

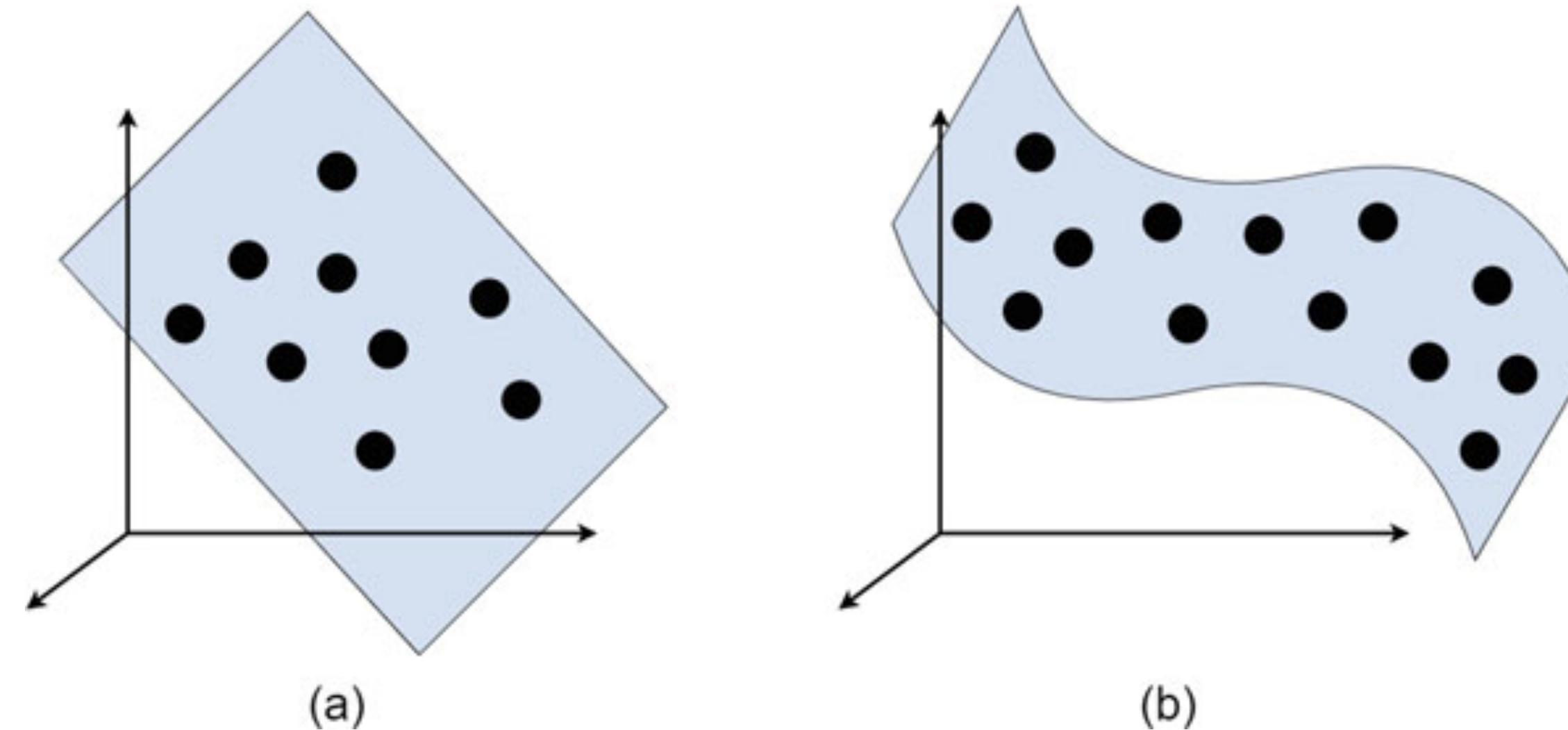
DATOS DE ALTA DIMENSIÓN

- Consideremos la medición de una cantidad. Por ejemplo:
 - datos personales de salud, como la tensión arterial, el azúcar en sangre y la grasa en sangre,
 - imágenes de una escena concreta pero tomadas desde distintas perspectivas,
 - imágenes de varias categorías de animales, como gato, perro, rana, etc,
 - imágenes médicas, como parches de imágenes patológicas digitales
- La cantidad puede ser multidimensional, y puede considerarse un punto de datos multidimensional en un espacio euclíadiano de dimensión d . Cada uno de estos d valores se denomina característica (feature)
- Habrá un conjunto de puntos de datos d -dimensionales, llamado conjunto de datos (dataset).
- Por ejemplo, la cantidad puede ser una imagen, cuyas características son sus píxeles. El conjunto de datos puede ser un conjunto de imágenes de una escena concreta pero con perspectivas y ángulos diferentes.

HIPÓTESIS DE SUBESPACIO (MANIFOLD HYPOTHESIS)

- Cada característica de un punto de datos no contiene la misma cantidad de información. Por ejemplo, algunos píxeles de una imagen son regiones de fondo con información limitada, mientras que otros píxeles contienen objetos importantes que describen la escena en la imagen.
- Esto significa que los puntos de datos pueden comprimirse considerablemente para conservar las características más informativas y eliminar las que tienen información limitada.
- En otras palabras, los puntos de datos d -dimensionales de un conjunto de datos no suelen cubrir todo el espacio euclíadiano d -dimensional, sino que se sitúan en una estructura específica de dimensiones inferiores en el espacio.
- Llamaremos *manifold*, subespacio o variedad a esta estructura de menor dimensión.

HIPÓTESIS DE SUBESPACIO (MANIFOLD HYPOTHESIS)



- En este ejemplo, los puntos del conjunto de datos tienen una estructura en un espacio bidimensional tanto en (a) como en (b).
- El espacio euclidiano tridimensional se denomina espacio de entrada, y el espacio bidimensional, que tiene una dimensionalidad menor que el espacio de entrada, se denomina subespacio, submanifold o espacio de incrustación (embedding).
- El subespacio puede ser lineal (a) o no lineal (b).

HIPÓTESIS DE SUBESPACIO (MANIFOLD HYPOTHESIS)



- El que los puntos del dataset estén en un espacio es solo una hipótesis, pero frecuentemente esa hipótesis es cierta porque los datos representan una señal natural, como por ejemplo una imagen
- De manera más formal, supondremos que los puntos de un dataset d -dimensional pertenecen a un subespacio de dimensión local menor que d con alta probabilidad.

INGENIERÍA DE CARACTERÍSTICAS

- Si se cumple la hipótesis anterior, un conjunto de datos puede comprimirse conservando la mayor parte de la información importante.
- La ingeniería de características puede considerarse una fase de preprocesamiento en la que se reduce la dimensionalidad de los datos.
- Supongamos que d y p denotan la dimensionalidad del espacio de entrada y del subespacio, respectivamente, donde $p \in (0, d]$. La ingeniería de características es un mapa de un espacio euclíadiano d -dimensional a un espacio euclíadiano p -dimensional, es decir, $\mathbf{R}^d \rightarrow \mathbf{R}^p$.
- La dimensionalidad del subespacio suele ser mucho menor que la dimensionalidad del espacio, es decir, $p \ll d$, porque la mayor parte de la información suele estar en unas pocas características.

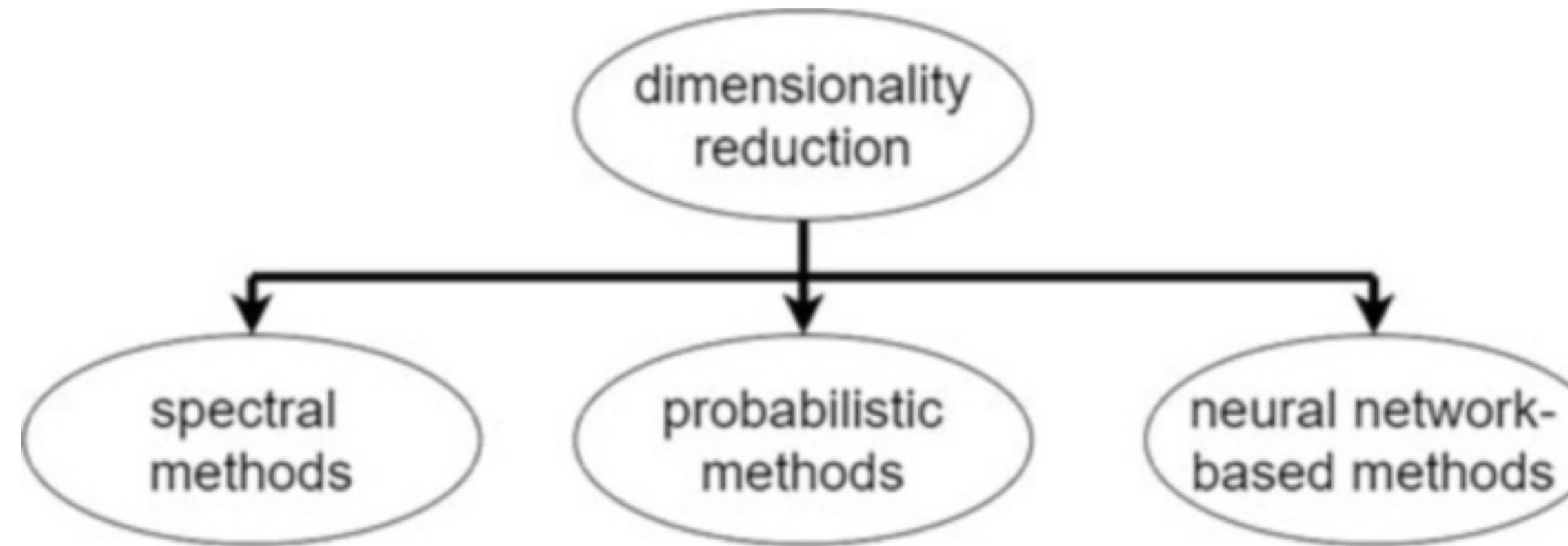
INGENIERÍA DE CARACTERÍSTICAS

- Hay dos formas de reducir el número de características de un dataset:
 - **selección de características:** se eligen unas características (las más informativas) y se descartan otras (por ejemplo, selección por información mutua)
 - **extracción de características:** se transforma el vector en otro con menor dimensión, pero todas las características iniciales intervienen en cada una de las características finales (por ejemplo, PCA)

REDUCCIÓN DE DIMENSIONALIDAD Y MANIFOLD LEARNING

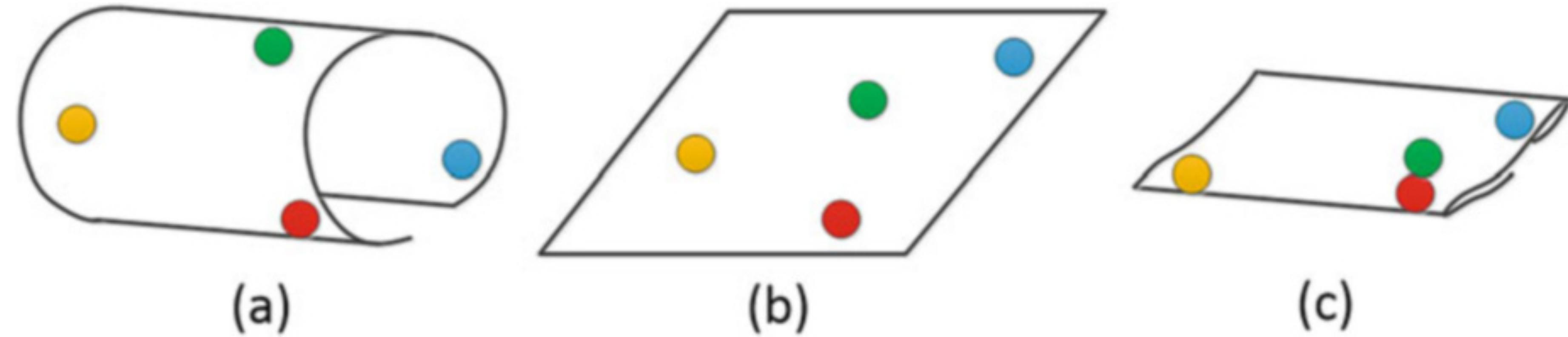
- La extracción de características también se conoce (en inglés) de forma indistinta como *dimensionality reduction, manifold learning, subspace learning, submanifold learning, manifold unfolding, embedding, encoding, representation learning*
- Las técnicas de manifold learning incluyen:
 - **Reducción de la dimensionalidad de los datos:** Producir una codificación compacta (comprimida) de baja dimensionalidad de un conjunto de datos de alta dimensionalidad dado.
 - **Preprocesamiento para el aprendizaje supervisado:** Simplificar, reducir y limpiar los datos para el posterior entrenamiento supervisado.
 - **Visualización de datos:** Ofrecer una interpretación de un conjunto de datos dado en términos de grados de libertad intrínsecos, normalmente como subproducto de la reducción de la dimensionalidad de los datos

REDUCCIÓN DE DIMENSIONALIDAD Y MANIFOLD LEARNING



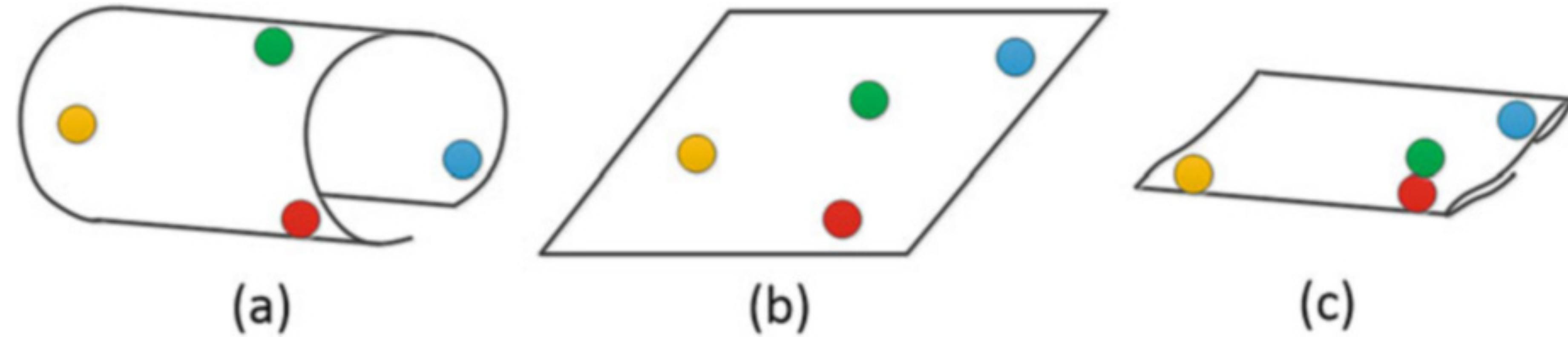
- En la reducción de la dimensionalidad, los puntos de datos se trasladan a un subespacio de menor dimensión, ya sea de forma lineal o no lineal.
- Los métodos de reducción de la dimensionalidad pueden agruparse en tres categorías: reducción de la dimensionalidad espectral, reducción de la dimensionalidad probabilística y reducción de la dimensionalidad basada en redes neuronales (artificiales)

REDUCCIÓN DE DIMENSIONALIDAD ESPECTRAL



- Consideremos varios puntos en un espacio euclídeo tridimensional. Supongamos que estos puntos se encuentran en un submanifold no lineal (a)
- Dos características pueden representar la mayor parte de la información de los puntos de datos, como se muestra en (b)
- Sin embargo, como se ve en (c), un método lineal de reducción de la dimensionalidad no puede encontrar correctamente una representación 2D subyacente correcta de los puntos de datos.

REDUCCIÓN DE DIMENSIONALIDAD ESPECTRAL



- Esto se debe a que un método lineal utiliza las distancias euclidianas entre los puntos, mientras que un método no lineal considera las distancias geodésicas a lo largo de la variedad no lineal.
- Los métodos espectrales de reducción de la dimensionalidad suelen tener una perspectiva geométrica e intentan encontrar el subespacio lineal o no lineal de los datos.
- Estos métodos suelen reducirse a un problema de valores propios generalizado.

REDUCCIÓN DE DIMENSIONALIDAD PROBABILÍSTICA

- Los datos pueden considerarse una variable aleatoria multidimensional.
- Es posible suponer una variable aleatoria latente de dimensión inferior, a la que está condicionada la variable aleatoria de los datos (véase el modo gráfico probabilístico en la figura).
- El punto de datos de alta dimensión x_i está condicionado a una variable aleatoria latente de baja dimensión z_i . Esta variable latente contiene la información más importante de los datos.

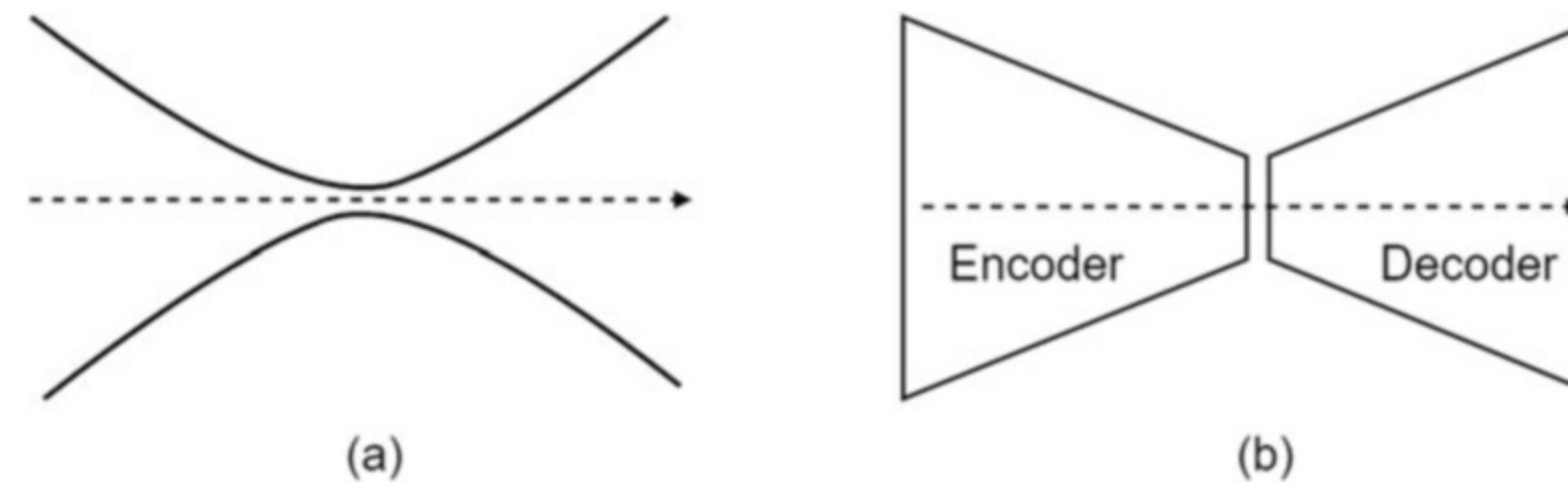


REDUCCIÓN DE DIMENSIONALIDAD PROBABILÍSTICA



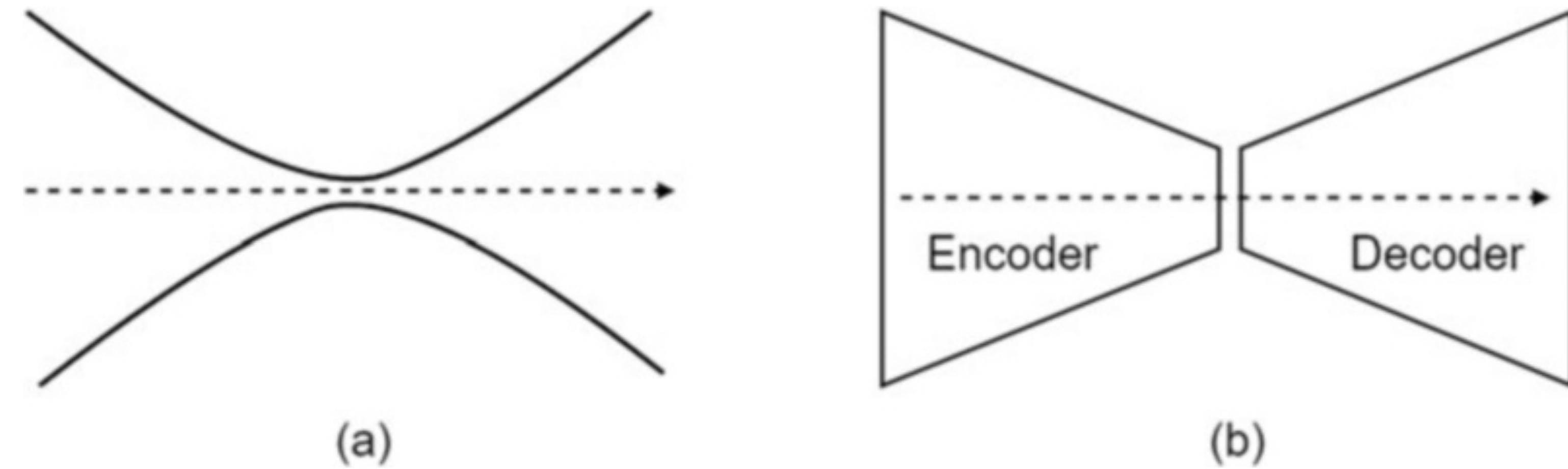
- En esta figura, el punto de vista puede modelarse mediante dos ángulos de izquierda a derecha y de abajo a arriba. Los puntos de datos son imágenes de alta dimensión y sus características son las luminosidades de los píxeles.
- Estas imágenes difieren sólo en términos de las dos perspectivas. Por lo tanto, los puntos de datos están condicionados a una variable latente bidimensional.
- Los métodos probabilísticos de reducción de la dimensionalidad intentan encontrar la variable latente de menor dimensión a partir de los datos de mayor dimensión.

REDUCCIÓN DE DIMENSIONALIDAD NEURONAL



- Es posible hacer pasar datos de alta dimensión a través de un cuello de botella de dimensión inferior y reconstruirlos después.
- Como los datos deben pasar por el cuello de botella y reconstruirse casi perfectamente después, las características más informativas de los datos deben conservarse en el cuello de botella. Por lo tanto, la representación de los datos en el cuello de botella es un candidato adecuado para la reducción de la dimensionalidad.
- Este cuello de botella puede implementarse utilizando una red neuronal autocodificadora, en la que los datos se comprimen entre el codificador y el decodificador

REDUCCIÓN DE DIMENSIONALIDAD NEURONAL



- Los métodos de reducción dimensional basados en redes neuronales adoptan una perspectiva basada en la teoría de la información. Estos métodos utilizan redes neuronales profundas o superficiales para la extracción de características.
- Dedicaremos un proyecto completo (en VA2) a este tipo de representaciones.

PERSPECTIVA HISTÓRICA DE LAS TÉCNICAS ESPECTRALES

- Varianza, lineal: Principal Component Analysis - PCA (Pearson, 1901)
- Varianza, lineal: Fisher Discriminant Analysis - FDA (Fisher, 1936)
- Similaridades / distancias, lineal: Multidimensional Scaling (MDS) (varios, sobre 1960)
- Distancias, no lineal: Mapa de Sammon (1969)
- Varianza, no lineal: Kernel PCA, Kernel FDA (1999)
- Geodésicas, no lineal: Isometric Mapping (Isomap) y Locally Linear Embedding (LLE) (2000)
- Grafos, no lineal: Clustering espectral - Laplacian Eigenmap (2001)
- Unificación de técnicas: Semidefinite embedding, maximum variance unfolding (2004-2006)

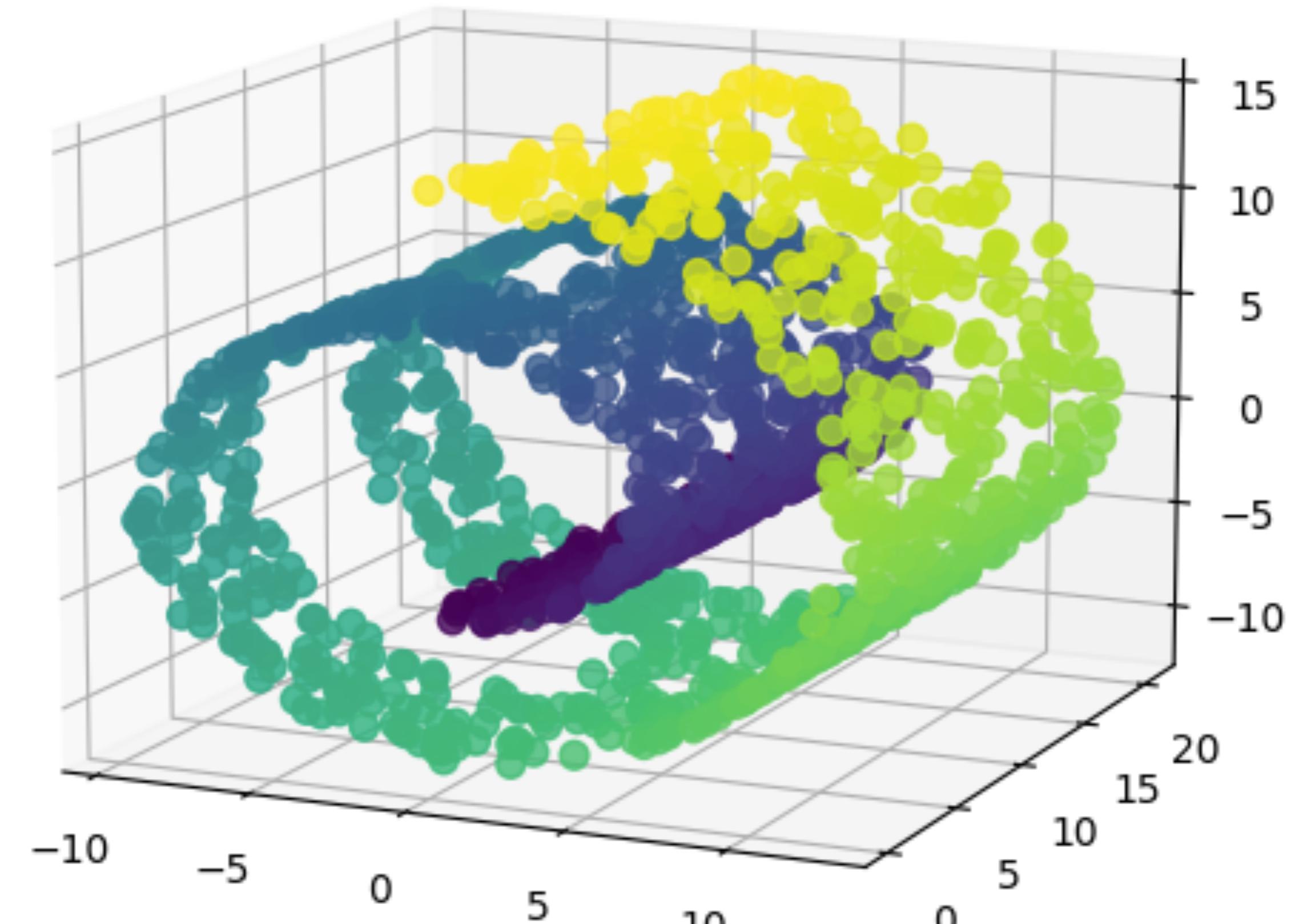
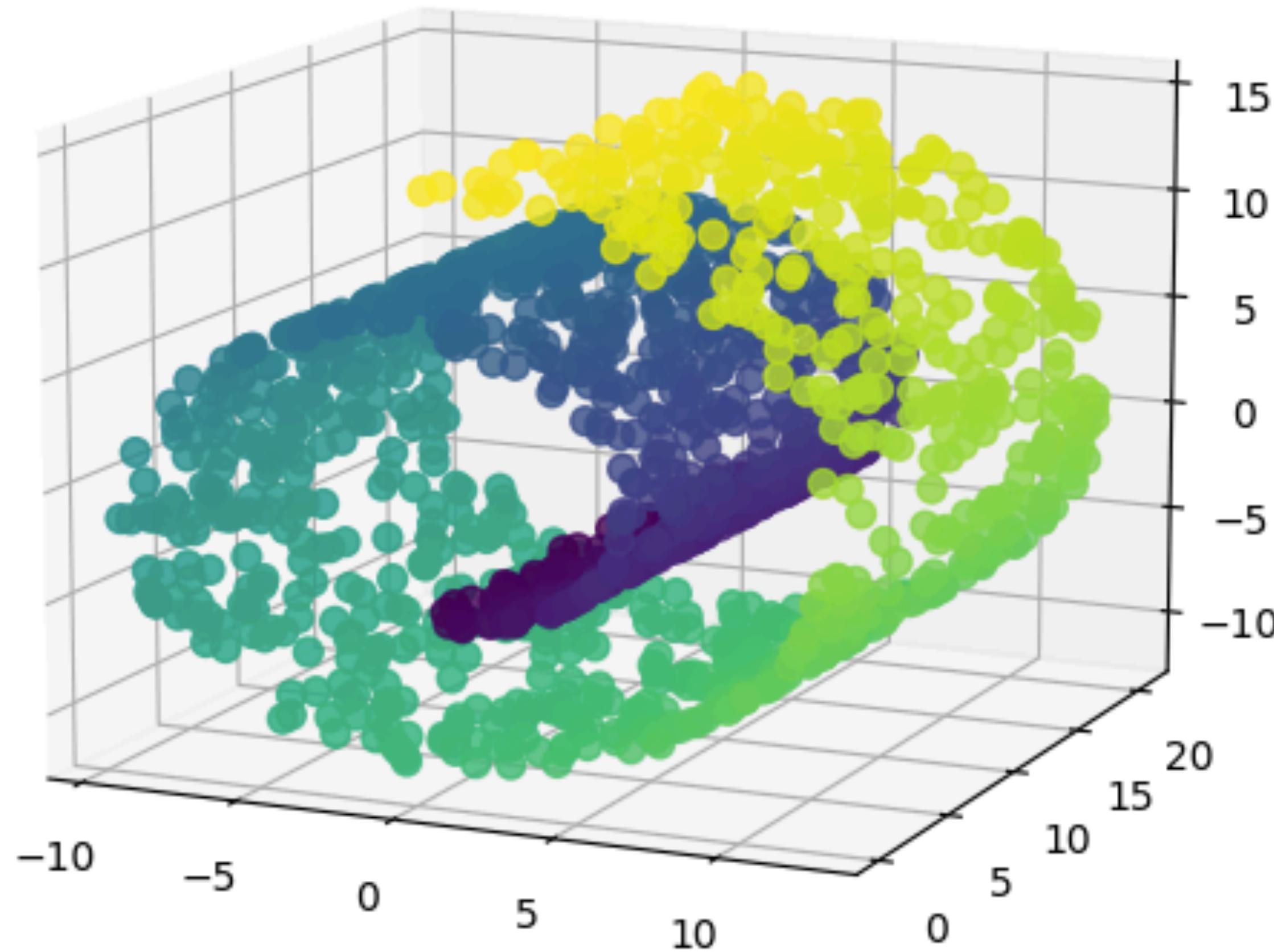
PERSPECTIVA HISTÓRICA DE LAS TÉCNICAS PROBABILÍSTICAS

