Universiteit Leiden

A Unified Toolkit for Evaluating Nonlinear Dimensionality Reduction Techniques

Kasra Amirani (s4084845)

Supervisors:
S.M.H. Huisman Ph.D. & Dr. E.P.L. van Nieuwenburg

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

June 30, 2026

Abstract

High-dimensional data appear in many fields, and visualizing them is often challenging. Nonlinear dimensionality reduction methods such as t-SNE and UMAP are commonly used for this, but it remains difficult to judge how good the resulting embeddings are. Each method optimizes its own internal objective, so these values are method-specific and do not give a fair way to compare embeddings across different methods. In this thesis we build on an existing diagnostics toolkit mainly for t-SNE and UMAP and turn it into a more accessible package for interested practitioners, while also extending it with diagnostics tools. The framework includes global plots, such as distance-fit plots using both Euclidean and geodesic distances, similarity plots, and rank-based measures, as well as local tools that let the user inspect a single selected point or a specific pair of points. We demonstrate the framework on four datasets with different structures: Iris, Swiss Roll, Digits, and a modified Blobs. The experiments show how these diagnostics can reveal structure that internal loss values hide, and help the user understand whether an embedding reflects the real data or introduce distortions of its own.

Contents

1 Introduction

High dimensional data appear in many fields, and visualizing or interpreting them is often challenging. Dimensionality reduction techniques help by representing such data in a lower-dimensional space while preserving meaningful structure. Classical methods such as Principal Component Analysis (PCA) are useful when the data are mostly linear, while nonlinear dimensionality reduction techniques such as t-SNE, UMAP and ISOMAP are often more suitable when the data lie on a nonlinear manifold [Hot33, AW10, TSL00, MH08, MHM18].

Although these methods are widely used, it remains difficult to assess the quality of the resulting embeddings. Most non-linear dimensionality reduction methods optimize their own internal objective, such as stress minimization, reconstruction error, or graph-preservation loss. However, these quantities are method specific and mainly reflect what the algorithm is designed to optimize. Because of this, they do not provide a fair basis for comparing embeddings across different methods [Wan25].

Several embedding quality measures have been proposed in the literature [LV08, LV09, VK06]. These usually focus on specific properties of an embedding, such as neighbourhood preservation, continuity, or rank-based agreement between the original and embedding spaces. While informative, they do not always provide a unified and broadly applicable framework for evaluating embeddings across methods and datasets.

This thesis builds on an existing diagnostics toolkit for t-SNE and UMAP embeddings [Hai25]. The aim is to transform this toolkit into a more usable package for the interested practitioners and to extend it with diagnostics that better assess global structure preservation. In this way, the project contributes toward a more informative and interpretable framework for evaluating nonlinear dimensionality reduction techniques.

The remainder of this thesis is structured as follows Chapter 2 reviews dimensionality reduction methods and existing embedding quality measures. Chapter describes the implementation of the proposed package and diagnostics. Chapter 4 presents the experimental setup and the results and Chapter 5 explains the conclusion and limitations of the current work.

2 Background and Related Work

This chapter introduces the theoretical background relevant to this thesis. It begins with dimensionality reduction in general and the important distinction between linear and non-linear approaches. It then reviews representative non-linear dimensionality reduction techniques, with particular focus on methods that preserve local neighbourhood structure or approximate manifold geometry. Finally, it discusses existing approaches for assessing embedding quality and identifies the gap in the field.

2.1 Dimensionality Reduction

Dimensionality reduction is commonly used when data are represented in a high-dimensional feature space and are therefore difficult to interpret, visualize, or process directly. Thus, the main goal is

to construct a lower-dimensional representation that preserves as much meaningful structure as possible. In practice, dimensionality reduction is widely used for visualization, exploratory analysis, noise reduction, and feature extraction. As discussed by Cunningham and Ghaharmani, dimension reduction remains one of the most important tools for understanding complex datasets [CG15].

In general, DR techniques can be divided into linear and non-linear approaches. Linear methods assume that the important structure of the data can be captured through a linear transformation of the original variables [CG15]. Non-linear methods are intended for settings in which the data may lie on or near a curved spaced. In such cases, linear projections may fail to preserve the intrinsic geometry of the data, which motivated the development of manifold learning methods such as ISOMAP, t-SNE, and UMAP [MH08, TSL00, MHM18].

