué forma podría ser esa.

ISOMAP

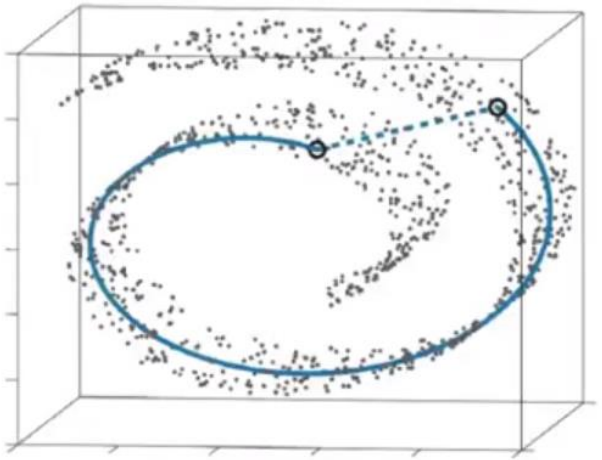
Existe un algoritmo que me permite aproxima la forma de los datos y calcular la distancia Geodésica:

ISOMAP

Supongamos que tengo esta distribución de datos



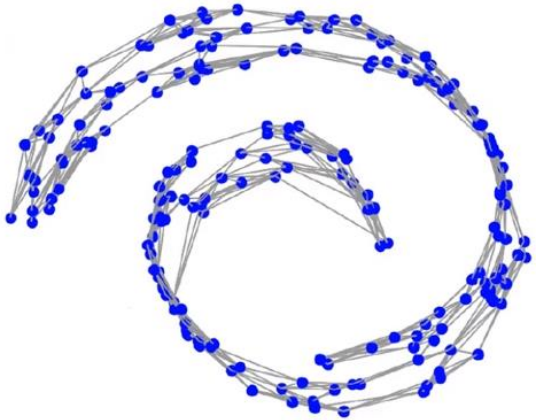
Solo a modo de ejemplo, para dos observaciones dadas, vemos la distancia Euclidiana (línea punteada) y la Geodésica



ISOMAP

Para intentar aproximar la distancia Geodésica:

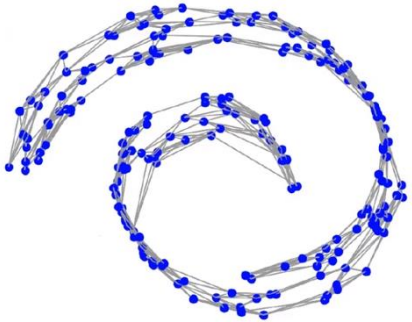
1. Construiremos un grafo pesado
    a. Para cada punto en el espacio de entrada, tomamos los **K** vecinos más cercanos (usamos la distancia Euclidiana para esto) **Nota**: k es un parámetro del método
    b. Trazamos aristas que conectan cada punto con sus vecinos más cercanos y ponderamos esas aristas según la distancia Euclidiana calculada

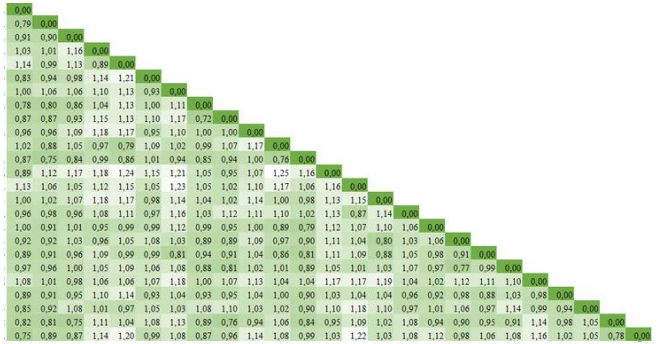


ISOMAP

Para intentar aproximar la distancia Geodésica:

1. Construiremos un grafo pesado
    a. Para cada punto en el espacio de entrada, tomamos los **K** vecinos más cercanos (usamos la distancia Euclidiana para esto) **Nota**: k es un parámetro del método
    b. Trazamos aristas que conectan cada punto con sus vecinos más cercanos y ponderamos esas aristas según la distancia Euclidiana calculada
2. Construimos una matriz de distancias entre todos los puntos, para dicha matriz usamos un algoritmo para encontrar la distancia más corta entre dos nodos en un grafo:
    a. Algoritmo de **Dijkstra** o de **Floyd-Warshall**





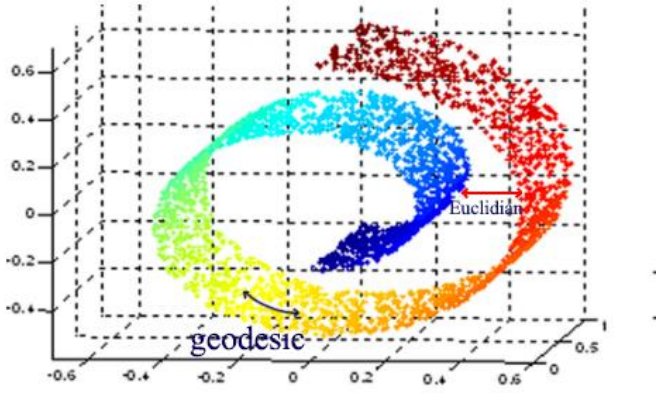
ISOMAP

Para intentar aproximar la distancia Geodésica:

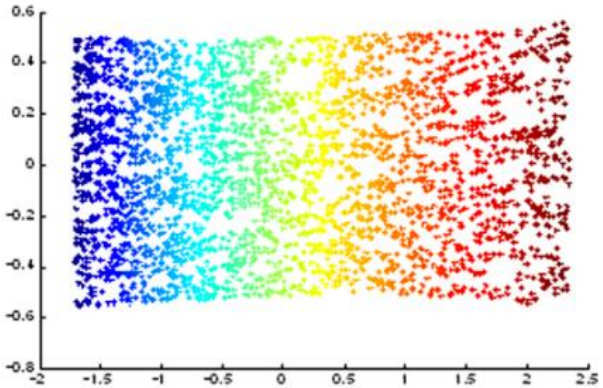
1. Construiremos un grafo pesado
    a. Para cada punto en el espacio de entrada, tomamos los **K** vecinos más cercanos (usamos la distancia Euclidiana para esto) **Nota**: k es un parámetro del método
    b. Trazamos aristas que conectan cada punto con sus vecinos más cercanos y ponderamos esas aristas según la distancia Euclidiana calculada
2. Construimos una matriz de distancias entre todos los puntos, para dicha matriz usamos un algoritmo para encontrar la distancia más corta entre dos nodos en un grafo:
    a. Algoritmo de **Dijkstra** o de **Floyd-Warshall**
3. Usamos MDS según se vio, pero con la matriz de distancia antes calculada.

ISOMAP

Resultado final:



(a)



ISOMAP

Debilidades:

- Es un algoritmo lento, a partir de 1000 observaciones comienza a notarse la lentitud.
- puede crear una proyección errónea si **k** es demasiado grande con respecto a la estructura de la variedad o si hay ruido en los datos, de tal forma que los puntos aparecen ligeramente movidos, fuera de la variedad. Incluso un solo dato mal medido puede alterar muchas entradas en la matriz de distancia geodésica.
- Si k es demasiado pequeño, el gráfico de vecindad puede volverse demasiado escaso para aproximar las trayectorias geodésicas con precisión.

Landmark ISOMAP

Landmark ISOMAP

Tengo N datos especiales, siendo que N es menor que el número total de datos. Y solo tomo las distancias que involucran a estos N datos especiales.

Los puedo elegir de forma uniformemente distribuida.

Luego se aplica Landmark-MDS (LMDS) en la matriz de distancias calculadas para encontrar una proyección euclidiana de todos los puntos en el conjunto original de datos.