- PCA probabilístico, NCA (Neighbourhood Component Analysis), Bayesian Metric Learning, Random Projection, Sufficient Dimension Reduction, Kernel Dimension Reduction y otras (muchas) técnicas relacionadas (desde 1997): se supone que los datos se obtienen añadiendo ruido a una transformación de una variable aleatoria latente
- Stochastic Neighbour Embedding - SNE (2003): mantiene la probabilidad de que un punto sea vecino de otros en el subespacio
- Uniform Manifold Approximation and Projection - UMAP (2018) - hipótesis similares a SNE

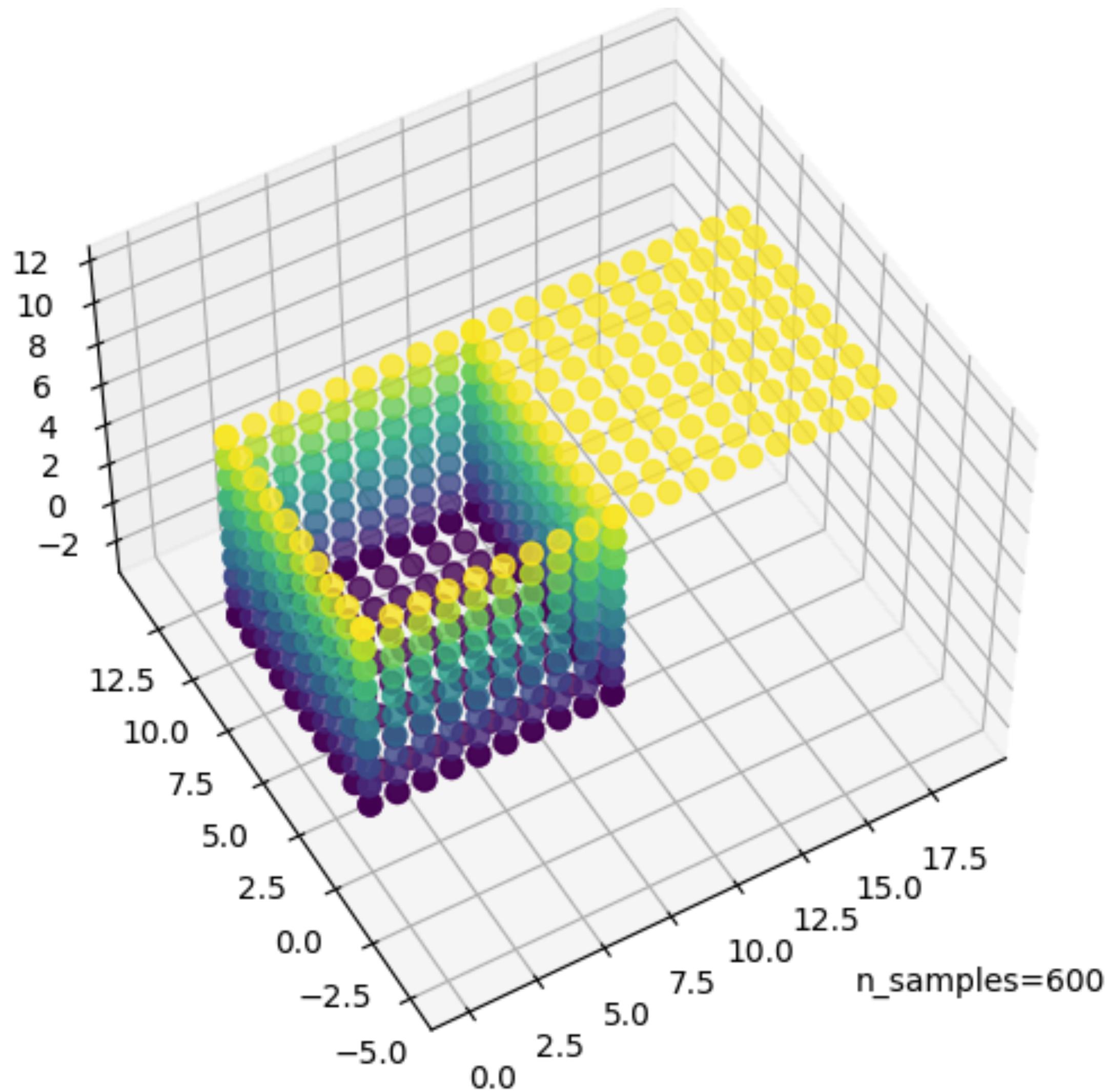
PERSPECTIVA HISTÓRICA DE LAS TÉCNICAS NEURONALES

- Deep Metric Learning (aprende matriz de distancias entre datos) (2019)
- Siamese networks / one shot learning / contrastive loss (aprende embeddings de imágenes y la forma de calcular las distancias entre los mismos) (desde 1993)
- Adversarial learning (2014)
- Autoencoders variacionales (2014) (veremos autoencoders y autoencoders variacionales en VA2)
- Graph Neural Networks - latent graph learning (problema abierto)

PROBLEMAS DE EJEMPLO: SWISS ROLL



PROBLEMAS DE EJEMPLO: OPEN BOX



MÉTODOS ESPECTRALES

PRESERVACIÓN DE LA DISTANCIA

- Distancias espaciales
 - Espacio métrico, distancias, normas y producto escalar
 - Escalamiento multidimensional
 - Cartografía no lineal de Sammon
- Distancias en grafos
 - Distancia geodésica y distancia en grafos
 - Isomap
- Otras distancias
 - Kernel PCA

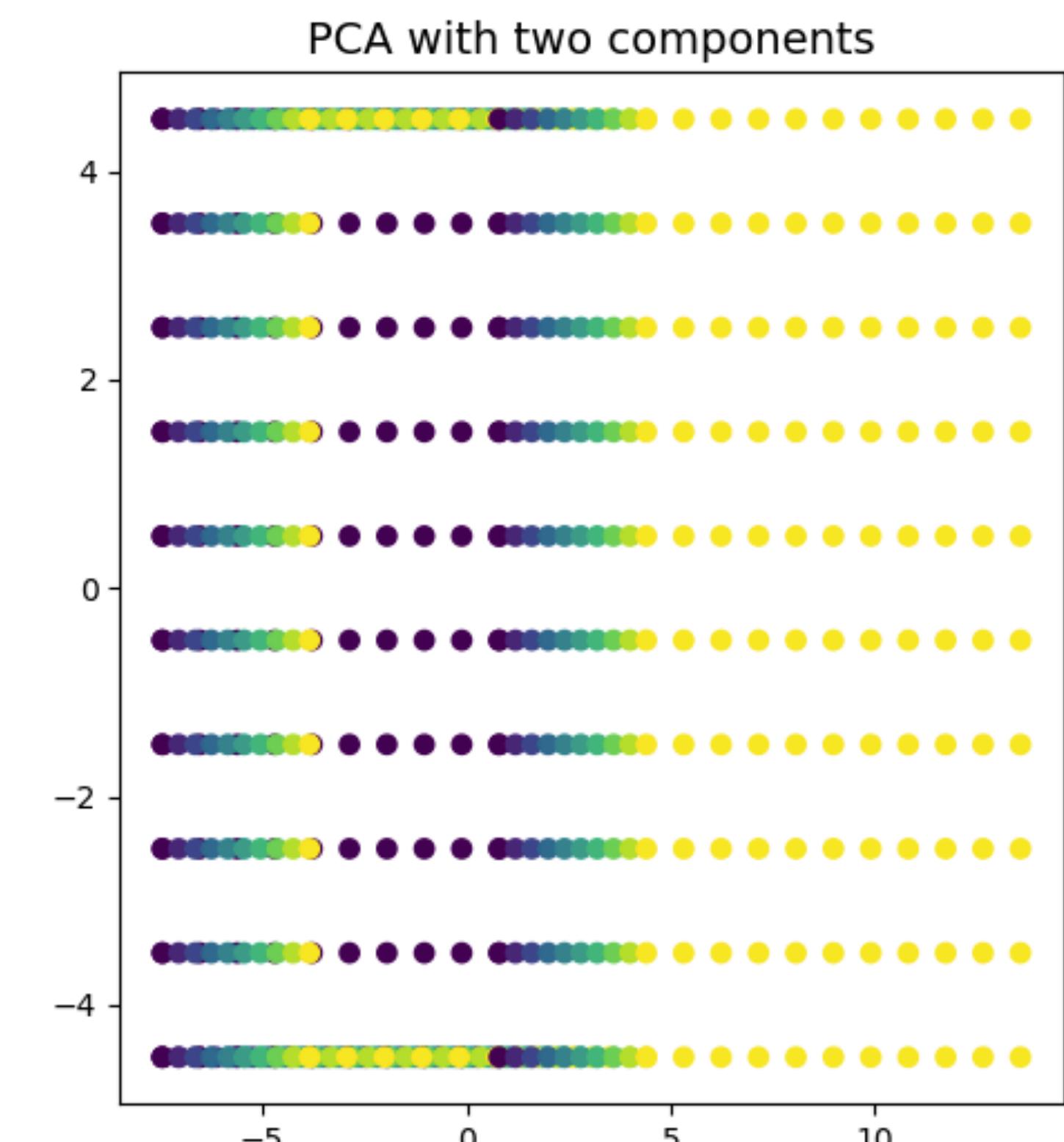
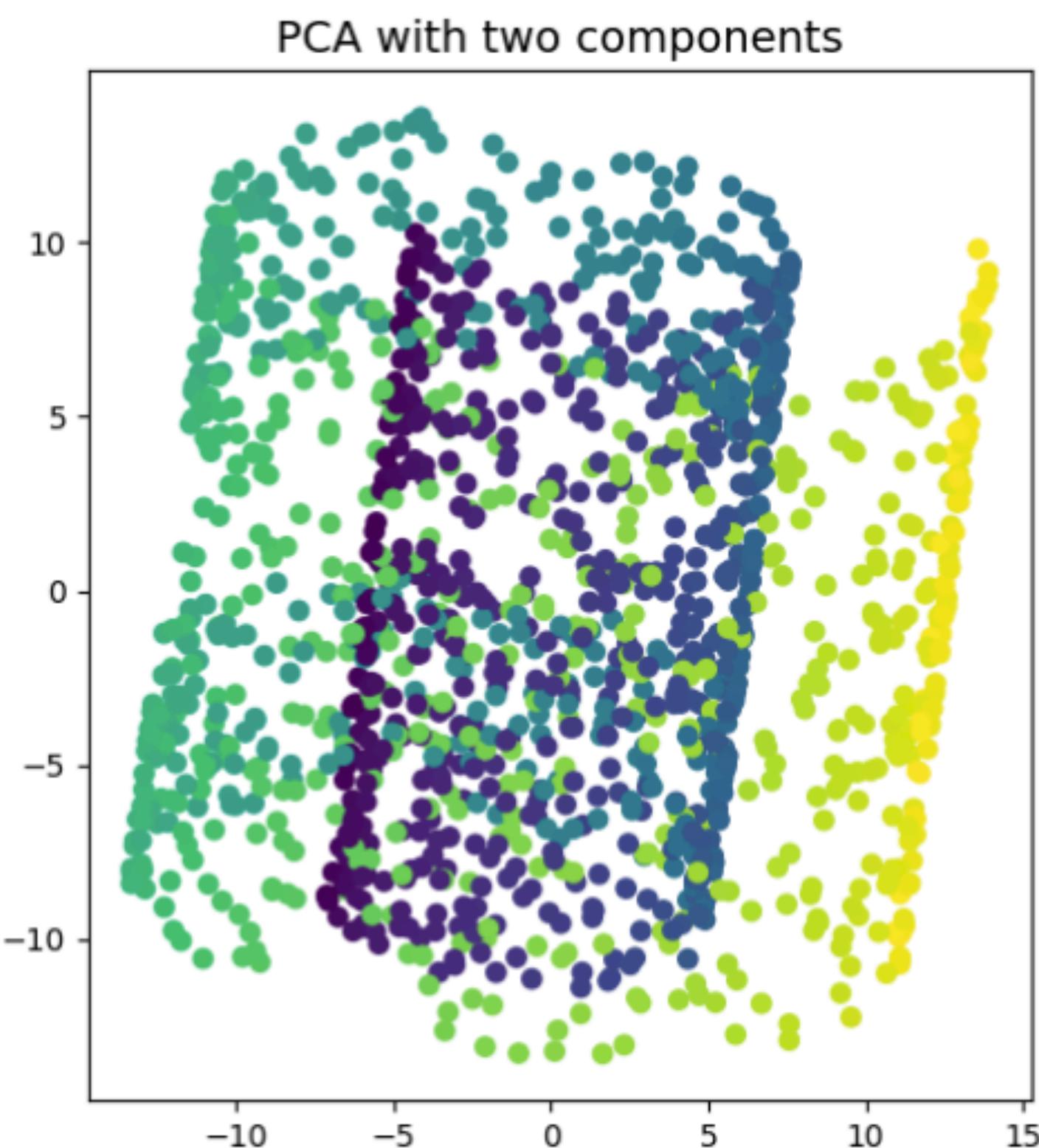
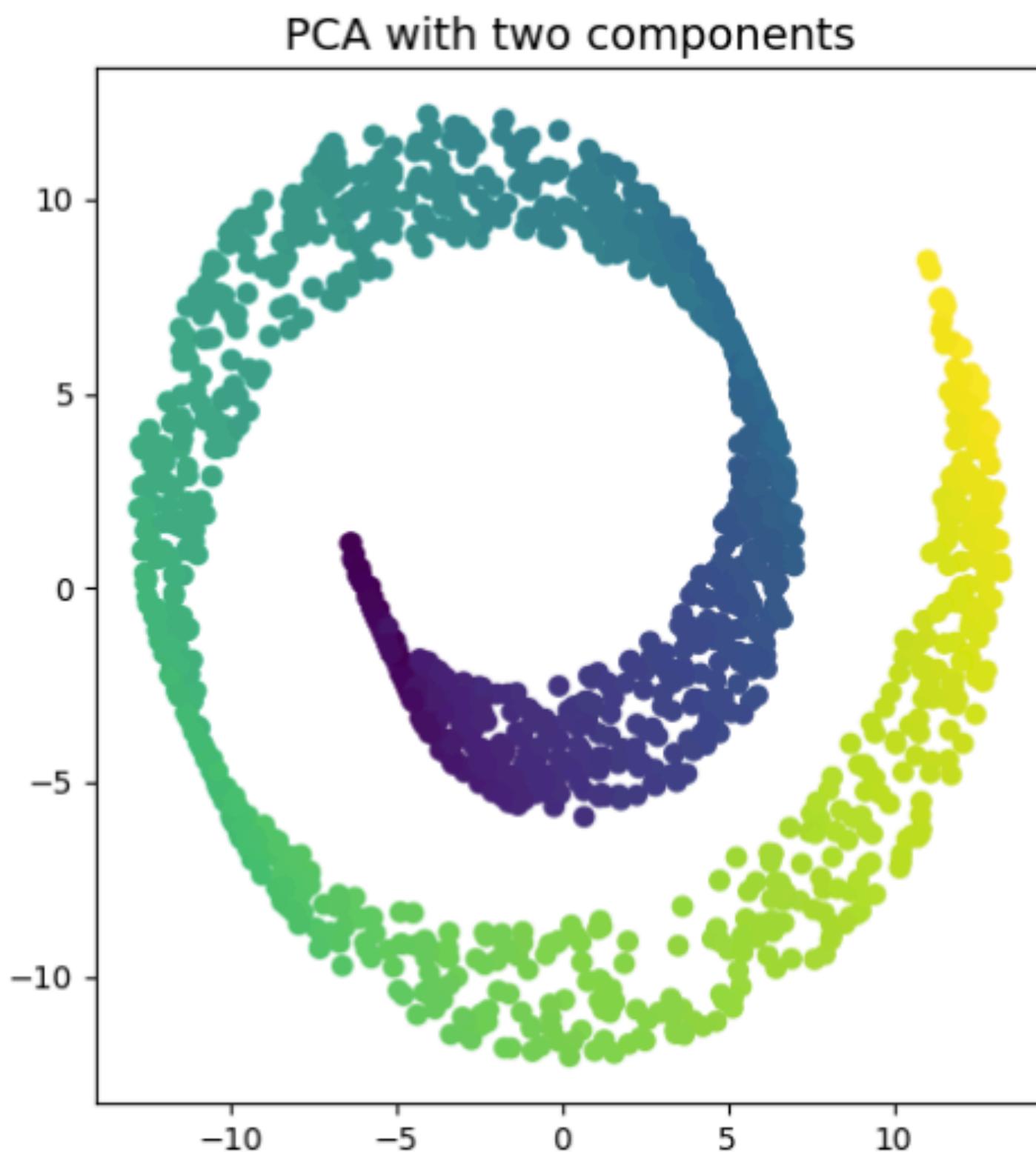
DISTANCIAS ESPACIALES

- Espacio métrico, distancias, normas y producto escalar
- Escalamiento multidimensional
- Cartografía no lineal de Sammon

ESPAZIO MÉTRICO, DISTANCIAS, NORMAS Y PRODUCTO ESCALAR

- Suponemos conocidos los conceptos de distancia, norma, producto escalar, y varianza / matriz de covarianzas
- Partimos de un espacio de dimensión alta + una función de distancia
- Queremos obtener una representación en un espacio de menor dimensión y en el que también haya una distancia definida
- La representación ideal debería cumplir que la matriz de distancias en el espacio final coincida con (o se parezca a) la matriz de distancias en el espacio inicial
- En el caso de que los espacios inicial y final estén compuestos de vectores de números reales y que la distancia sea la euclídea, la solución de este problema consiste en diagonalizar la matriz de covarianzas de los datos (PCA)

SOLUCIONES PCA DE LOS PROBLEMAS DE EJEMPLO



ESCALAMIENTO MULTIDIMENSIONAL

- MDS es similar a PCA (en lugar de mantener las distancias euclídeas entre los puntos se mantienen los productos escalares entre los mismos)
- PCA requiere que se conozcan las coordenadas de los puntos, pero MDS solo necesita una matriz de similaridades (o distancias) entre los elementos del dataset,
- Requiere más memoria y es más difícil de generalizar a puntos no vistos
- Se puede extender a distancias no euclídeas o a diferentes tipos de disimilaridades entre los objetos

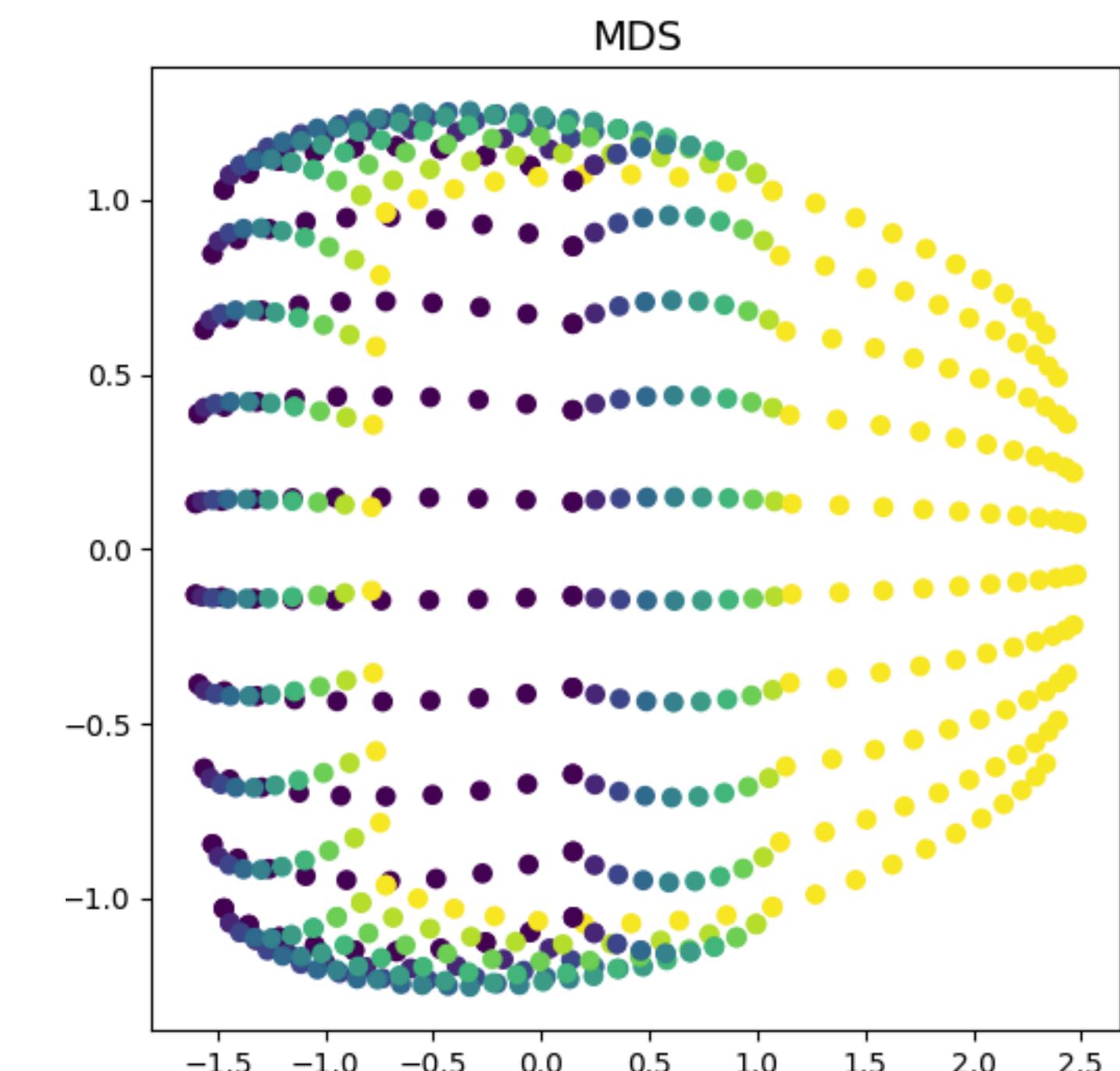
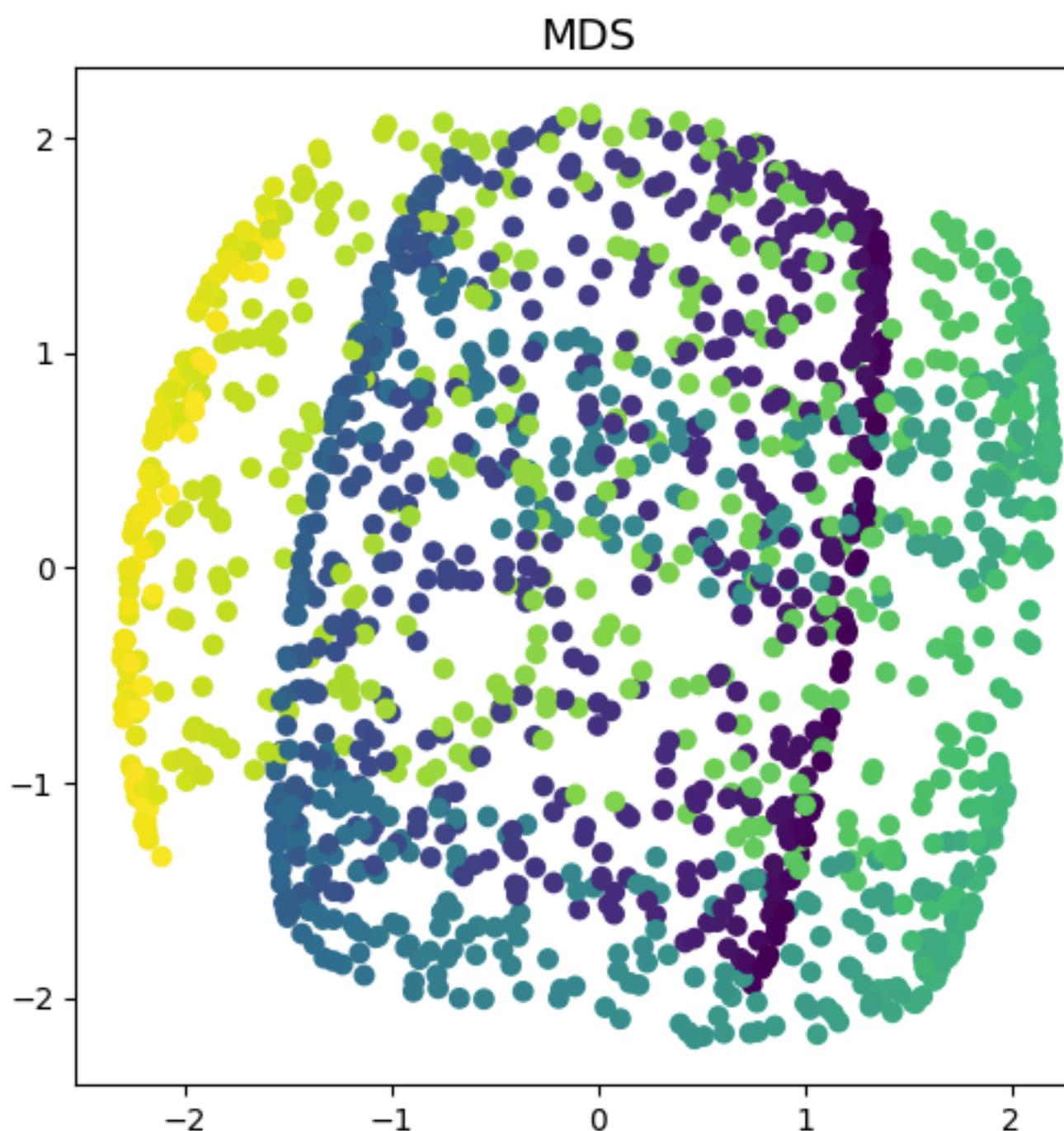
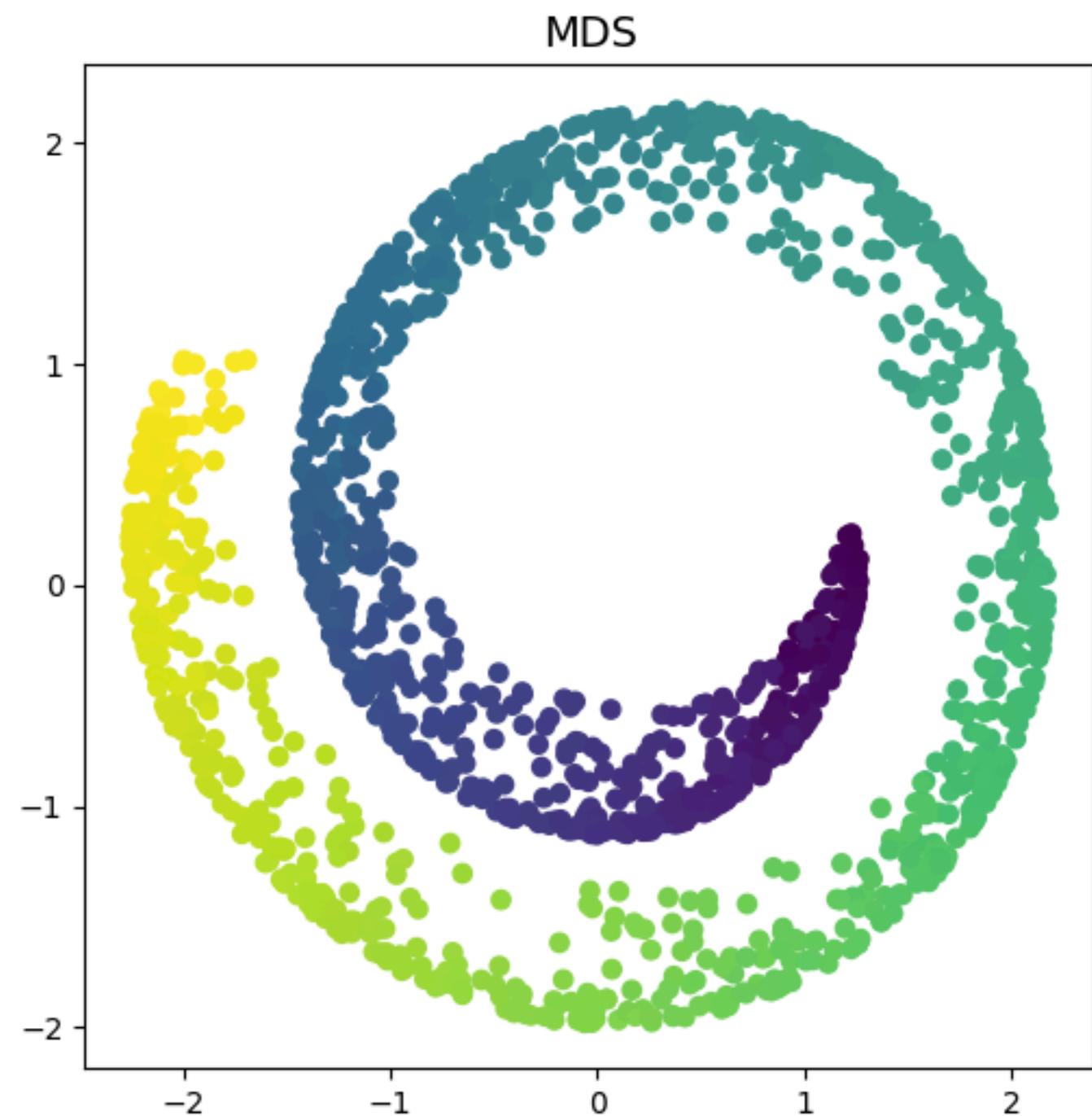
ESCALAMIENTO MULTIDIMENSIONAL