In a mathematical sense, to simplify what DR does we can let the points in the dataset (high dimension) to be $X = \{x_1, x_2, \ldots, x_n\}$, where each point $x_i \in \mathbb{R}^D$. The goal of DR is to find a mapping $f : \mathbb{R}^D \mapsto \mathbb{R}^d$, with $d < D$,such that the low-dimensional representations $Y = \{y_1, y_2, \ldots, y_n\}$, with $y_i \in \mathbb{R}^d$, retain important information from the original data.

2.2 Linear Dimensionality Reduction

One of the most established DR methods is Principal Component Analysis (PCA), originally formalized by Hotelling [Hot33]. PCA transforms the data into a new coordinate system defined by orthogonal directions, called principal components, such trhat the first component captures the greatest variance in the data, the second captures the greatest remaining variance orthogonal to the first, and so on. Abdi and Williams describe PCA as one the standard approaches for summarizing correlated variabels by a smaller set of orthogonal components [AW10]. For a centered data matrix $X \in \mathbb{R}^{n \times D}$, the PCA is based on the eigendecomposition of the covariance matrix

$$\Sigma = \frac{1}{n} X^\intercal X$$

. And the lower dimensional representation is obtained as

$$Y = XW_d$$

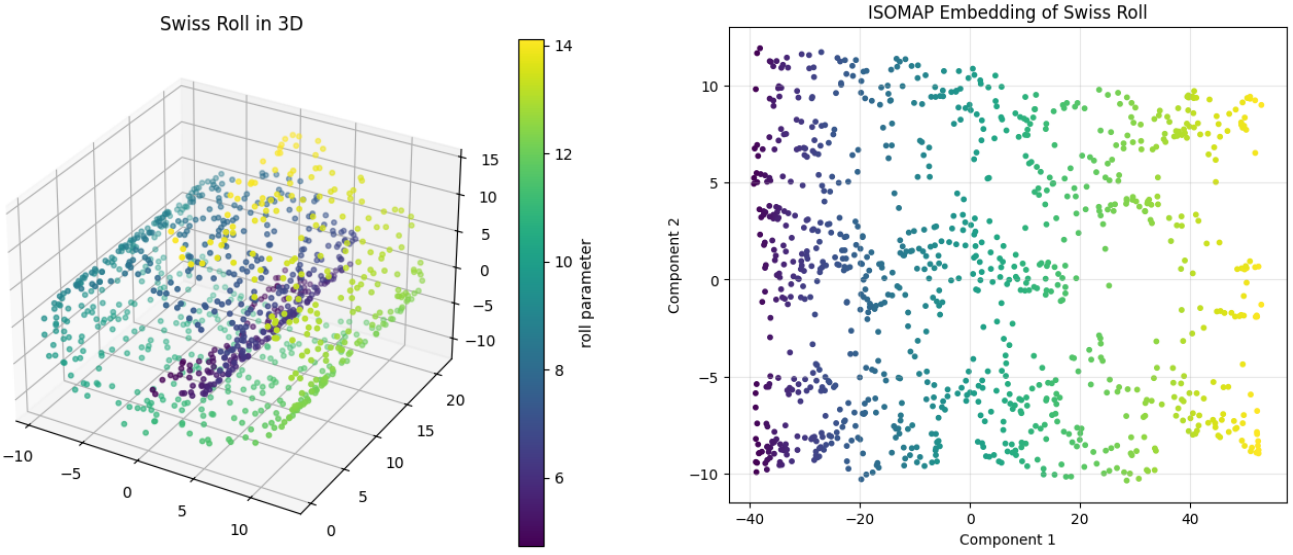
where W_d contains the eigenvectors associated with the largest values and the top eigenvectors are the principal components mentioned.

Another important classical method is Multidmensional Scaling (MDS). As explained by Kruskal, MDS aims to place points in a low-dimensional space so that pairwise distances are preserved as much as possible [KW78]. Although PCS and MDS are important classical methods, they do not explicitly model non-linear manifold structure. When data lie on a curved manifold, Euclidean distances in the ambient space can be misleading [TSL00].

2.3 Non-Linear Dimensionality Reduction

The main idea behind manifold learning is that although observed data may lie in a high-dimensional ambient space, the intrinsic degrees of freedom governing the data can be much lower-dimensional. Methods in this sphere attempt to uncover that structure by preserving neighbourhood relations or approximating geodesic geometry rather than relying only on global Euclidean distance. This

perspective is very important for methods such as ISOMAP [TSL00]. It is important to note the difference between Euclidean and geodesic distance. In manifold learning, two points may appear close in Euclidean space even though they are far apart when measured along the manifold, this can be imagined as a series of stations along a long railway. Tenenbaum motivates for ISOMAP by arguing that geodesic distance is often more appropriate when data lie on a non-linear manifold as seen in Figure 1. Calculating the geodesic distance may prove to be challenging if the manifold in which the data is located is unknown.