- El punto de partida es la matriz de productos escalares entre todos los elementos del dataset (matriz de Gram)
- La solución es la misma que PCA en este caso (se diagonaliza la matriz de Gram en MDS o la matriz de covarianzas en PCA, en ambos casos se obtiene la misma matriz de paso si los datos están centrados)
- Si el punto de partida es una matriz de distancias, se emplea el que la distancia al cuadrado entre dos vectores sea el producto escalar de la diferencia de ambos por sí misma y se obtiene la matriz de Gram a partir de la matriz de distancias
- Si se emplea una disimilitud no generada a partir de un producto escalar, hay que asegurarse de que la matriz de Gram siga siendo simétrica y definida positiva

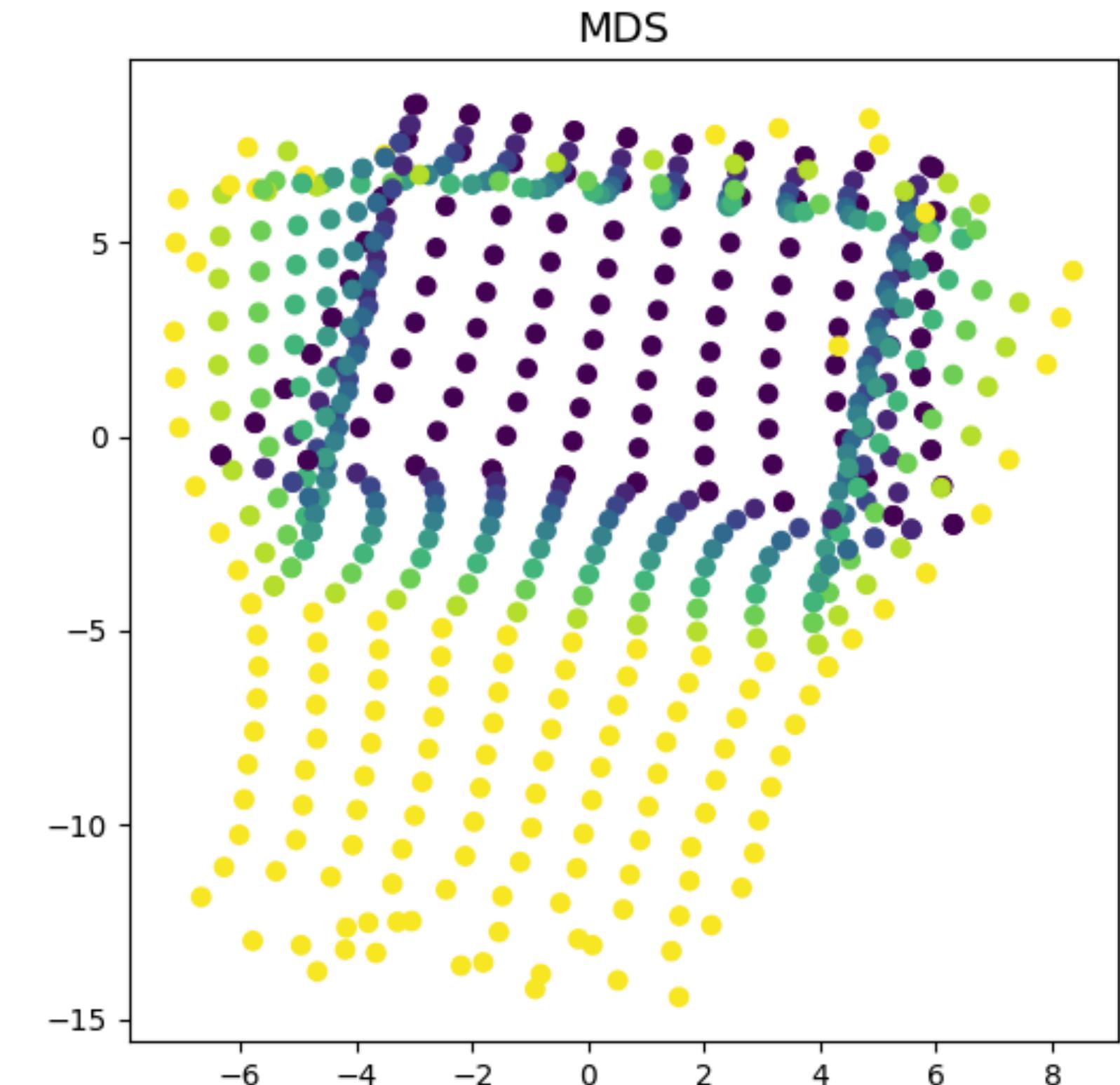
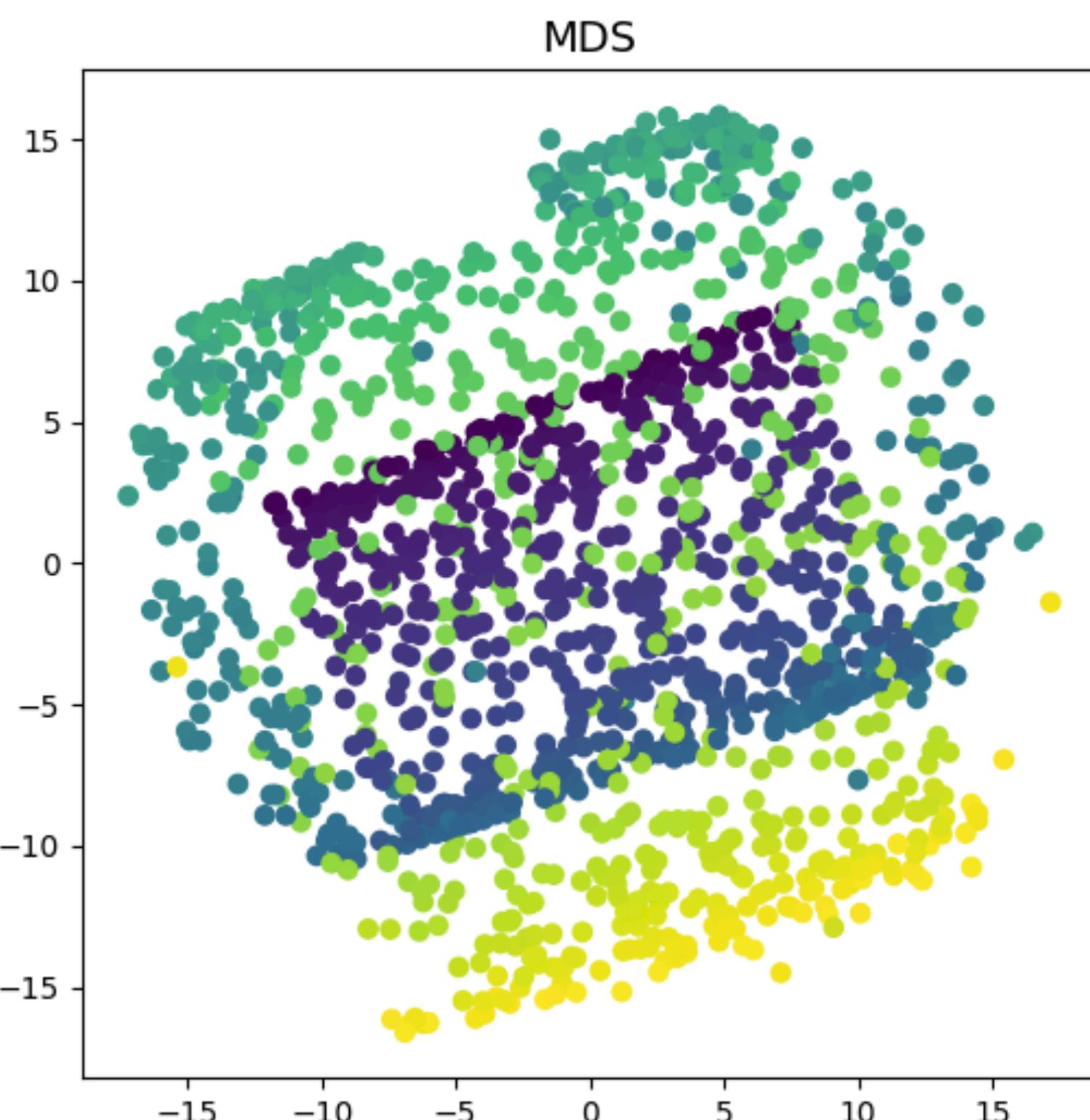
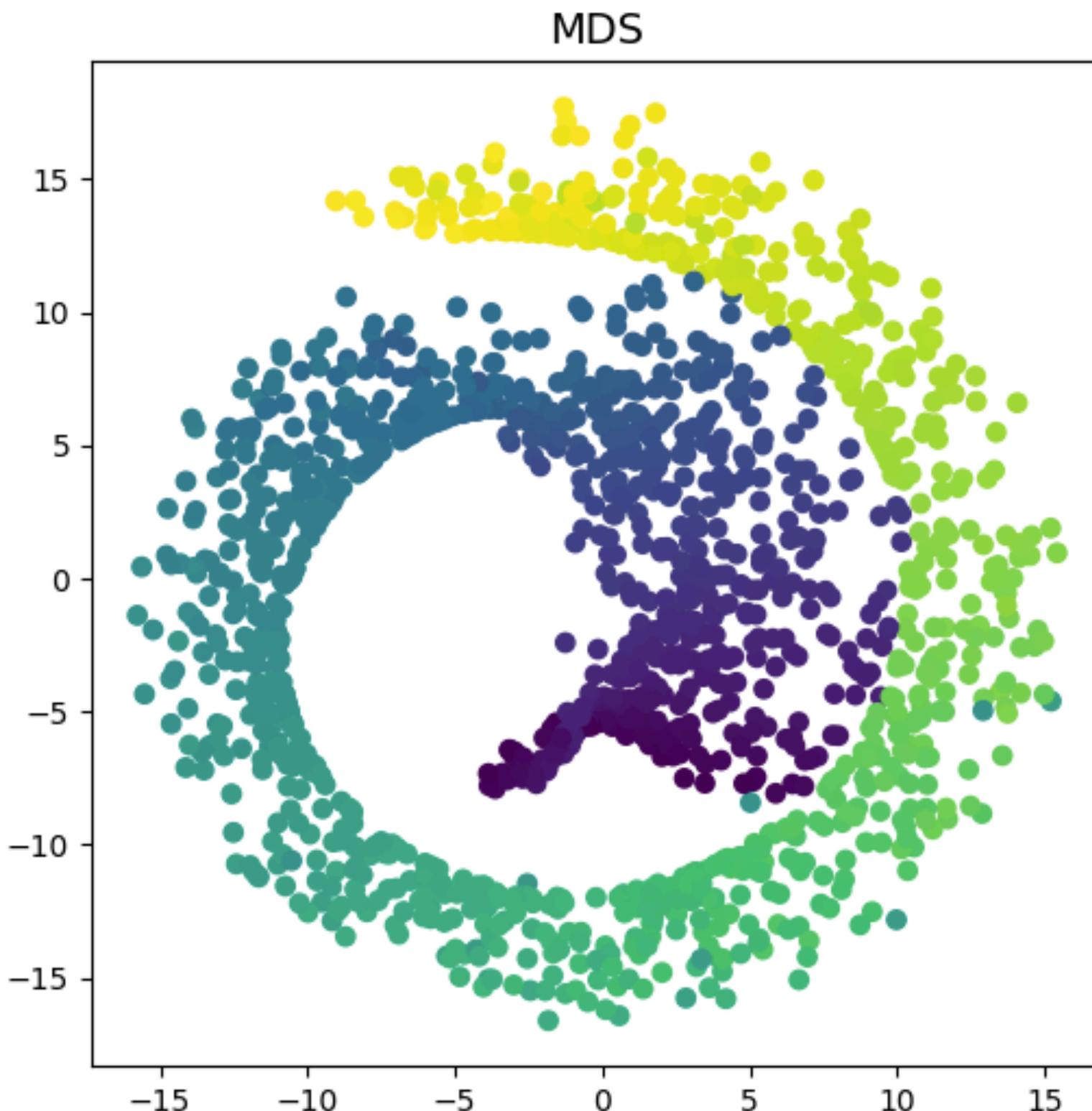
ESCALAMIENTO MULTIDIMENSIONAL

- sklearn implementa el algoritmo SMACOF, que es diferente de la versión métrica de MDS
- SMACOF minimiza iterativamente una función de stress, y es un algoritmo no determinista
- Solamente requiere que la matriz de disimilaridades sea simétrica

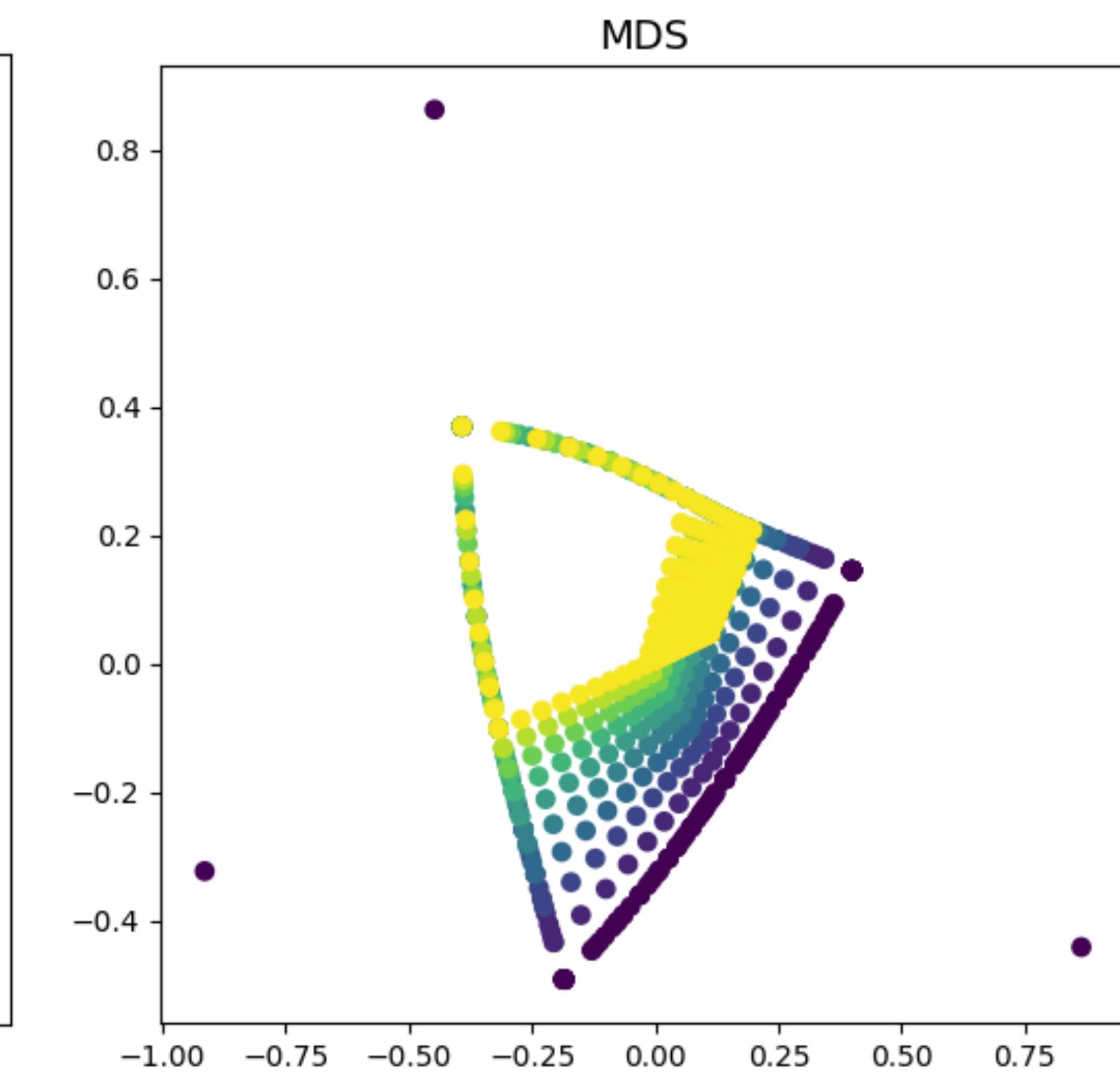
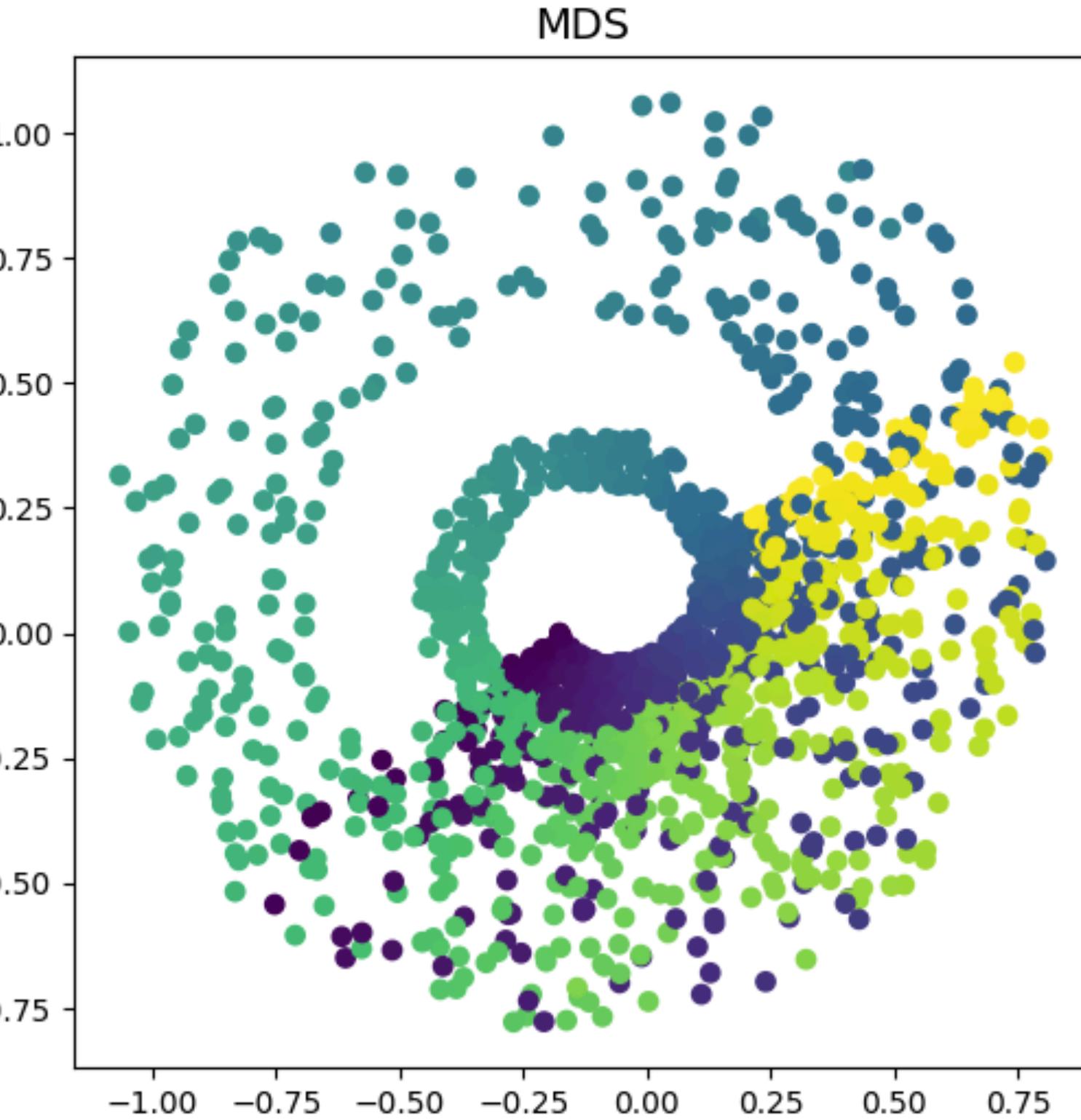
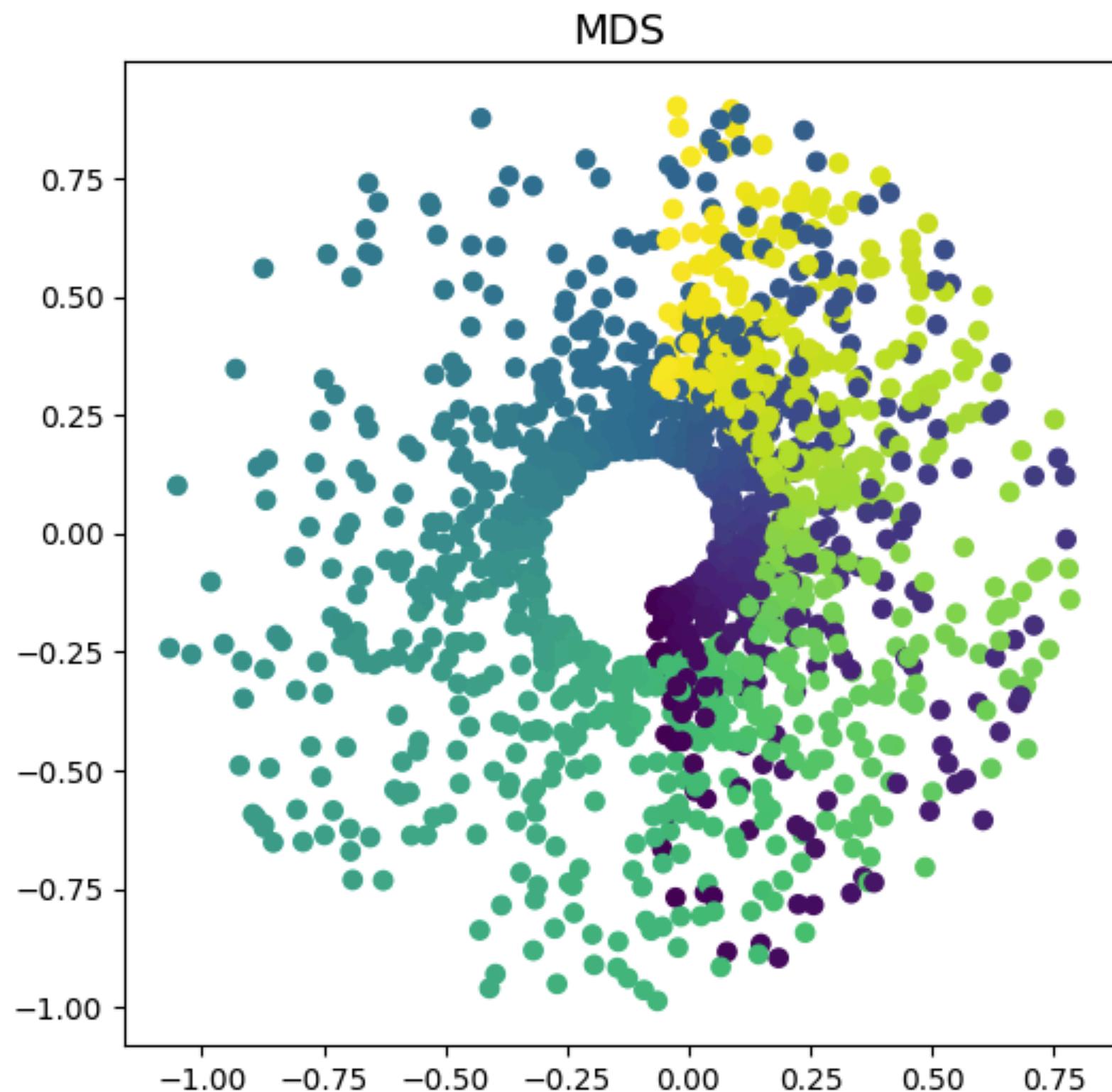
SOLUCIONES MDS DE LOS PROBLEMAS DE EJEMPLO (MDS MÉTRICO)



SOLUCIONES MDS DE LOS PROBLEMAS DE EJEMPLO (SMACOF)



SOLUCIONES MDS DE LOS PROBLEMAS DE EJEMPLO (SMACOF + D. COSENO)

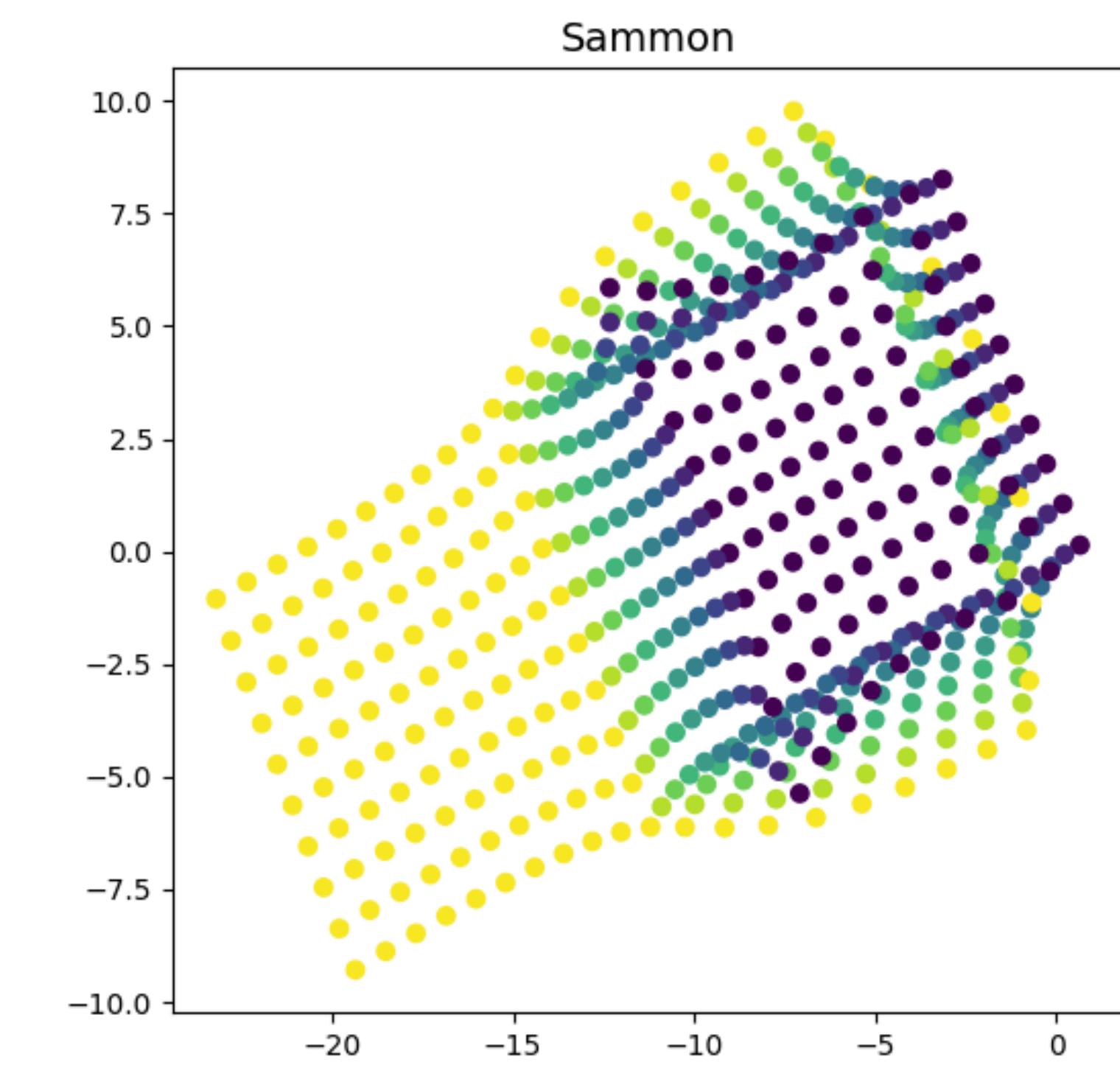
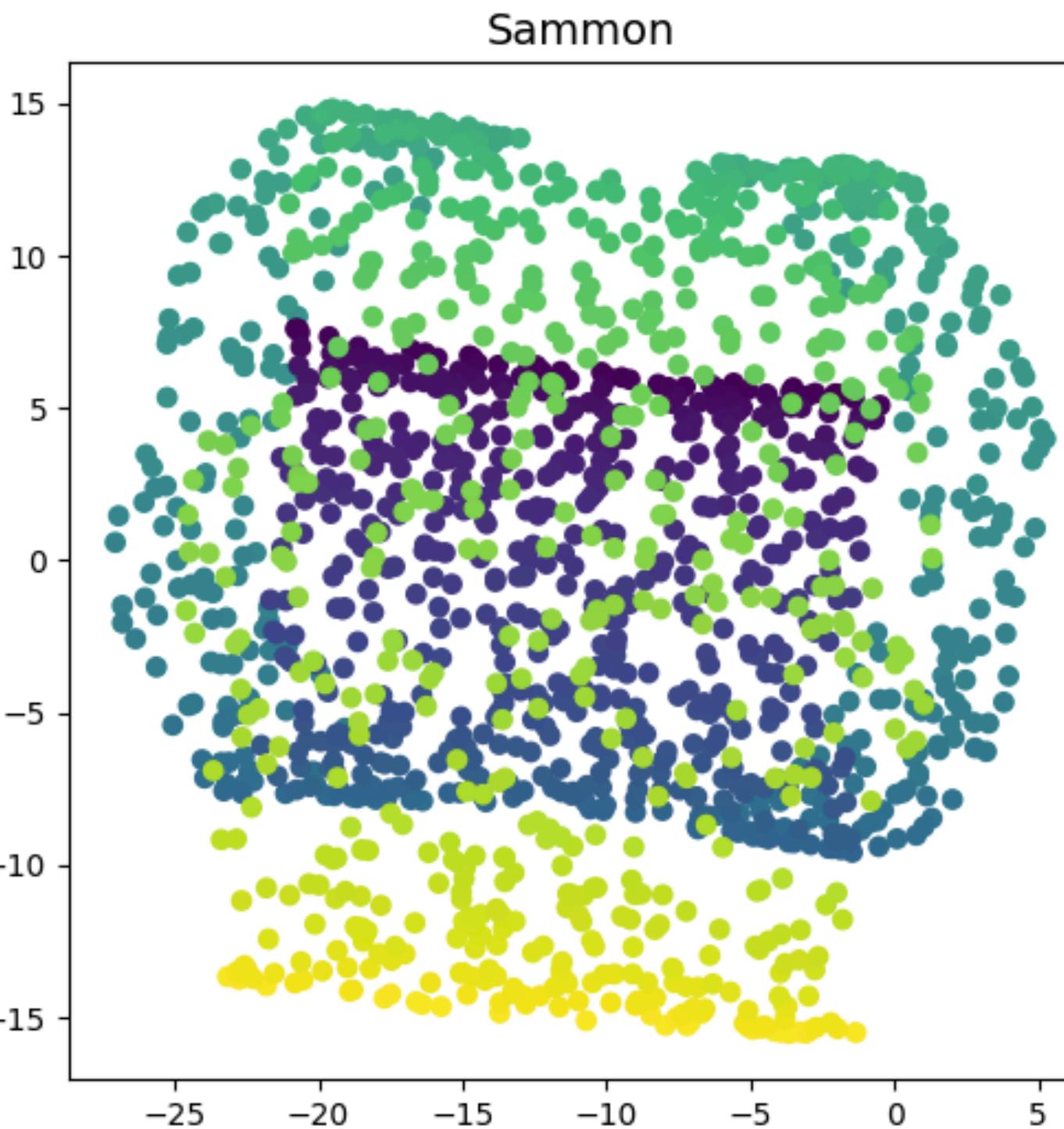
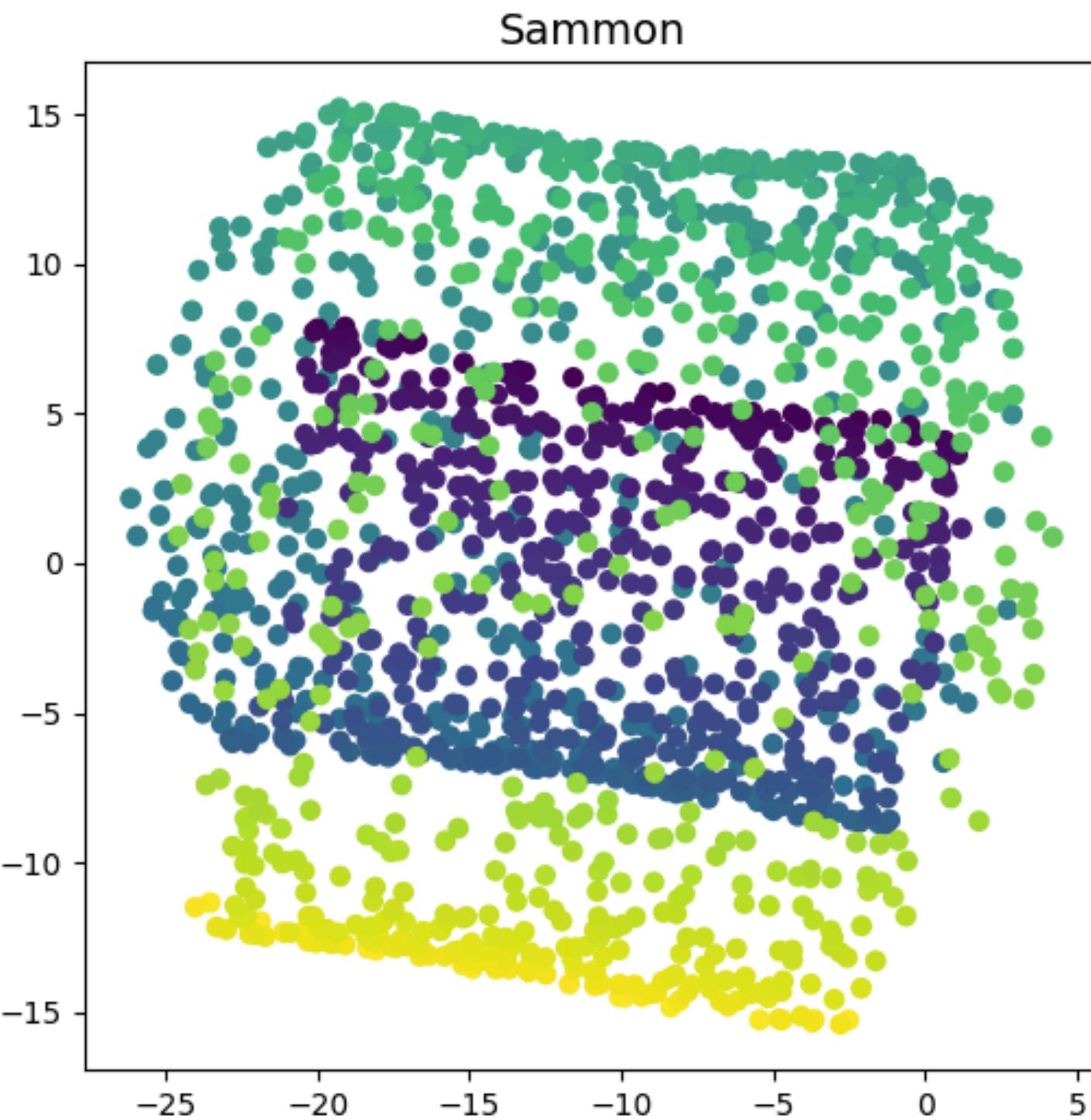


CARTOGRAFÍA NO LINEAL DE SAMMON

$$E_{\text{NLM}} = \frac{1}{c} \sum_{\substack{i=1 \\ i < j}}^N \frac{(d_{\mathbf{y}}(i, j) - d_{\mathbf{x}}(i, j))^2}{d_{\mathbf{y}}(i, j)}$$

- minimiza una función de esfuerzo (diferencia relativa entre las distancias en la proyección y en el espacio inicial)
- maneja subespacios no lineales (que no estén excesivamente doblados)
- no puede generalizarse a puntos nuevos
- procedimiento de optimización ineficiente comparado con PCA

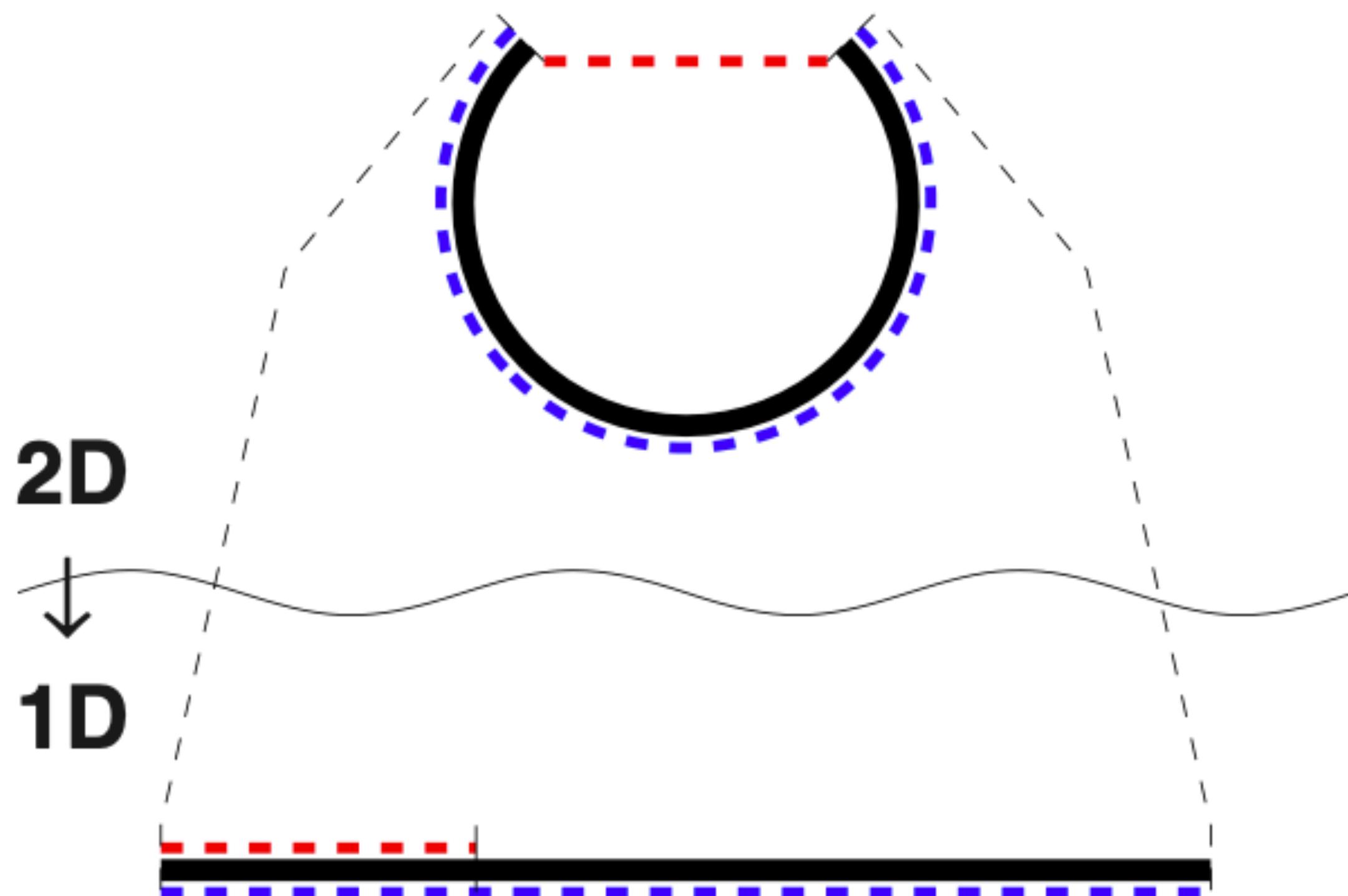
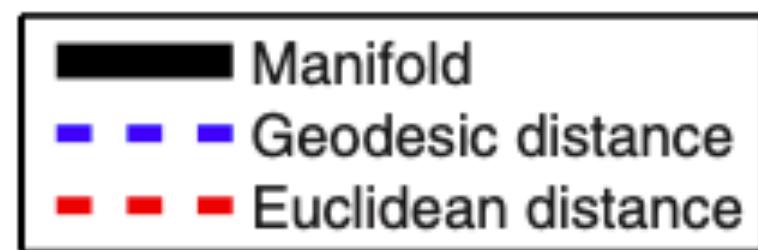
SOLUCIONES SAMMON DE LOS PROBLEMAS DE EJEMPLO



DISTANCIAS EN GRAFOS

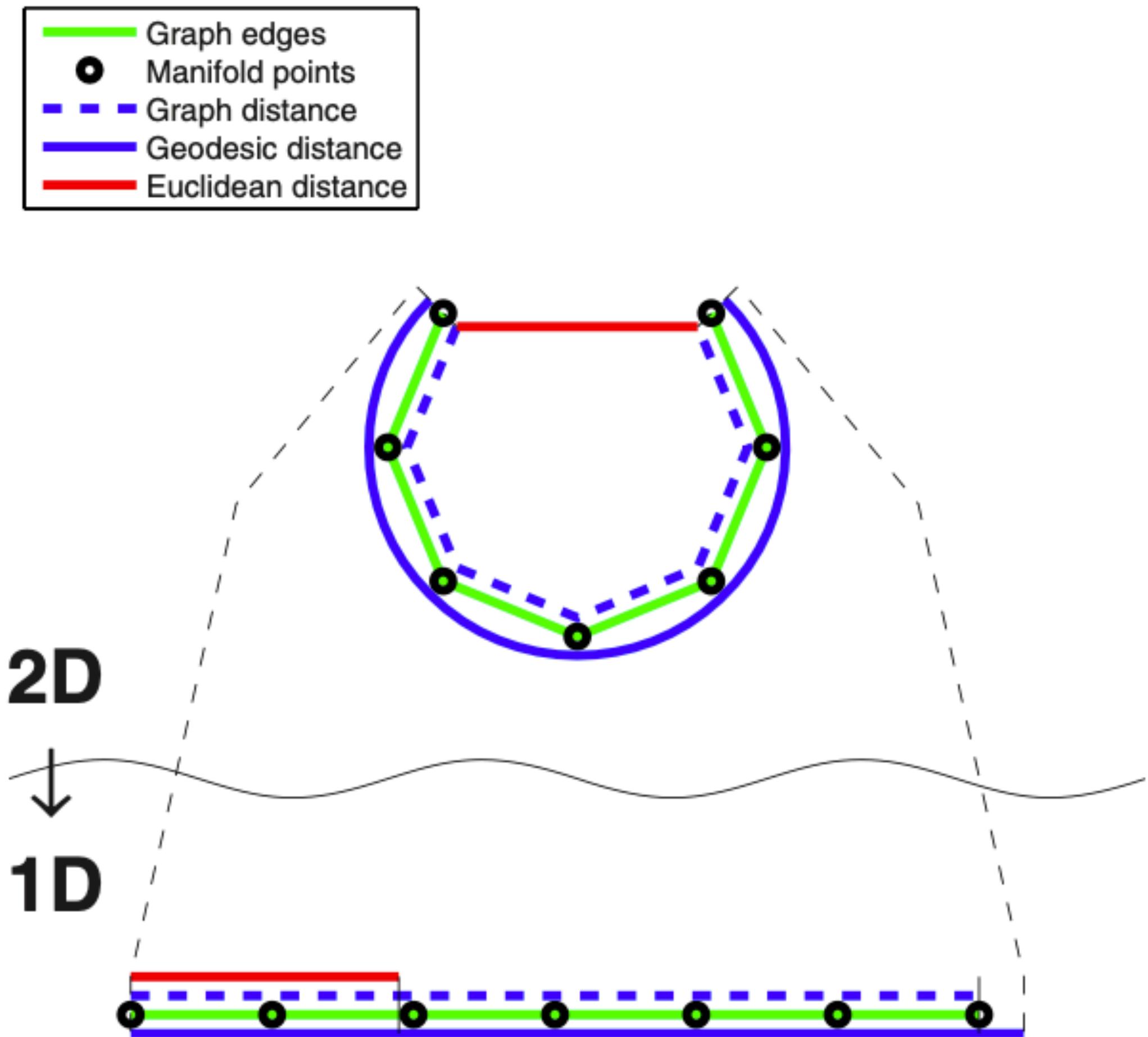
- Distancia geodésica y distancia en grafos
- Isomap

DISTANCIA GEODÉSICA Y DISTANCIA EN GRAFOS



- La distancia euclíadiana no puede conservarse fácilmente, salvo para distancias muy pequeñas
- En el espacio bidimensional, la distancia entre los extremos de la curva esta distancia es corta porque el colector está plegado sobre sí mismo.
- A diferencia de la distancia euclíadiana, la distancia geodésica se mide a lo largo del colector.
- Por consiguiente, no depende tanto como la métrica euclídea de una incrustación concreta de la variedad. En el caso de la curva C, la distancia geodésica es la misma en espacios unidimensionales y bidimensionales.

DISTANCIA GEODÉSICA Y DISTANCIA EN GRAFOS



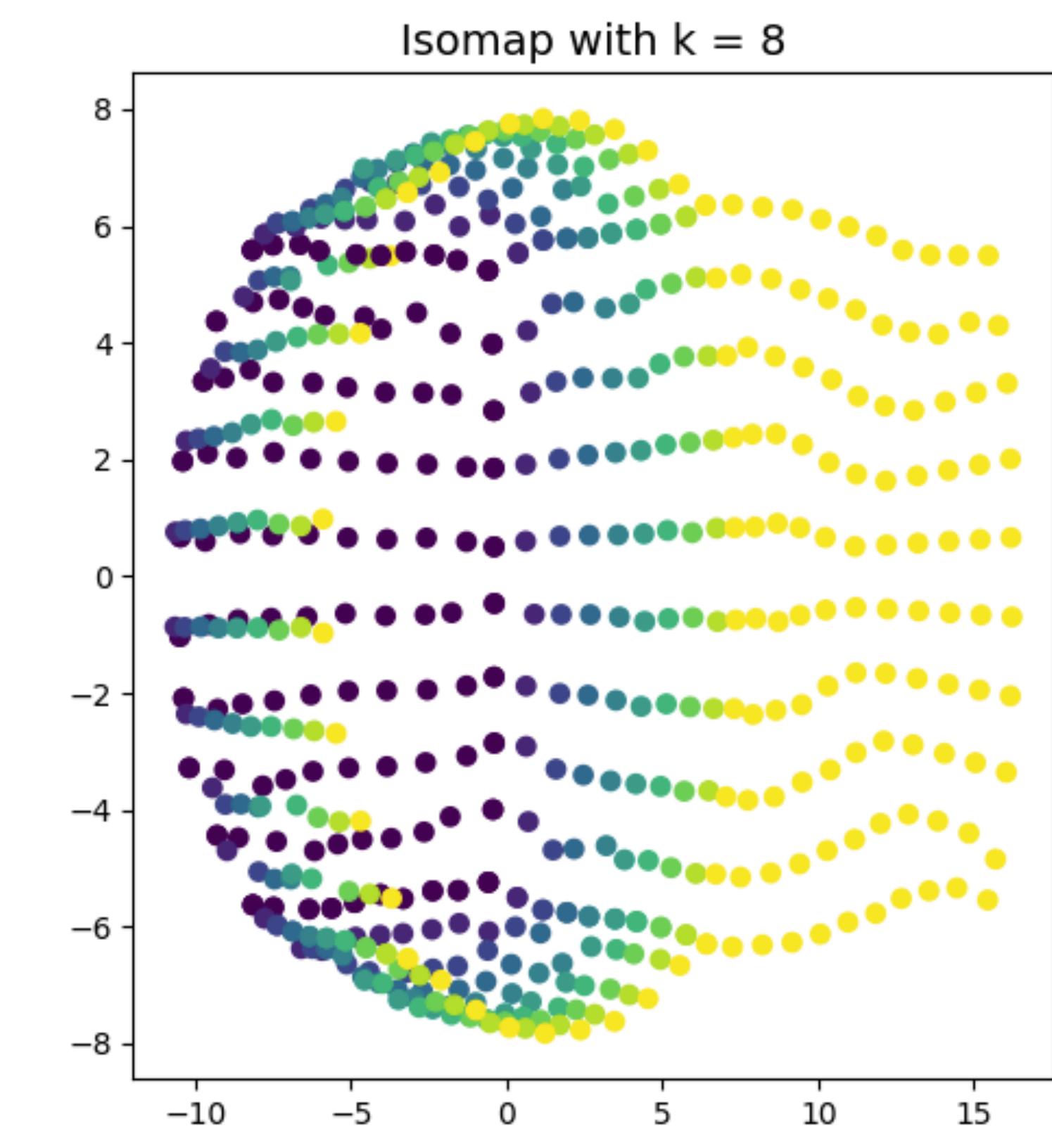
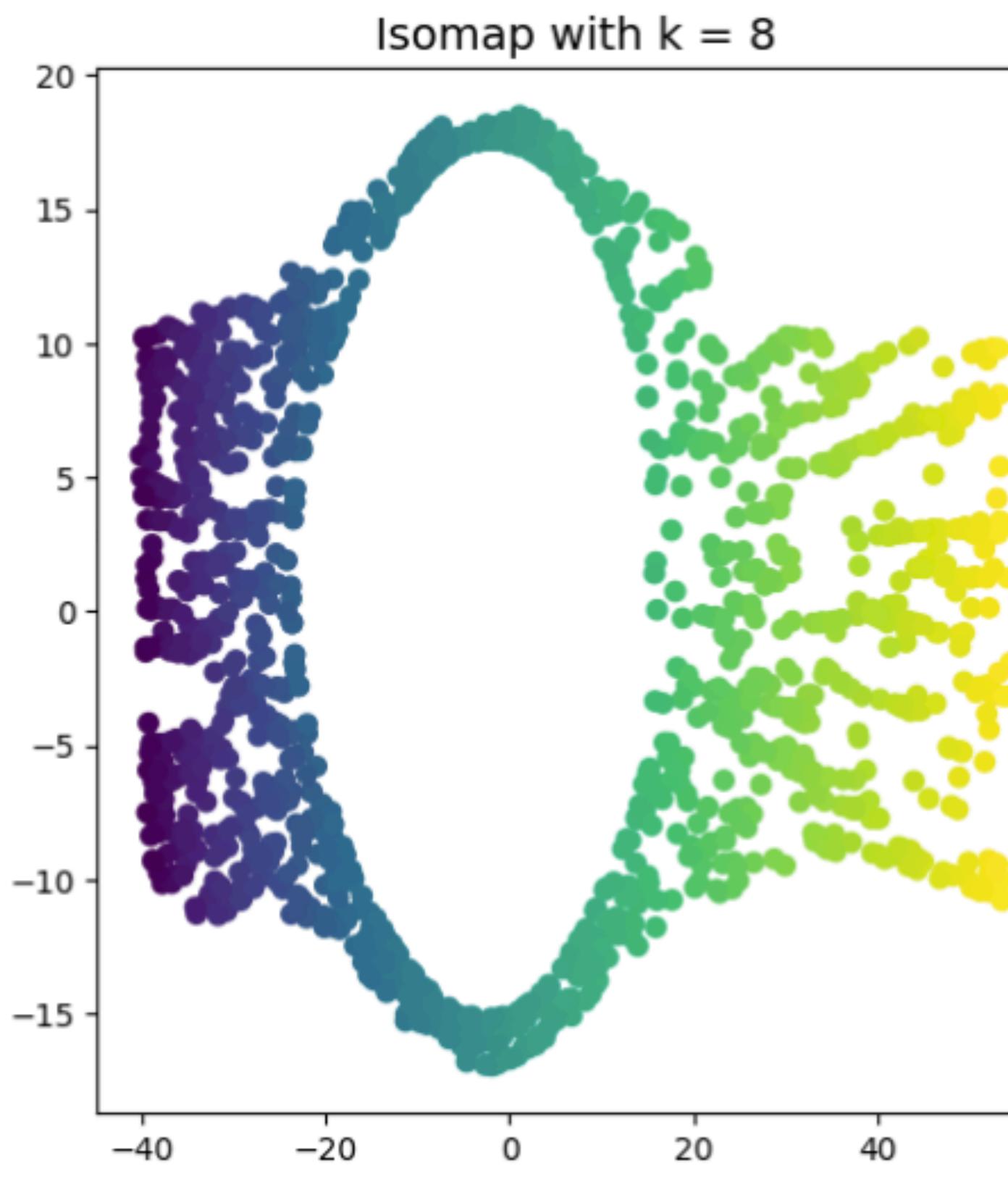
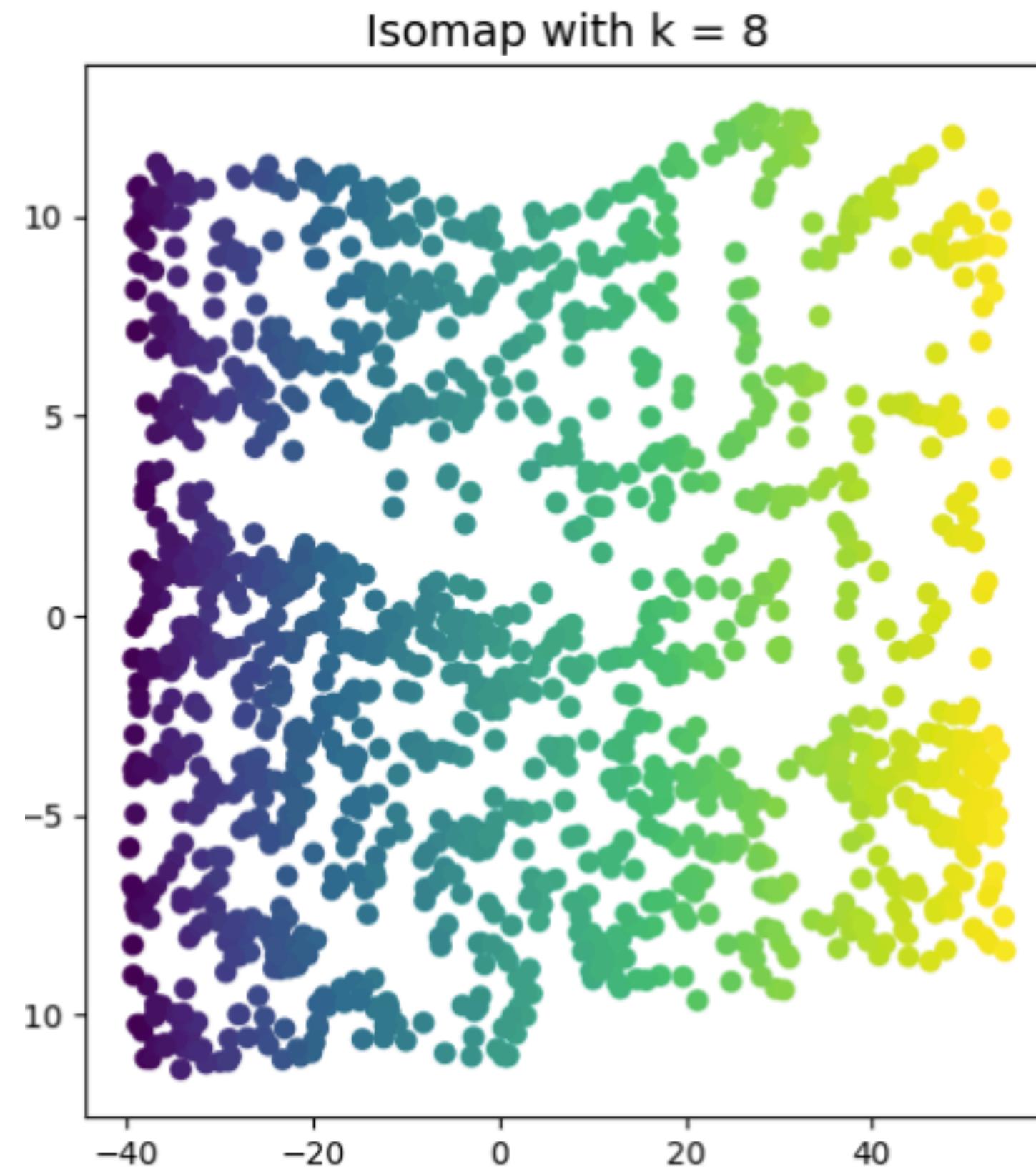
- En este caso, sólo se conocen algunos puntos.
- Para aproximar la distancia geodésica, se asocian vértices a los puntos y se construye un grafo.
- La distancia del grafo puede medirse sumando las aristas del grafo a lo largo del camino más corto entre ambos extremos de la curva.
- Ese camino más corto puede calcularse mediante el algoritmo de Dijkstra. Si el número de puntos es suficientemente grande, la distancia gráfica proporciona una buena aproximación a la distancia geodésica real.

ISOMAP

1. Build a graph with either the K -rule or the ϵ -rule.
2. Weight the graph by labeling each edge with its Euclidean length.
3. Compute all pairwise graph distances with Dijkstra's algorithm, square them, and store them in matrix \mathbf{D} .
4. Convert the matrix of distances \mathbf{D} into a Gram matrix \mathbf{S} by double centering.
5. Once the Gram matrix is known, compute its spectral decomposition $\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^T$.
6. A P -dimensional representation of \mathbf{Y} is obtained by computing the product $\hat{\mathbf{X}} = \mathbf{I}_{P \times N}\Lambda^{1/2}\mathbf{U}^T$.

-
- El procedimiento es similar a MDS, la única diferencia es la métrica en el espacio de los datos, que se calcula mediante el grafo
 - Es posible añadir puntos no vistos

SOLUCIONES SAMMON DE LOS PROBLEMAS DE EJEMPLO



OTRAS DISTANCIAS: KPCA

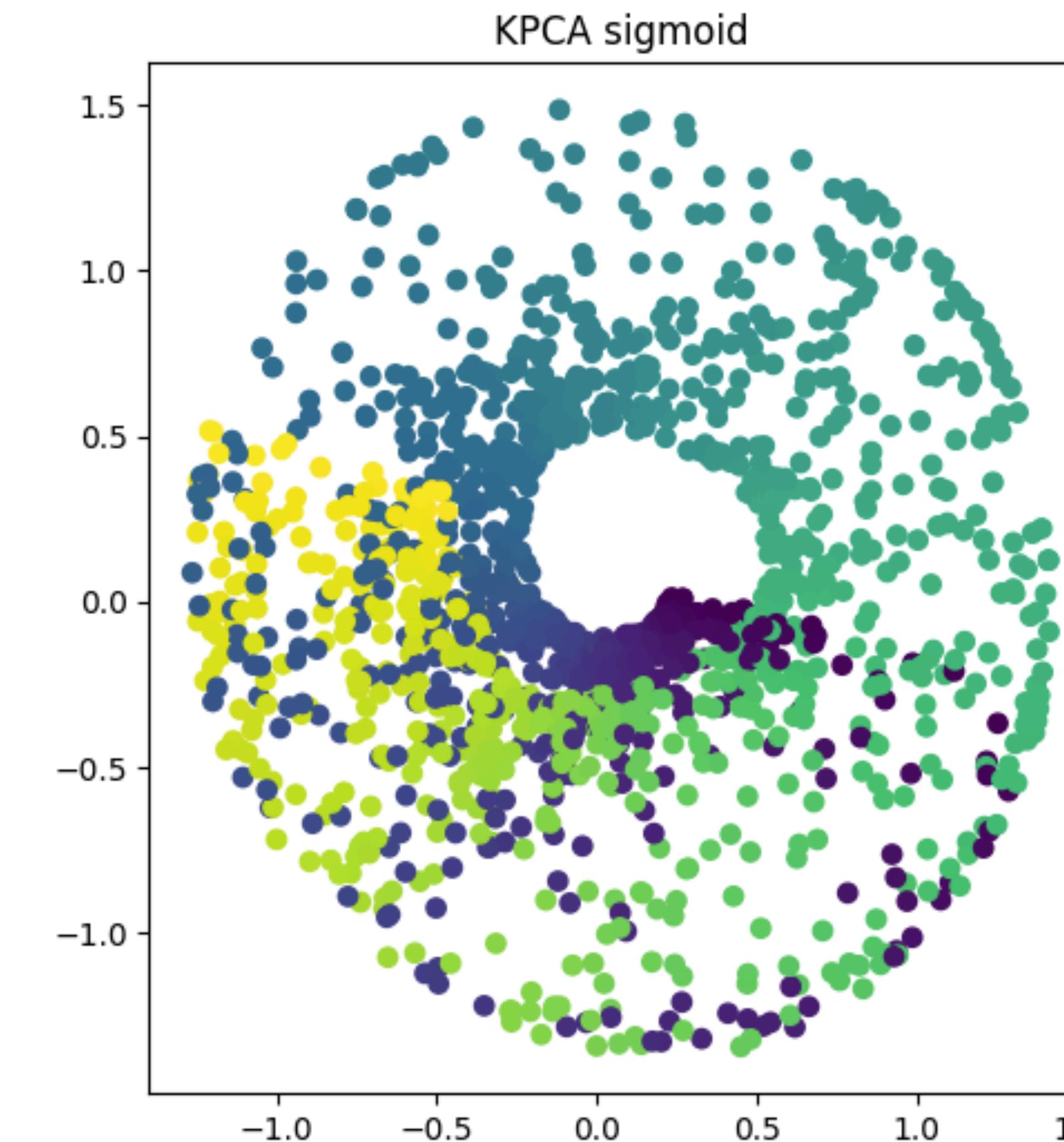
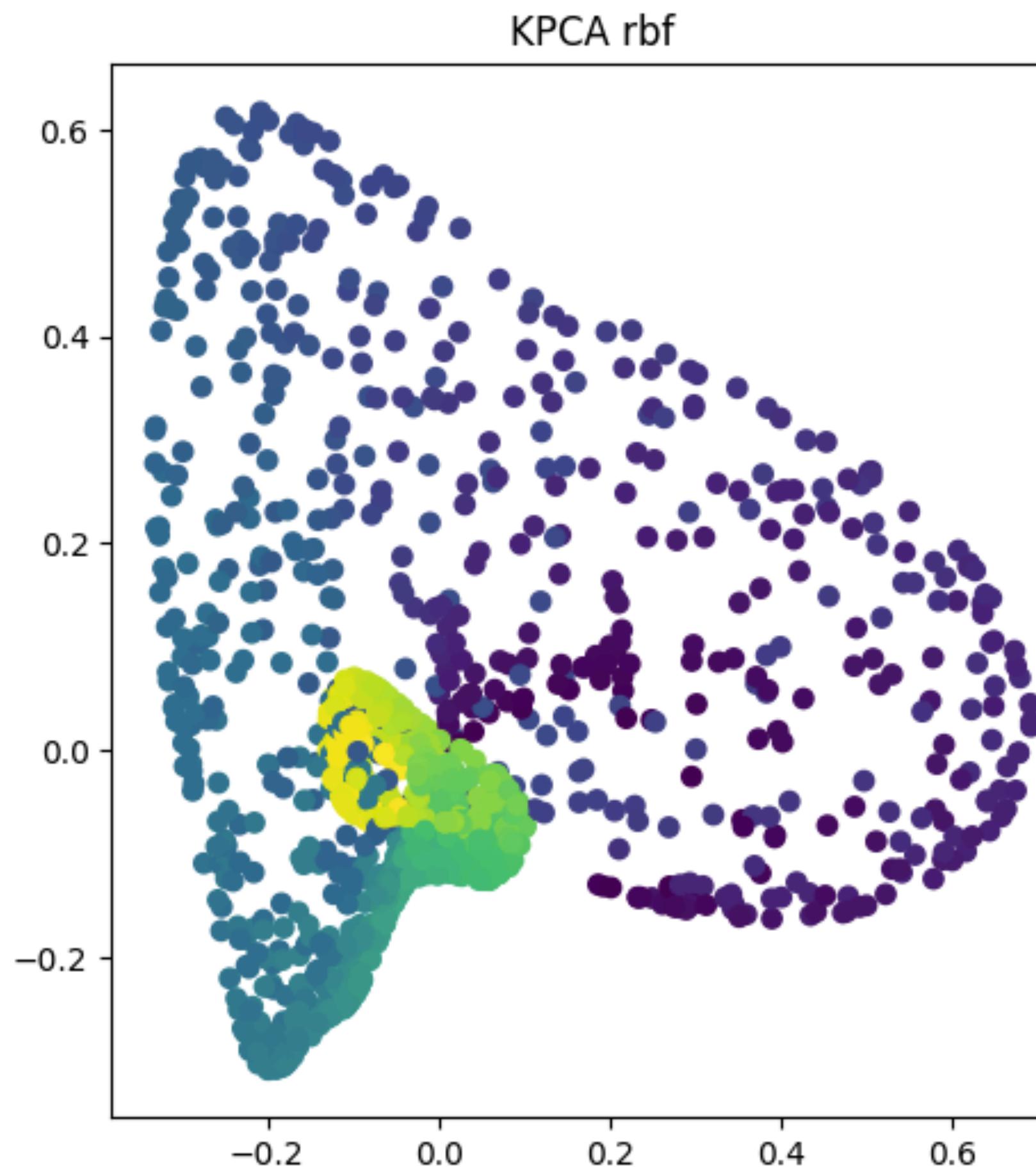
- Kernel PCA: el nombre "Kernel PCA" es una generalización de MDS a variedades no lineales que hace uso del mismo "truco del kernel" usado en las máquinas de vectores de soporte
- Al contrario que PCA, KPCA linealiza la variedad que contiene los puntos en un espacio de mayor dimensión que el espacio inicial. Ese espacio no se usa de forma explícita, sino que trabajamos con la matriz de Gram de productos escalares de los vectores transformados a través de una matriz de aplicaciones de una función kernel a cada pareja de valores del dataset en el conjunto inicial (la función kernel debe cumplir determinadas propiedades -teorema de Mercer- por lo que las elegimos de un catálogo; normalmente se usan kernel polinomiales, de base radial/gaussianos o sigmoidales)

OTRAS DISTANCIAS: KPCA

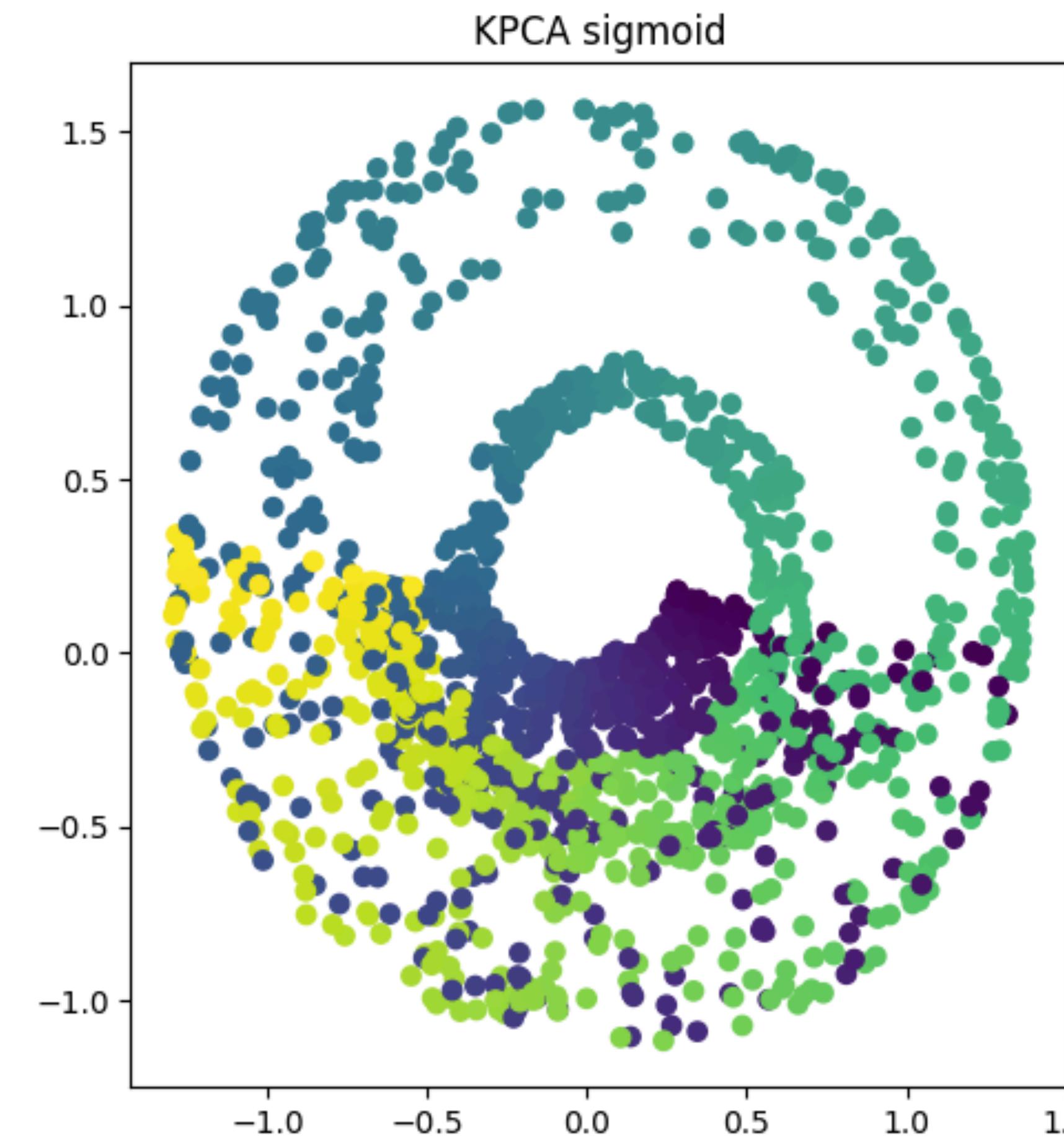
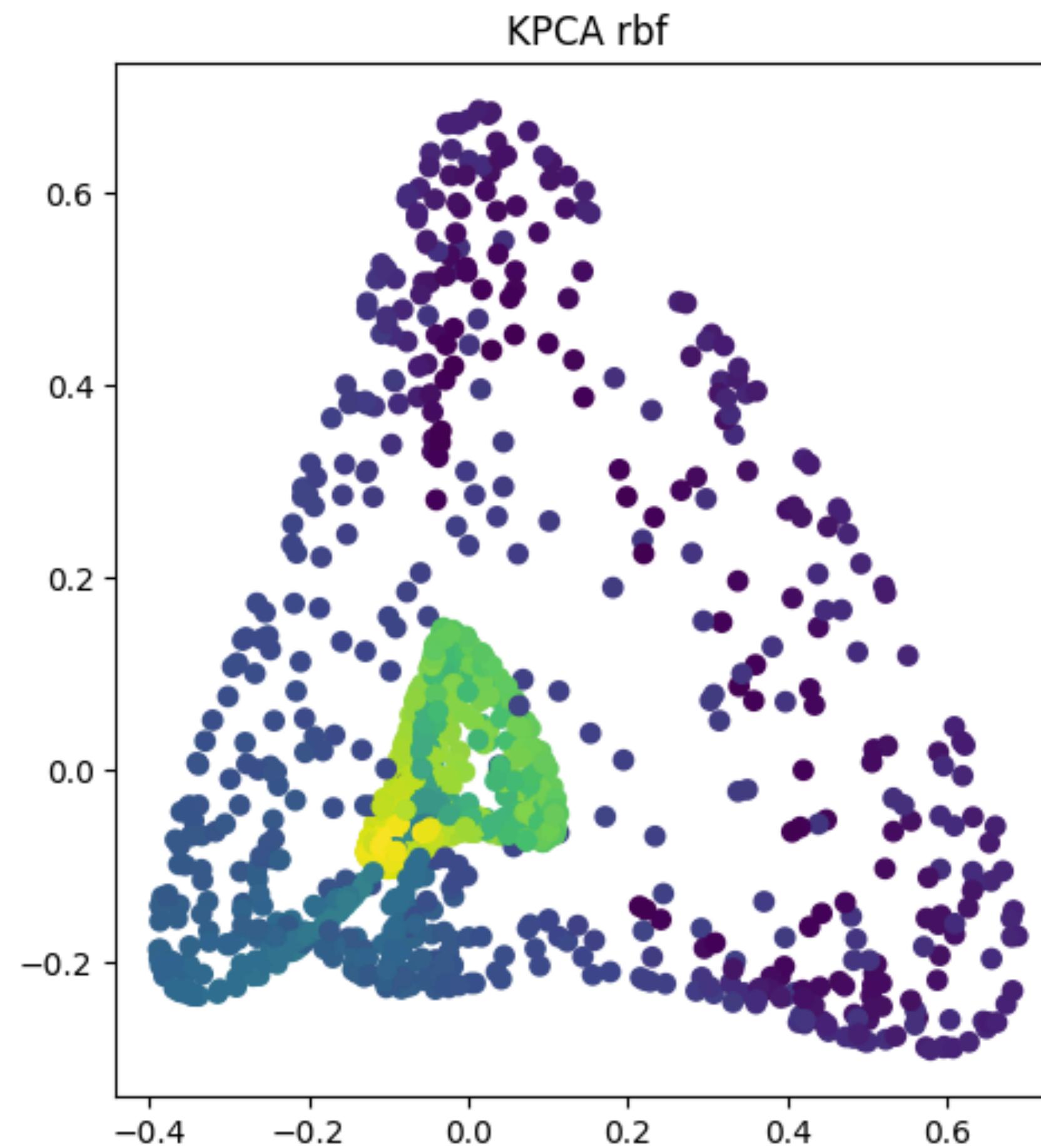
1. Compute either the matrix \mathbf{S} (scalar products) or the matrix \mathbf{D} (squared Euclidean distances), depending on the chosen kernel.
 2. Compute the matrix of kernel values Φ .
 3. Double-center Φ :
 - Compute the mean of the rows, the mean of the columns, and the grand mean.
 - Subtract from each entry the mean of the corresponding row and the mean of the corresponding column, and add back the grand mean.
 4. Decompose the double-centered Φ into eigenvalues and eigenvectors.
 5. A P -dimensional representation of \mathbf{Y} is obtained by computing the product $\hat{\mathbf{X}} = \mathbf{I}_{P \times N} \boldsymbol{\Lambda}^{1/2} \mathbf{U}^T$.
-

- La elección del kernel es arbitraria y trata de conseguir que el cambio de variable implícito linealice la variedad, de forma que la aplicación de PCA en el espacio transformado revele las componentes principales del dataset

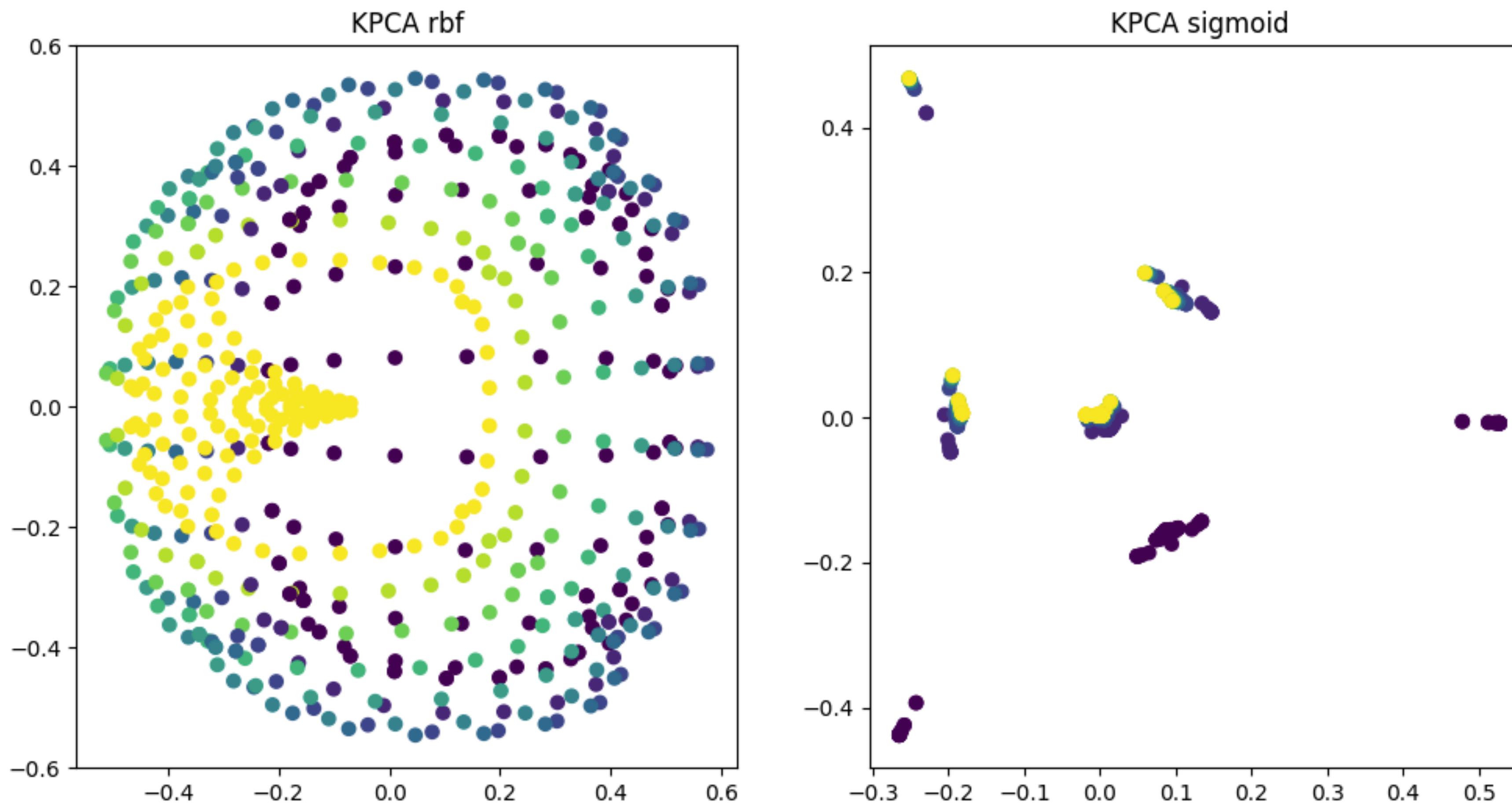
KPCA SWISS ROLL



KPCA SWISS ROLL HOLE



KPCA SWISS OPEN BOX



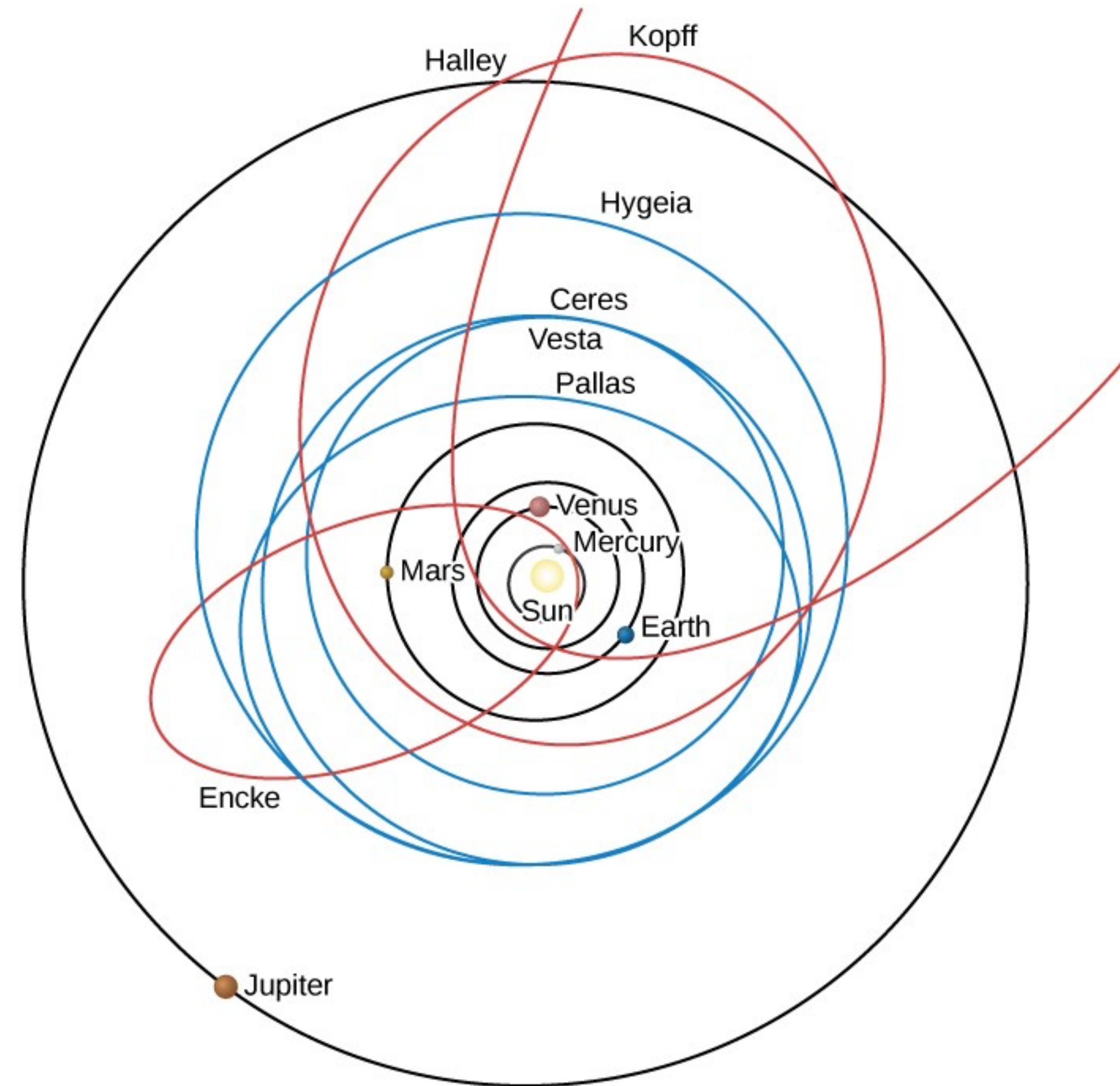
CONSERVACIÓN DE LA TOPOLOGÍA

CONSERVACIÓN DE LA TOPOLOGÍA

- Estado de la técnica
- Retículo predefinido
 - Mapas autoorganizados
- Retículo basado en datos
- Incrustación lineal local (LLE) y spectral embedding

ESTADO DE LA TÉCNICA

- La reducción de la dimensionalidad puede conseguirse mediante transformaciones que conserven las distancias. Este principio es simple intuitivamente, pero tiene algunas limitaciones. En muchos casos, la incrustación de una variedad requiere que algunas regiones se estiren o se compriman para que la visualización sea informativa.
- La conservación de la topología consiste en mantener las proximidades comparadas: una región está cercana a una segunda región y lejos de la tercera.
- Las distancias pueden aportar información innecesaria
- Las relaciones de vecindad solamente se consideran dentro de la variedad (las distancias pueden calcularse en una línea que no esté contenida en el subespacio, como se ha visto en las geodésicas y en las distancias medidas con grafos)



ESTADO DE LA TÉCNICA

- La mayoría de estos métodos operan en un modelo discreto o retículo (*lattice*), que pueden ser un conjunto de puntos repartidos uniformemente en un plano o en un grafo (con pesos en los arcos)
- Los primeros métodos se basan en un retículo predefinido, como SOM (Self Organized Map)
- Los métodos más modernos trabajan con un retículo definido por datos, que puede modificarse cuando el método se ejecuta. Hay métodos que ajustan la importancia de los arcos y otros que pueden introducir nuevos vértices.
- Actualmente este tipo de problemas se resuelve con Graph Neural Networks (https://en.wikipedia.org/wiki/Graph_neural_network)

RETÍCULO PREDEFINIDO: SOM

- Junto con el perceptrón multicapa (MLP), el mapa autoorganizado es quizá el método más conocido en el campo de las redes neuronales artificiales.
- Los SOM realizan simultáneamente la combinación de dos subtareas: la cuantificación vectorial y la representación topográfica (es decir, la reducción de la dimensionalidad).
- Por ejemplo, los SOM también pueden utilizarse en cierta medida para la separación ciega no lineal de fuentes, así como para la reducción no lineal de dimensionalidad. Tienen numerosas aplicaciones en varios campos del análisis de datos, como la predicción de series temporales o la visualización de datos.
- SOM sigue teniendo aplicaciones, pero en este curso estudiaremos en su lugar los métodos en los que el retículo se calcule a partir de los datos.

RETÍCULO BASADO EN DATOS

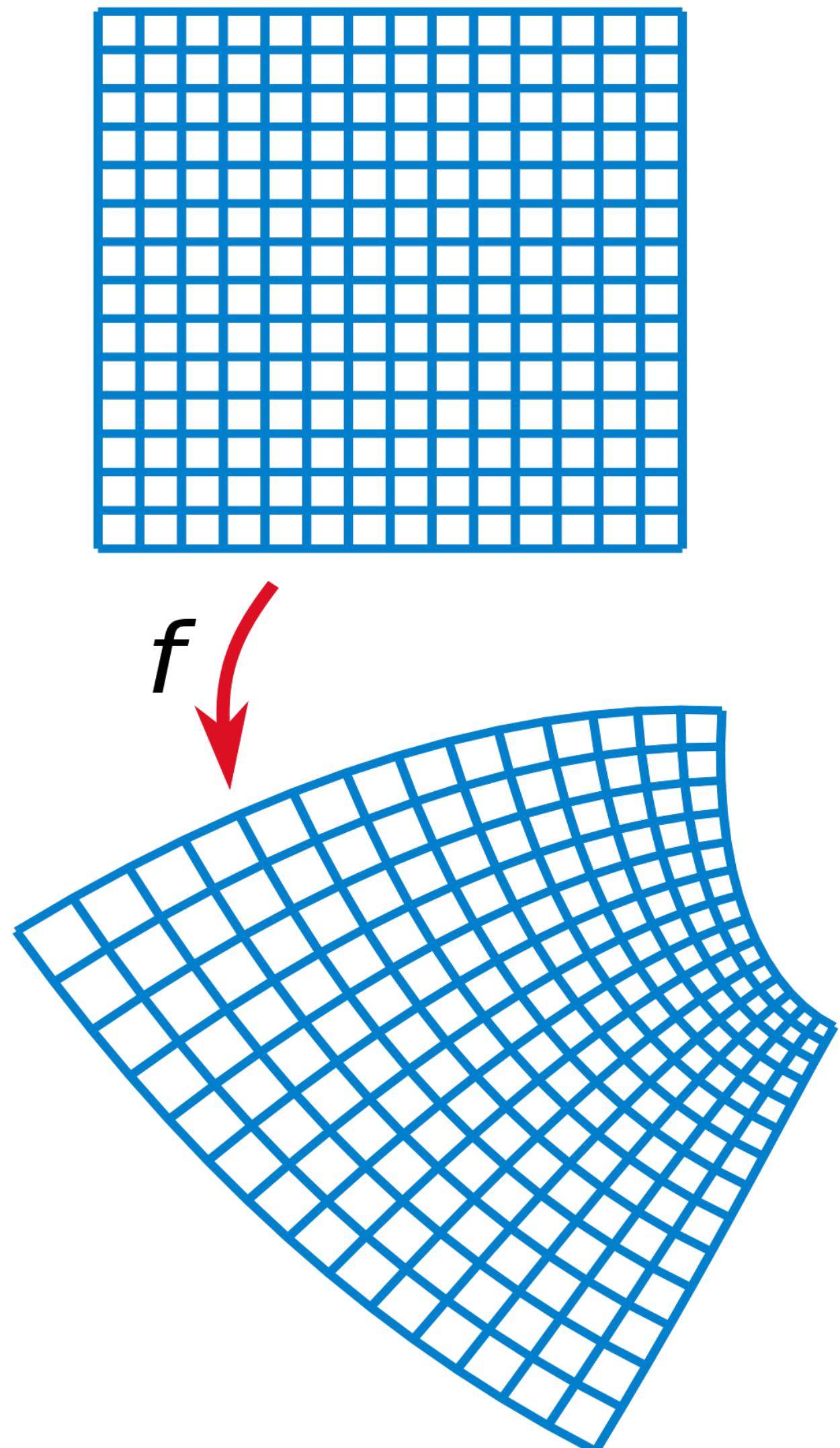
- A diferencia de los métodos que utilizan un retículo predefinido, los métodos basados en datos no hacen ninguna suposición sobre la forma y topología de la incrustación. En su lugar, utilizan la información contenida en los datos para establecer la topología del conjunto de datos y calcular la forma de la incrustación en consecuencia. De este modo, la incrustación no está limitada en modo alguno y puede adaptarse para capturar la forma de la variedad.
- En todos los métodos que se detallan en las secciones siguientes, el entramado de datos se formaliza mediante un grafo cuyos vértices son los puntos de datos y cuyas aristas representan las relaciones de vecindad.

INCRUSTACIÓN LINEAL LOCAL (LLE)

- Los métodos como SOM intentan preservar la topología manteniendo los puntos vecinos cerca unos de otros (los vecinos en la red se mantienen cerca en el espacio de datos). En otras palabras, para estos métodos, la noción cualitativa de topología se traduce concretamente en proximidades relativas: los puntos están cerca o lejos unos de otros.
- LLE propone otro enfoque basado en los mapeados conformes (*conformal mapping*). Un mapa conforme (o mapa biholomorfo) es una transformación que preserva los ángulos locales.
- Hasta cierto punto, la preservación de los ángulos locales y la de las distancias locales están relacionadas y pueden interpretarse como dos formas diferentes de preservar los productos escalares locales.

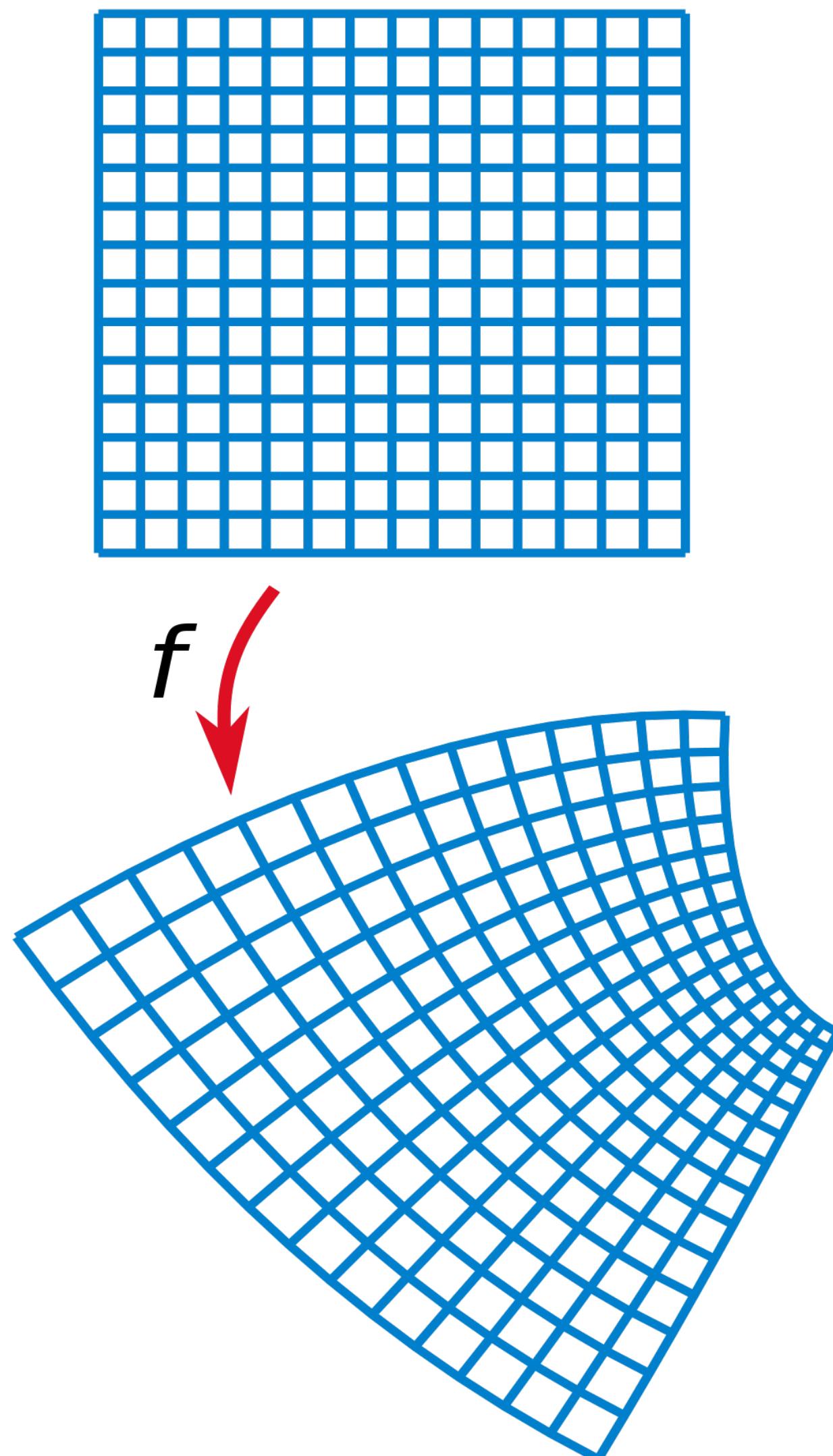
LLE

- Para determinar qué ángulos son relevantes en la representación, LLE selecciona varios vecinos para cada punto del dataset (generalmente, los más cercanos)
- Si el dataset es lo bastante grande y poco ruidoso, se supone que hay un número de vecinos donde la superficie que los contiene es lineal. LLE reemplaza entonces cada punto por una combinación lineal de sus vecinos y almacena los coeficientes de cada punto como filas de una matriz

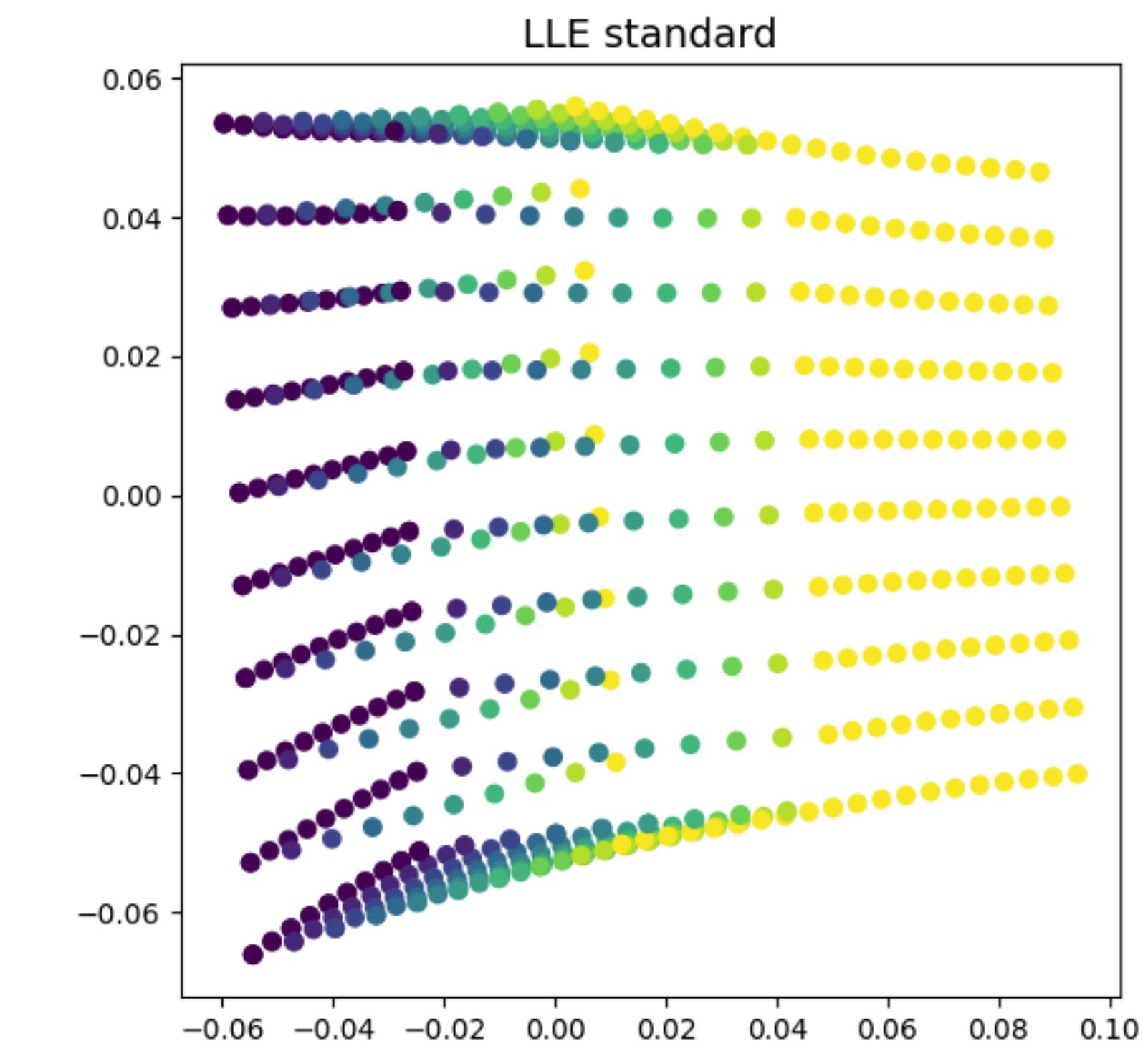
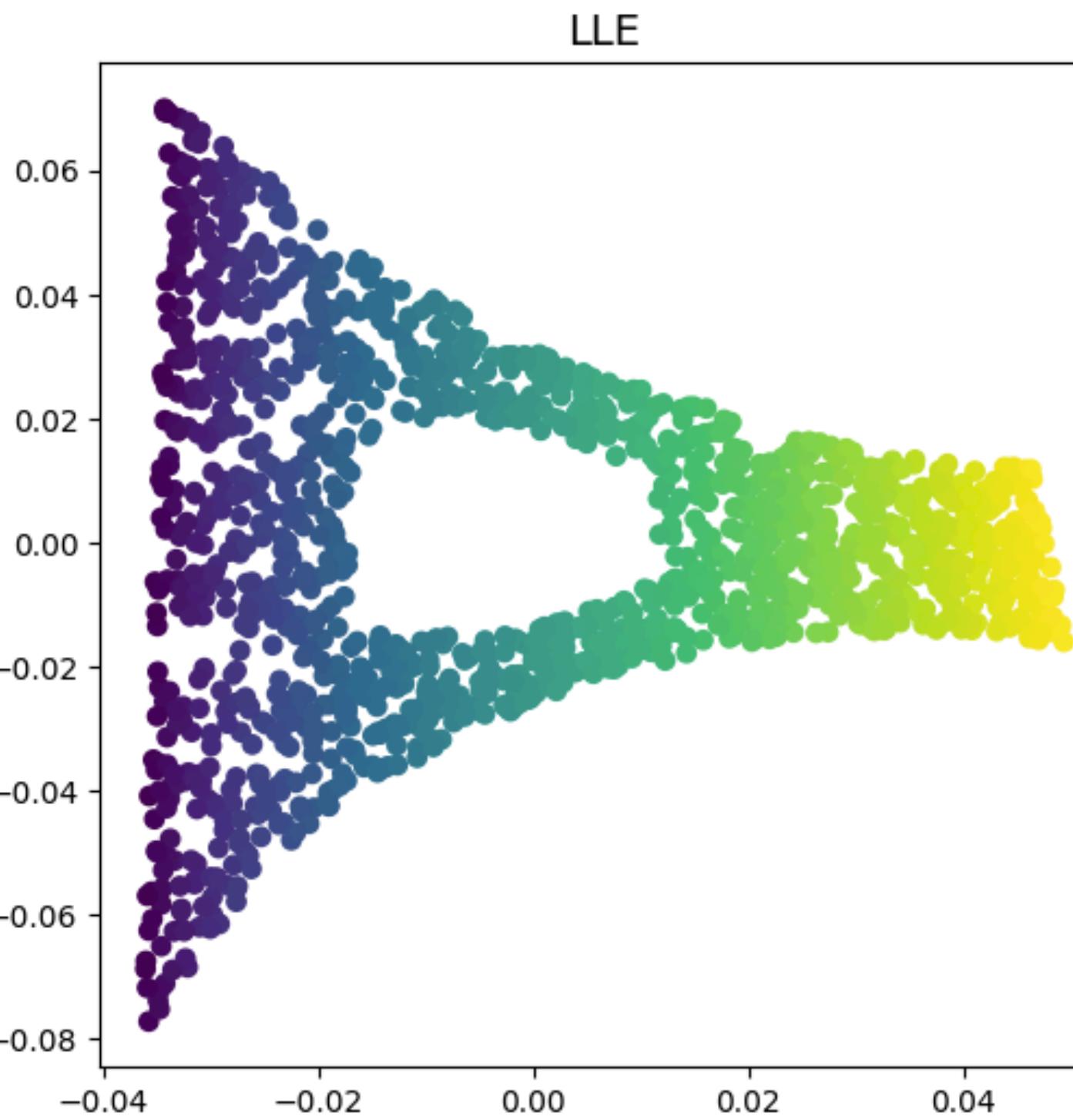
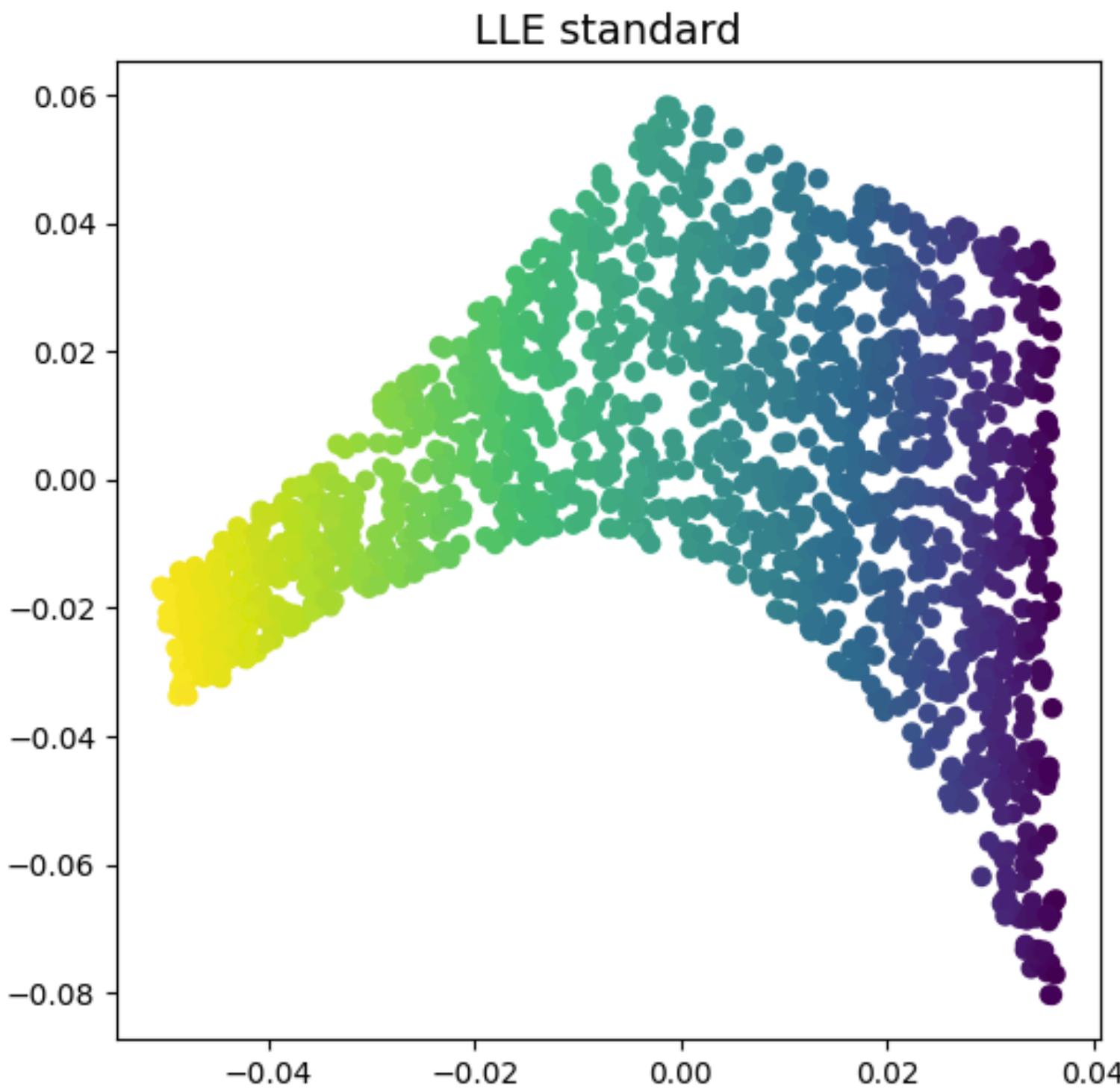


LLE

- Por último, se determina un mapa donde cada punto pueda representarse también (de forma aproximada) como combinación lineal de sus vecinos
- Existen versiones de LLE que permiten interpolar puntos no vistos y situarlos en el mapa
- LLE puede considerarse como una variante de Local PCA (las coordenadas de los puntos con respecto a sus vecinos tienen relación con la covarianza local)
- Varios algoritmos de manifold learning pueden considerarse modificaciones de LLE (LTSA, Hessian Eigenmapping, Laplacian Eigenmapping, etc.)



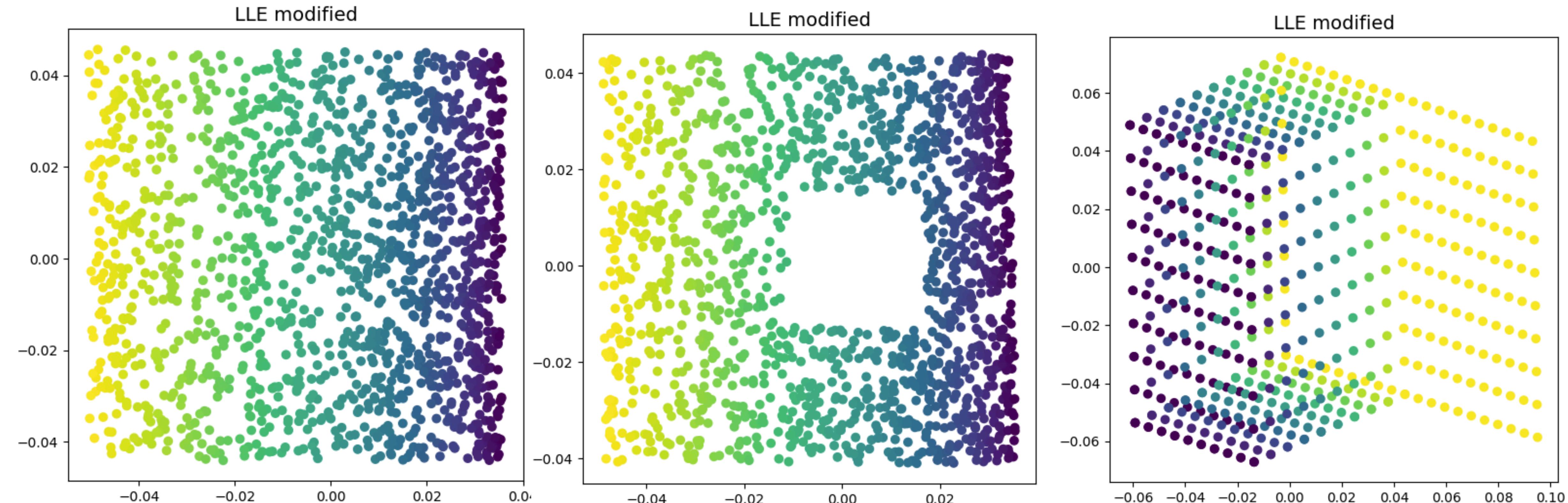
LLE EN LOS PROBLEMAS DE EJEMPLO



LLE MODIFICADO

- Cuando el número de vecinos es mayor que el número de dimensiones, la determinación de los coeficientes es un problema indeterminado, por lo que se añade un parámetro de regularización que penaliza la traza de la matriz local de pesos
- MLLE usa múltiples vectores de pesos en cada vecindario para reducir la distorsión de la geometría que se introduce con el término de regularización
- Sólo se aplica cuando el número de vecinos es mayor que el número de dimensiones del espacio inicial

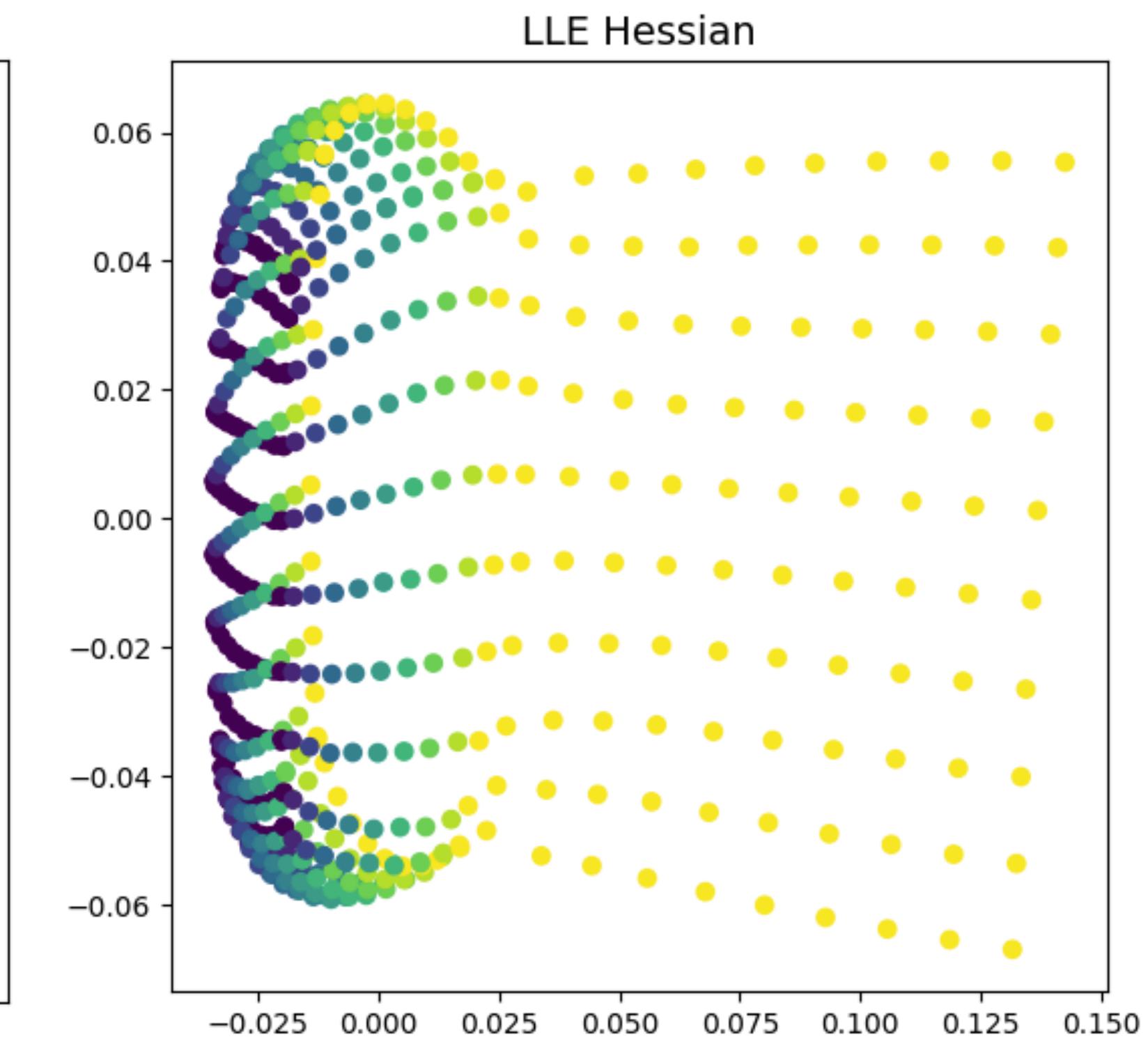
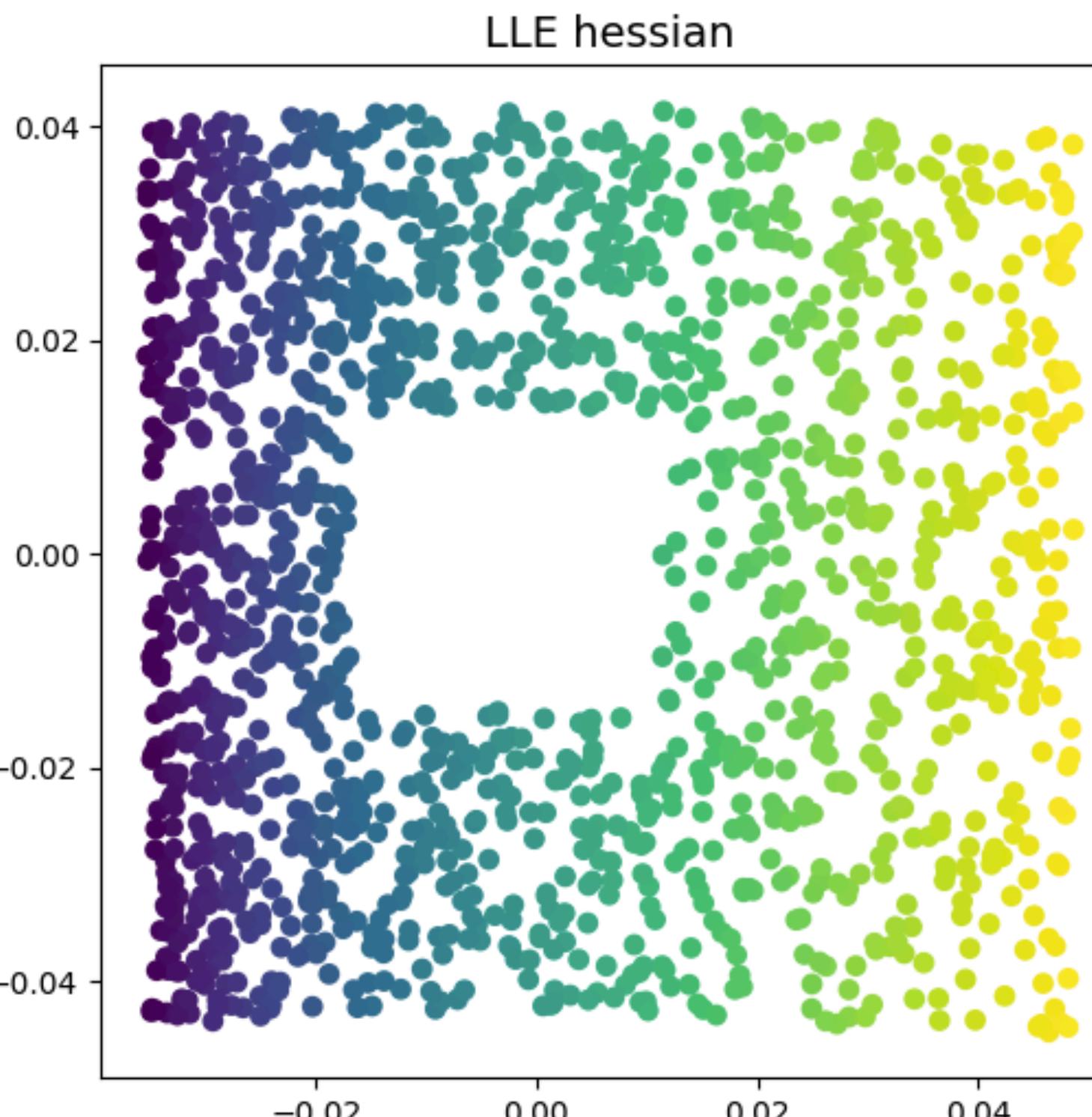
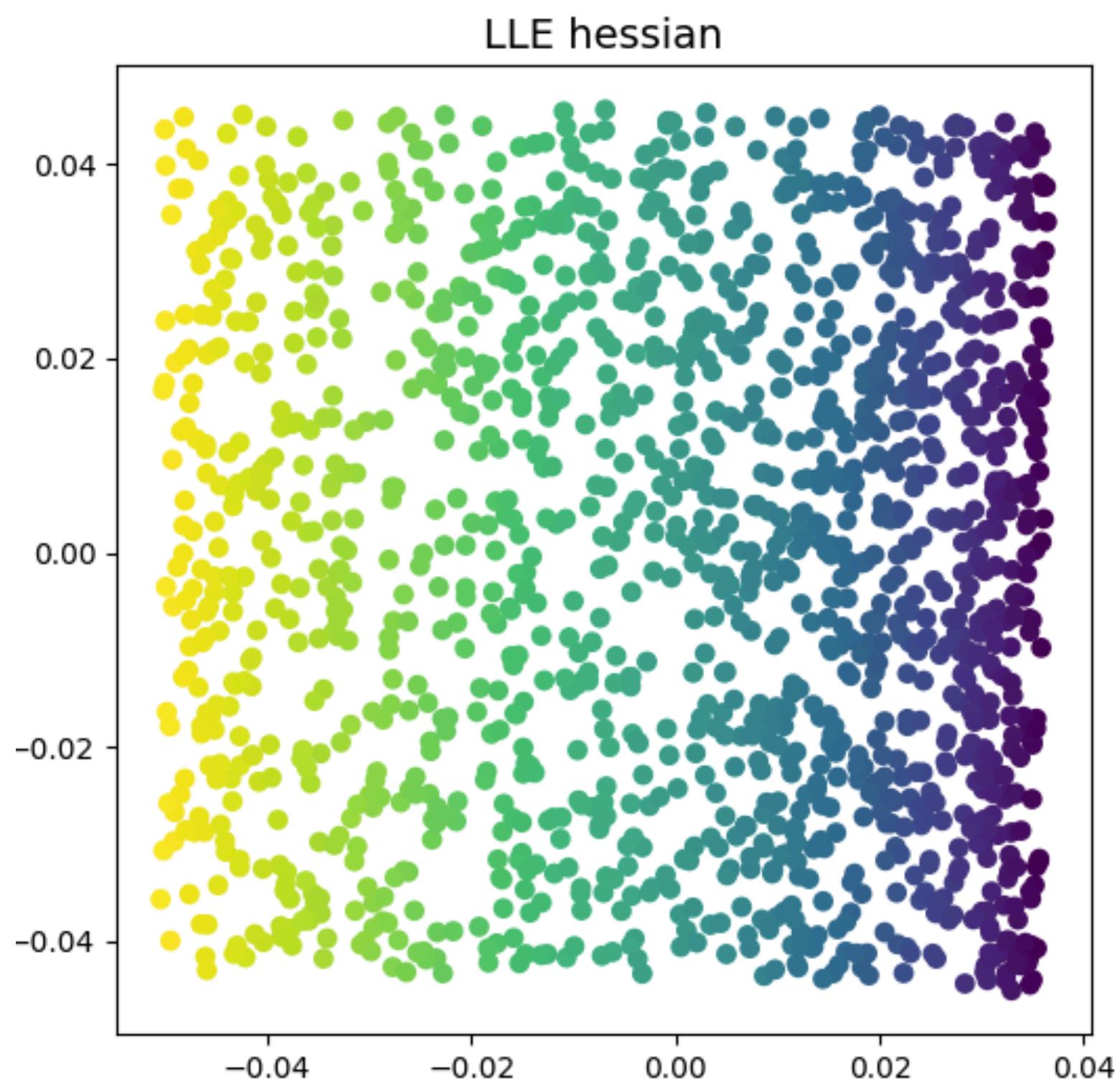
LLE MODIFICADO



HLLE (HESSIAN EIGENMAPPING)

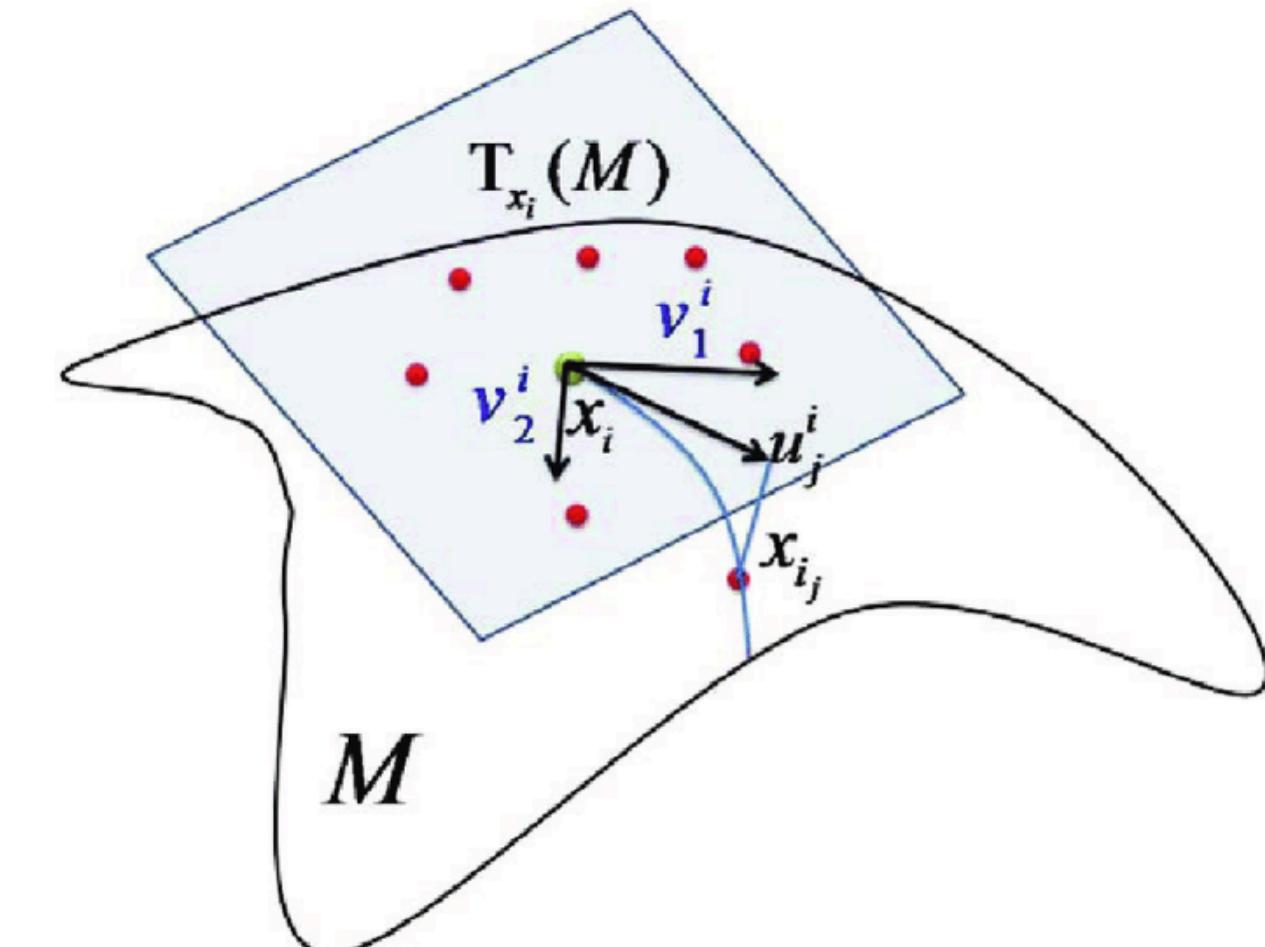
- Una segunda modificación de LLE consiste en reemplazar la aproximación lineal local por una forma cuadrática
- La aproximación lineal aproxima localmente la variedad con el plano tangente, la aproximación cuadrática requiere estimar las derivadas segundas (matriz Hessiana) tanto en el espacio inicial como en el mapa proyectado

HLLE PARA LOS PROBLEMAS DE EJEMPLO

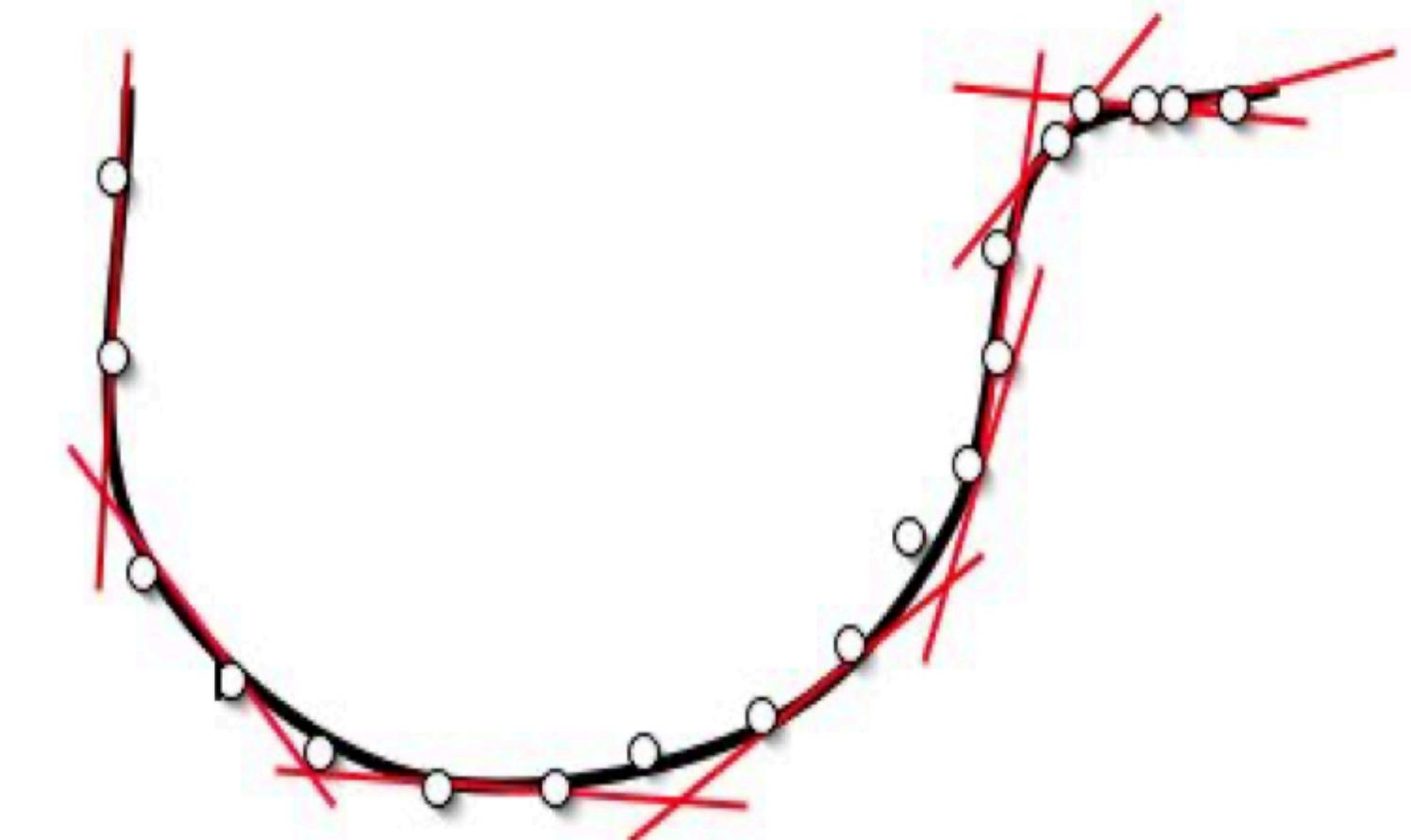


LOCAL TANGENT SPACE ALIGNMENT (LTSA)

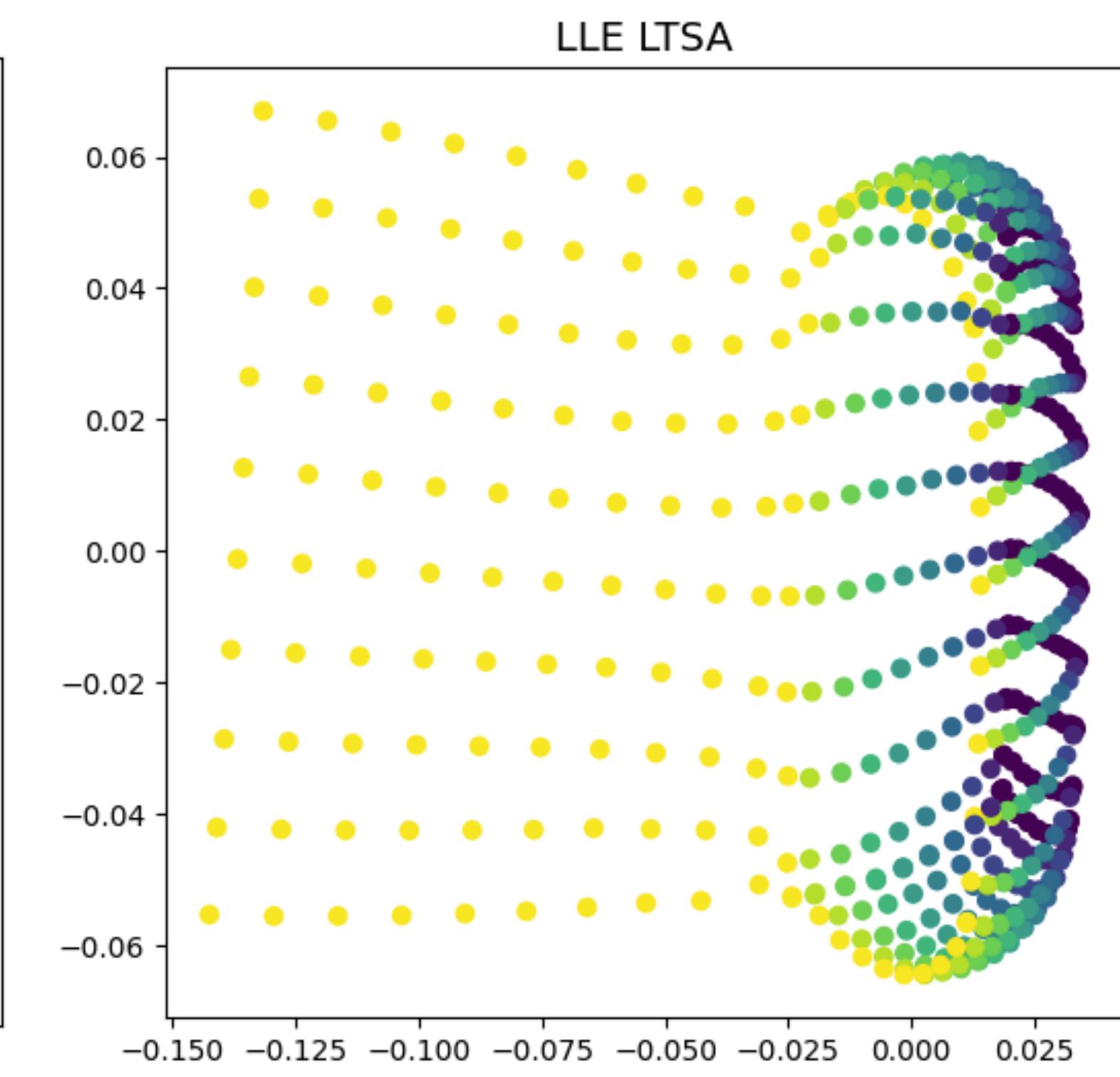
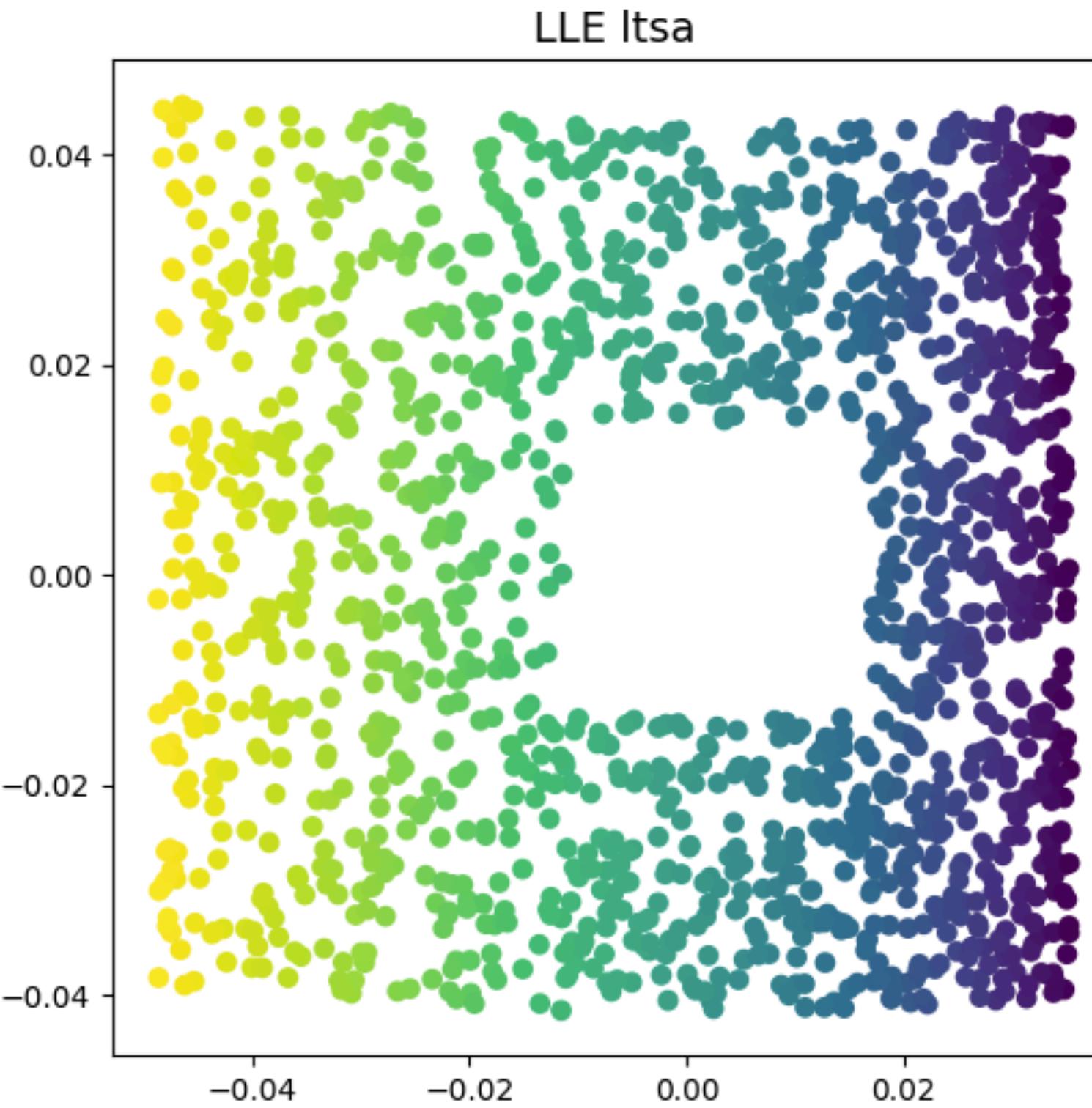
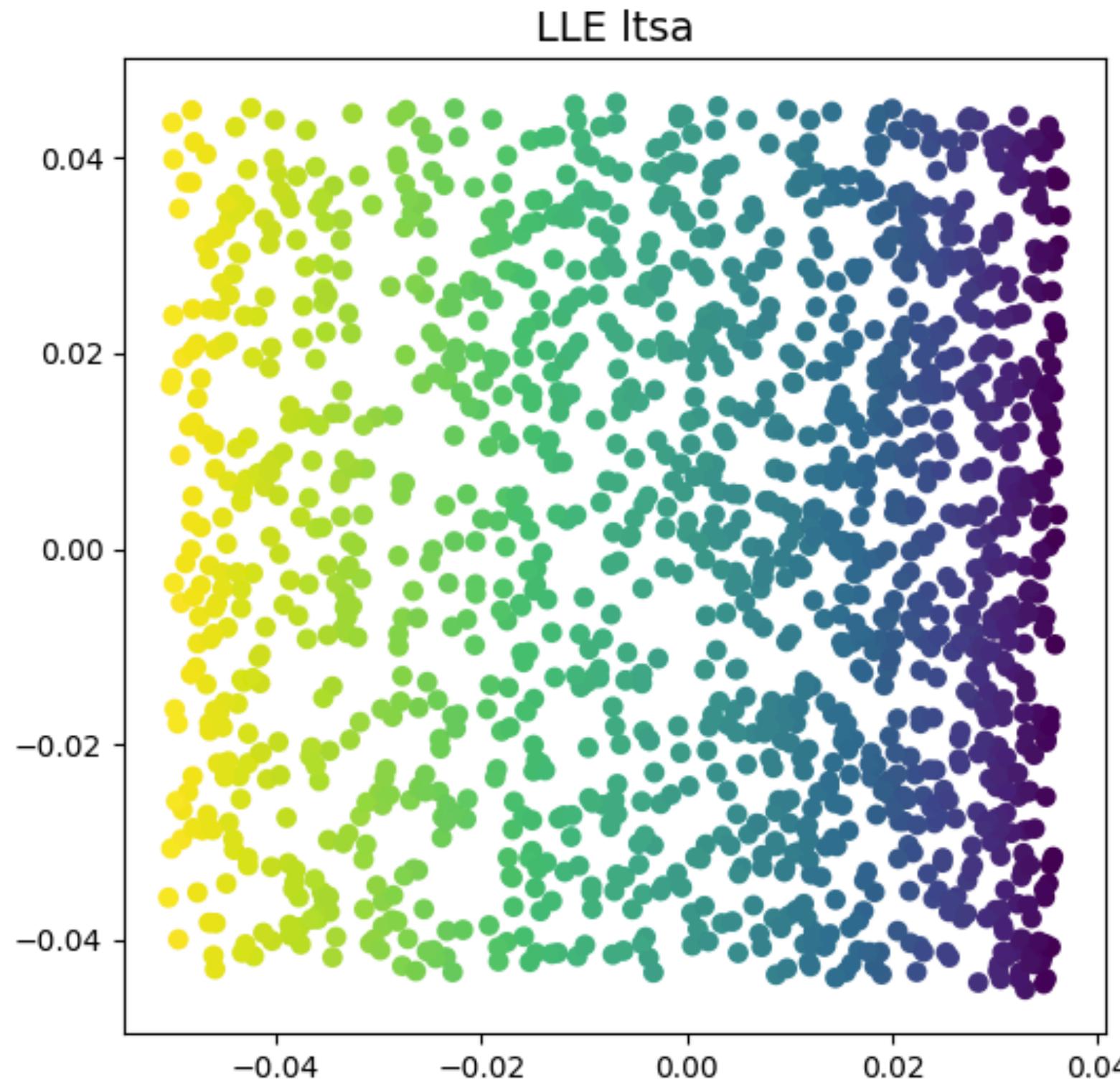
- LTSA es otra variante de LLE basada en una aproximación lineal de la variedad en el entorno de cada punto
- Para los puntos de datos que se encuentran en una variedad d -dimensional, LTSA utiliza una aproximación lineal dentro de cada vecindad para construir un sistema de coordenadas d -dimensional para la vecindad y, a continuación, alinea estos sistemas de coordenadas locales superpuestos para obtener un sistema de coordenadas global.



Local Tangent space approximation



LLE-LTSA APLICADO A LOS PROBLEMAS DE EJEMPLO



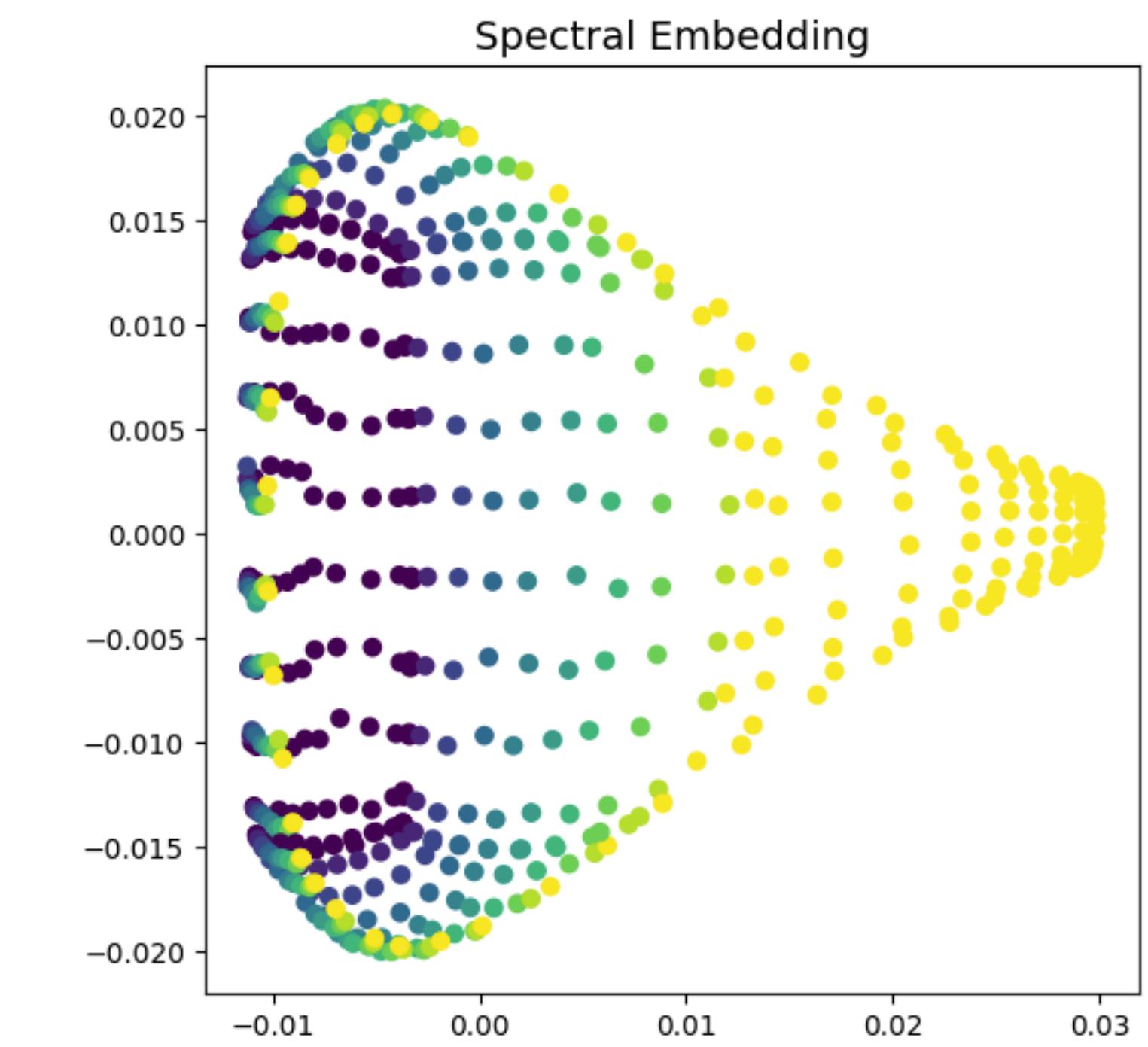
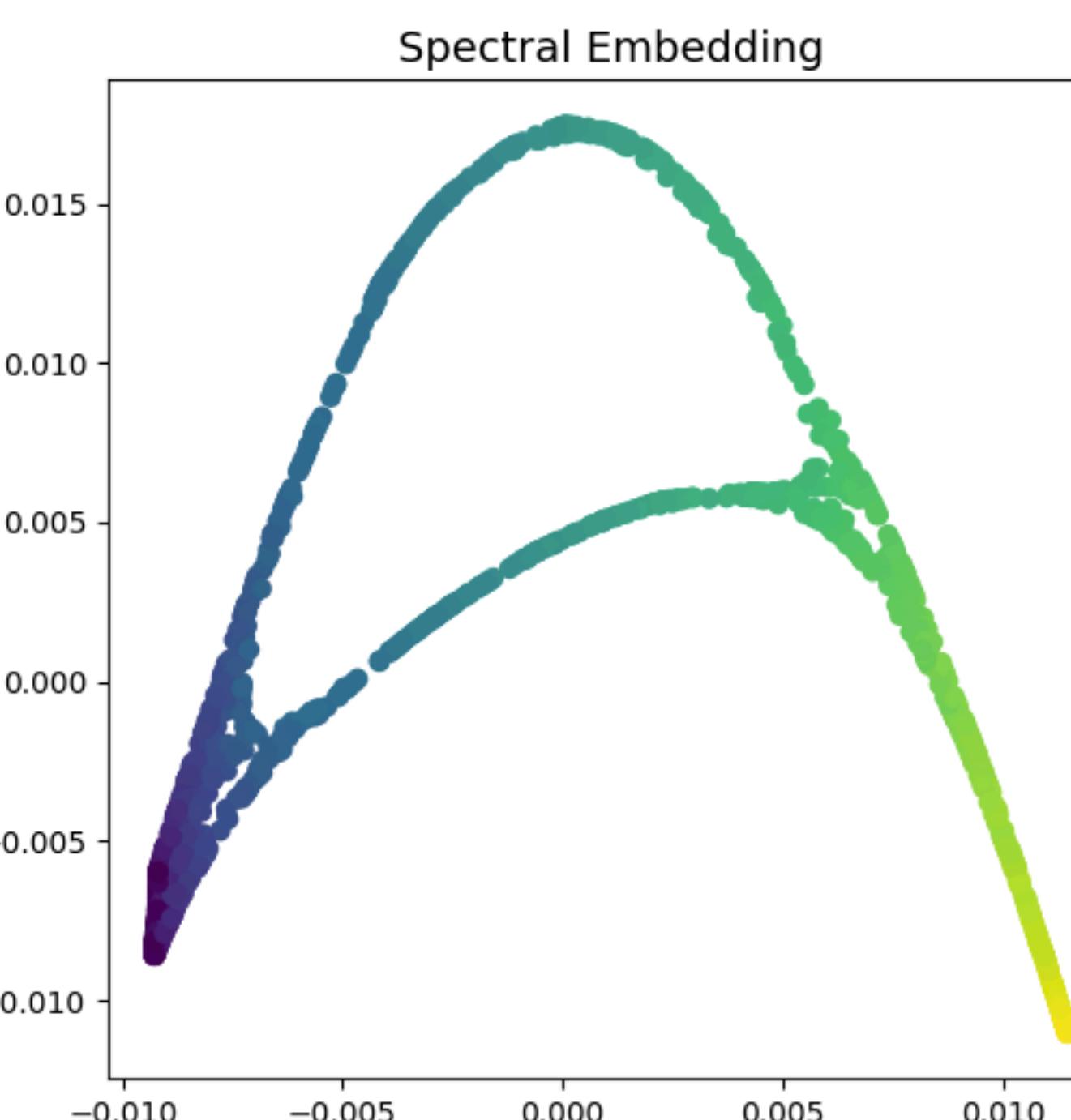
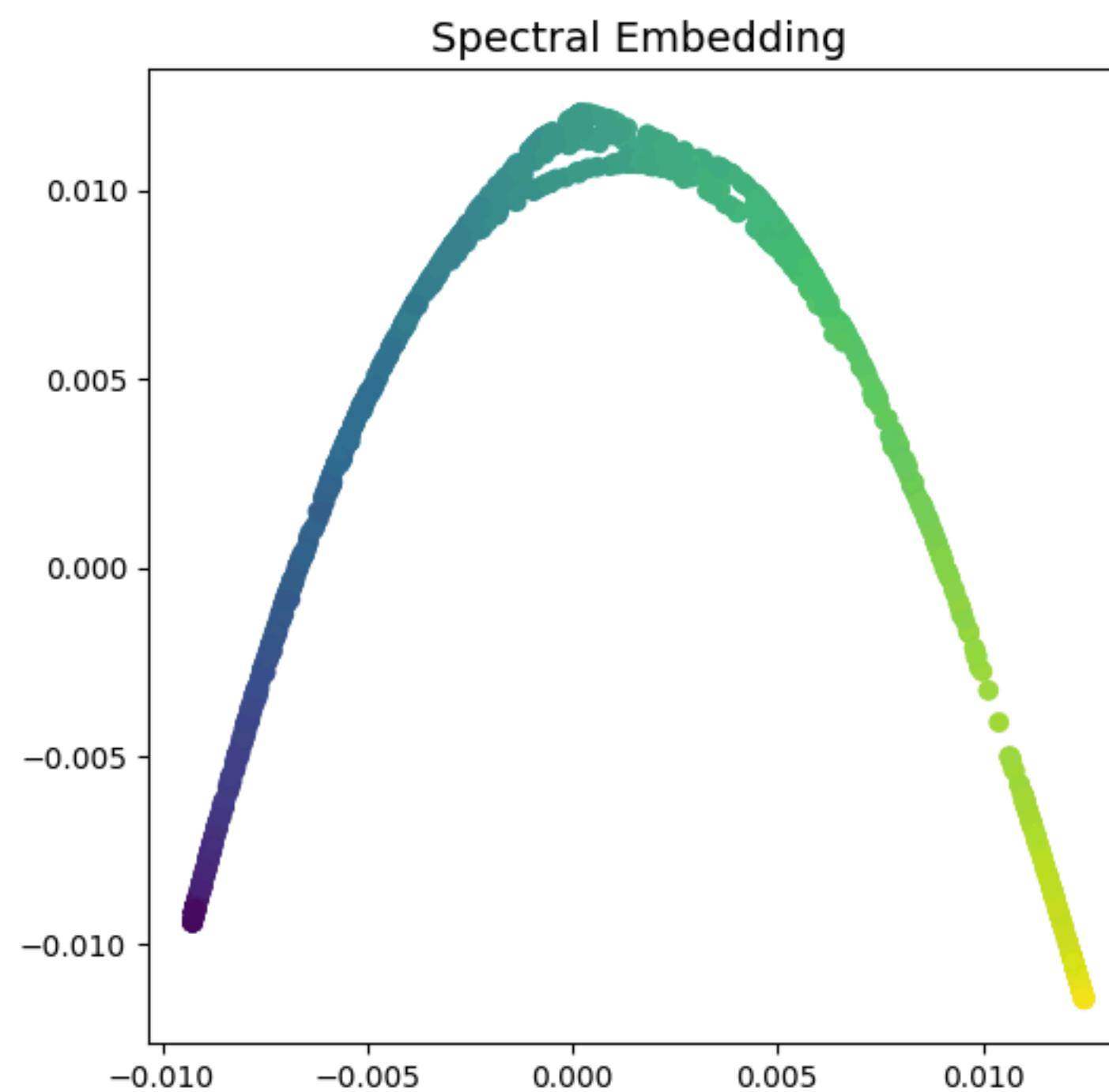
LAPLACIAN EIGENMAPS

- LE está relacionado con LLE, pero es un algoritmo de descomposición espectral como Isomap
- En vez de tomar pequeños entornos lineales de cada punto, usa el operador Laplaciana en el grafo
- Se considera que los datos pueden representarse mediante un grafo que solamente tiene arcos entre puntos vecinos (lo mismo que LLE). Se usa una grafo no dirigido, luego la matriz de adyacencias es simétrica.
- El objetivo de LE es conseguir un mapa en el que las relaciones de vecindad entre los puntos den lugar al mismo grafo

LAPLACIAN EIGENMAPS

- Se construye una aproximación continua al grafo del mapa, que le asigna a cada elemento una similaridad con los demás calculada mediante un kernel gaussiano
- Se define una función de coste donde las distancias entre dos puntos del mapa estén penalizadas por el elemento correspondiente del mapa de similaridades (es decir, no se penaliza la distancia entre puntos lejanos y sí se penalizan que en el mapa haya una gran distancia entre las proyecciones de puntos conectados en el grafo)
- La función de coste puede expresarse en términos de las coordenadas de los puntos del mapa y de la matriz Laplaciana del grafo de conexiones (lo cual le da nombre al método)
- Como KPCA, no suele producir buenos embeddings, es mejor para clustering que para reducción de dimensionalidad.

LAPLACIAN EIGENMAPS PARA LOS PROBLEMAS DE EJEMPLO



CONSERVACIÓN DE LAS PROBABILIDADES

- Los modelos de aprendizaje pueden dividirse en modelos discriminativos y generativos.
- Los modelos discriminativos discriminan las clases de datos para separar mejor las clases, mientras que los modelos generativos aprenden un espacio latente que genera los puntos de datos.
- La inferencia variacional es una técnica que encuentra el límite inferior de la verosimilitud de los datos y lo maximiza. Este límite inferior suele denominarse límite inferior de evidencia (ELBO). El aprendizaje de los parámetros del espacio latente suele realizarse mediante la maximización de expectativas (algoritmo EM)

ESTADO DE LA TÉCNICA

- No repetimos los conceptos de inferencia variacional, ni otros como el algoritmo EM y el análisis factorial, porque ya se estudian en otras asignaturas de este curso
- En el bloque VA2 particularizaremos la inferencia variacional a un tipo de red neuronal de interés en visualización gráfica
- Tampoco explicaremos PCA probabilístico, NCA (Neighbourhood Component Analysis), Bayesian Metric Learning, Random Projection, Sufficient Dimension Reduction, Kernel Dimension Reduction y otras (muchas) técnicas relacionadas
- Por el contrario (aunque UMAP ya se ha mencionado en otras asignaturas de la carrera) introduciremos los métodos de aprendizaje de variedades y reducción de dimensionalidad SNE y UMAP, porque se utilizan ampliamente en visualización.

STOCHASTIC NEIGHBOUR EMBEDDING (SNE)

- En SNE se considera una probabilidad gaussiana centrada en cada punto del dataset
- Esta probabilidad modela la aceptación de que un punto cualquiera del espacio sea vecino del punto correspondiente del dataset
- La distancia entre dos puntos del dataset es

$$\mathbb{R} \ni d_{ij}^2 := \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma_i^2}.$$

- La probabilidad de que el elemento i-ésimo del dataset sea vecino del j-ésimo es

$$\mathbb{R} \ni p_{ij} := \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)},$$

SNE

- La definición de la probabilidad requiere sumar la distancia en todos los elementos del dataset y esto hace que este algoritmo sea impracticable para problemas de gran tamaño
- En la proyección, se sigue un mecanismo similar para decidir la probabilidad de que dos puntos sean vecinos (no se utiliza la varianza)

$$\mathbb{R} \ni q_{ij} := \frac{\exp(-z_{ij}^2)}{\sum_{k \neq i} \exp(-z_{ik}^2)}, \quad \mathbb{R} \ni z_{ij}^2 := \|\mathbf{y}_i - \mathbf{y}_j\|_2^2.$$

SNE

- El objetivo es que las distribuciones de probabilidad en los espacios inicial y en el mapa sean similares, para lo cual se minimiza una divergencia entre ambas evaluada en los puntos de la muestra.
- Si se emplea la divergencia de Kullback-Leibler, la expresión es

$$\mathbb{R} \ni c_1 := \sum_{i=1}^n \text{KL}(P_i || Q_i) = \sum_{i=1}^n \sum_{j=1, j \neq i}^n p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right),$$

T-SNE

- En la versión t-SNE se emplea una distribución t de Student o una distribución de Cauchy en vez de una distribución normal en el mapa
- Esto es porque las distribuciones t / Cauchy tienen las colas más pesadas que la distribución gaussiana y los datos quedan mejor repartidos visualmente
- Las probabilidades de que dos elementos del mapa sean vecinos se derivan de la distribución elegida

$$q_{ij} = \frac{(1 + z_{ij}^2)^{-1}}{\sum_{k \neq l} (1 + z_{kl}^2)^{-1}}, \quad q_{ij} = \frac{(1 + z_{ij}^2/\delta)^{-(\delta+1)/2}}{\sum_{k \neq l} (1 + z_{kl}^2/\delta)^{-(\delta+1)/2}}.$$

UMAP

- UMAP es, junto con t-SNE uno de los métodos "estado del arte" en visualización de datos
- UMAP comienza construyendo un grafo con los k vecinos de cada elemento, como se ha visto en los métodos que conservan la distancia. La similaridad entre cada pareja de puntos conectados de forma similar a SNE, con un kernel gaussiano.
- En UMAP se calcula el parámetro de la varianza para cada punto de manera que se cumpla que

$$\sum_{j=1}^k \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_{i,j}\|_2 - \rho_i}{\sigma_i}\right) = \log_2(k).$$

- donde $\rho(i)$ es la distancia del punto i -ésimo a su punto más cercano en el dataset. De esta manera se evita calcular el denominador de SNE y el método es mucho más rápido en problemas de gran tamaño.

- Las probabilidades en el mapa proyectado se calculan mediante la función

$$\mathbb{R} \ni q_{ij} := (1 + a \|y_i - y_j\|_2^{2b})^{-1},$$

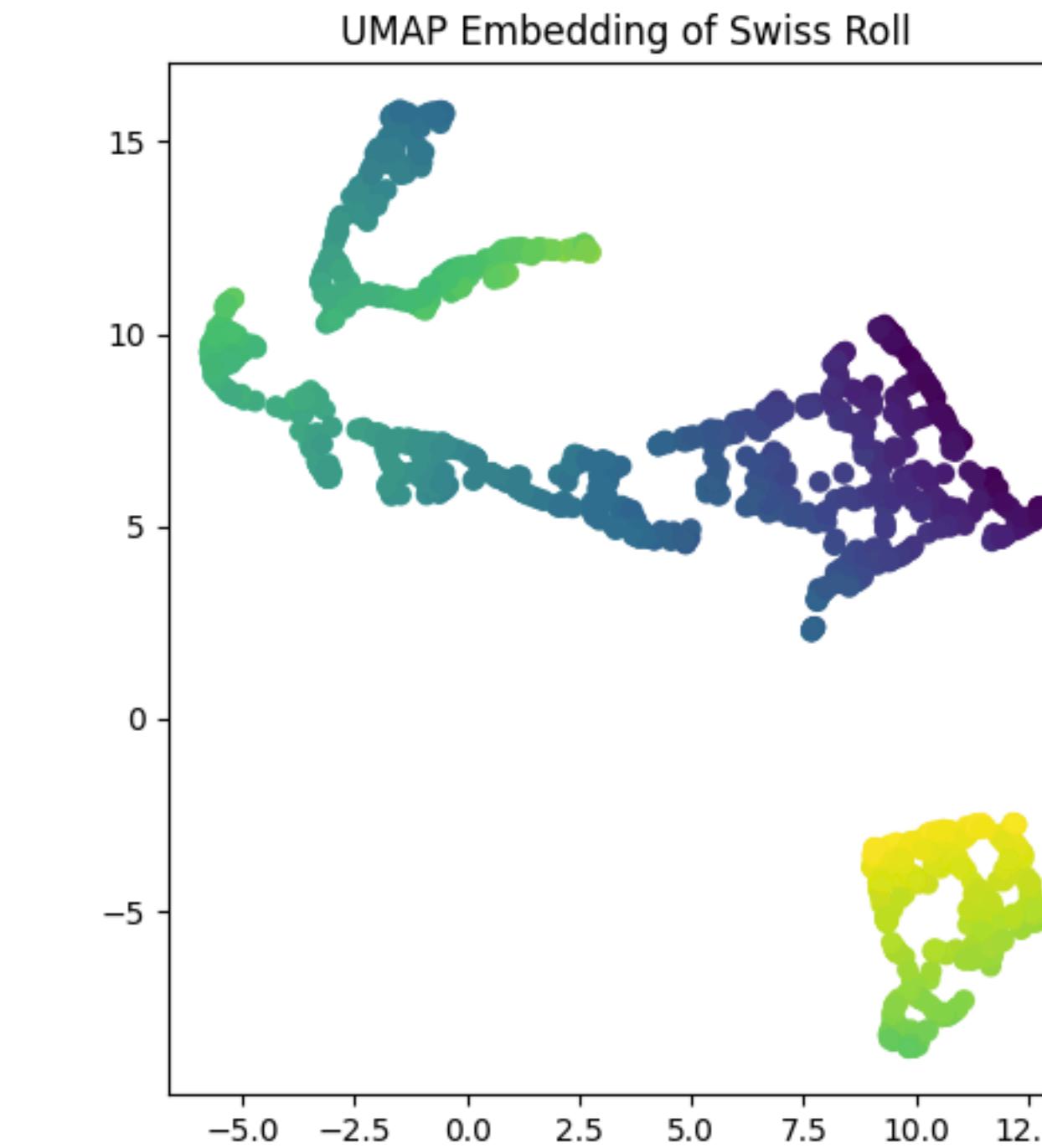
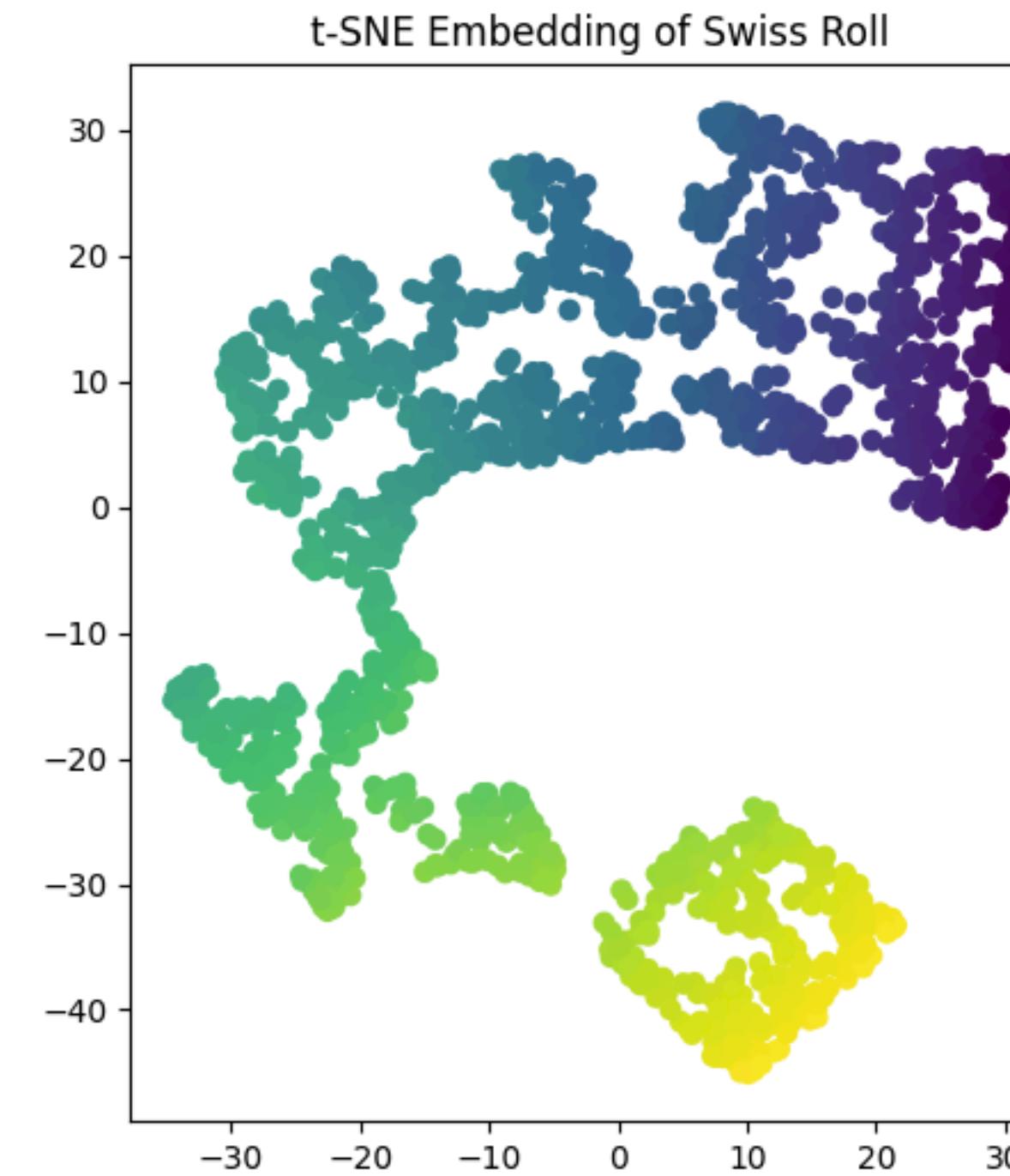
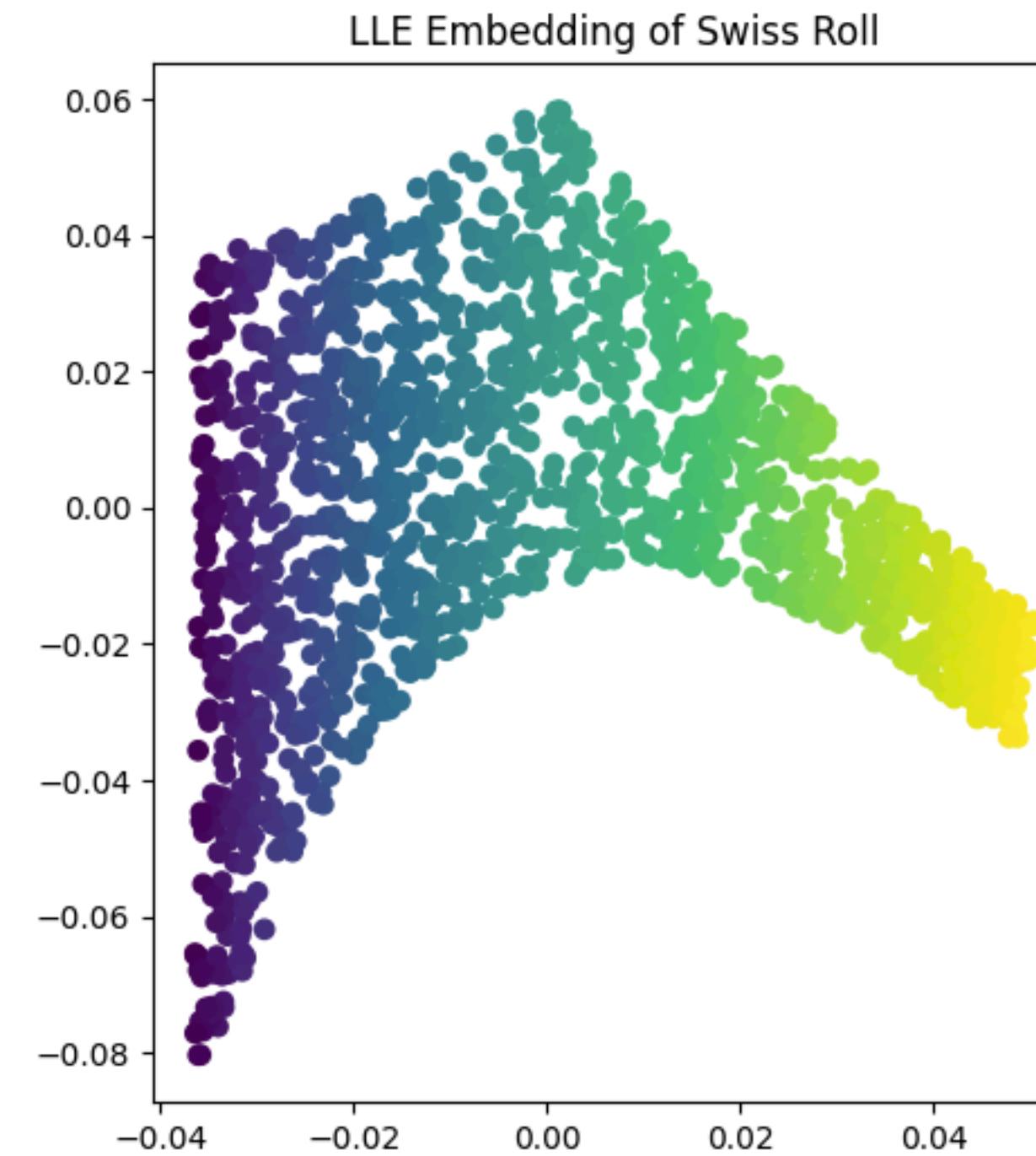
- "a" y "b" son hiperparámetros. La función de coste es la fuzzy cross-entropy,

$$c_1 := \sum_{i=1}^n \sum_{j=1, j \neq i}^n \left(p_{ij} \ln\left(\frac{p_{ij}}{q_{ij}}\right) + (1 - p_{ij}) \ln\left(\frac{1 - p_{ij}}{1 - q_{ij}}\right) \right)$$

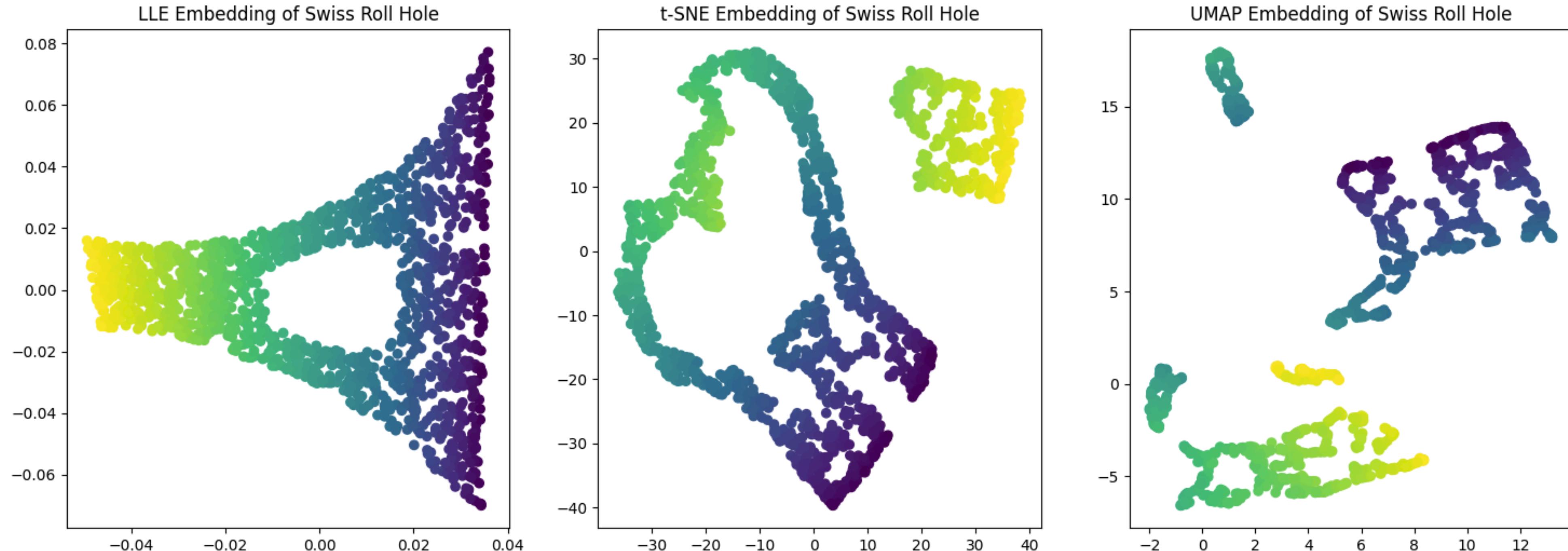
UMAP

- UMAP parte del Laplacian Eigenmap visto anteriormente para inicializar los puntos del mapa, y optimiza iterativamente la entropía cruzada entre el mapa y el espacio mediante descenso de gradiente estocástico
- Los pasos de este algoritmo tienen una justificación teórica consistente (no son decisiones heurísticas) pero, aunque su concepción es muy diferente de SNE, su implementación numérica sigue siendo un algoritmo de grafos de vecindades, como este último, con algunas diferencias que le hacen notablemente más eficiente en problemas de tamaño grande.

COMPARACIÓN EN PROBLEMAS DE EJEMPLO



COMPARACIÓN EN PROBLEMAS DE EJEMPLO



COMPARACIÓN EN PROBLEMAS DE EJEMPLO