(a) Swiss roll dataset

(b) ISOMAP embedding of the swiss roll

Figure 1: These two figures display the importance of measuring the geodesic distance along the manifold as the euclidean distance of two pairs of points in the different layers of the roll is misleading.

2.3.1 t-Distributed Stochastic Neighbour Embedding

t-distributed Stochastic Neighbour Embedding (t-SNE) was introduced by van der Maaten and Hinton as an extension of earlier stochastic neighbour based embedding method. The method aims to preserve local neighbourhood structure by converting pairwise euclidean distances into probabilities and then finding a low-dimensional embedding whose probability distribution matches them as closely as possible [MH08]. A key difference between t-SNE and SNE is the use of t-distribution in the low-dimensional space, which helps alleviate the crowding problem and generally produces more interpretable visualizations.

To define these probability distributions more formally, t-SNE first represents similarities between points in the high-dimensional space using conditional probabilities. The conditional probability that point x_i selects x_j as its neighbour is given by

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

Here σ_i controls the width of the Gaussian distribution centred at x_i, and therefore determines the scale of the local neighbourhood around that point which is also known as the local density. Local density plays an important role in the behaviour of the technique which will be explained later.

An important parameter in t-SNE is the perplexity, which controls the effective number of neighbours considered around each point. Instead of fixing σ_i directly, t-SNE chooses it separately for each point so that the entropy of the conditional Distribution P_i matches a user-defined perplexity value. The perplexity is defined as

$$\text{Perp}(P_i) = 2^{H(P_i)}$$

where $H(P_i)$ is the Shannon entropy

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

. In this way, perplexity can be interpreted as a smooth measure of the effective neighbourhood size. Smaller perplexity values emphasise a very local structure, whereas larger values take into account a broader neighbourhood.

Additionally, t-SNE uses a symmetric version of these similarities, the joint probabilities in the high-dimensional space are defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

where n is the number of the data points and uses a factor of 2 for normalisation which ensures that $\Sigma_j p_{ij} > \frac{1}{2n}$ for all data points. Symmetrising the conditional probabilities allows all the data points to have an equal influence on the cost function. Additionally it greatly simplifies the gradient of the cost function to help save computation time. In the low-dimensional space, the similarity between the embedded points y_i and y_j is defined using a Student t-distribution:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

Compared to a Gaussian distribution, the heavier tails of the Student t-distribution allow moderately distant points to be placed farther apart, which helps reduce the crowding problem [MH08].

The final embedding is obtained by minimizing Kullback-Liebler divergence between the probability distribution in the high-dimensional space and the one in the low-dimensional space:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The gradient of the KL-divergence between symmetrised P and Q (student-t distributed joint probabilities Q) is given by:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}$$

And the result of this gradient descent on each data point can be interpreted as a set of springs connecting all the data points to each other with positive values corresponding to attraction between two points and negative values as repulsion force.

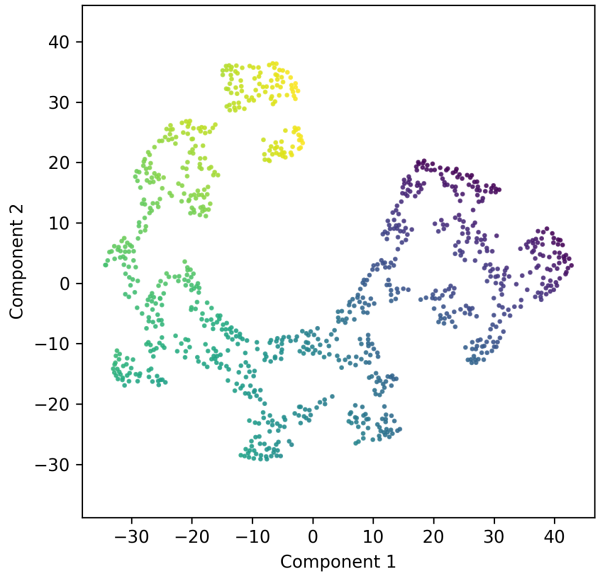


Figure 2: t-SNE embedding on the Swiss roll dataset

2.3.2 UMAP

Uniform Manifold Approximation and Projection, or UMAP, is another non-linear dimensaionality reduction method proposed by McInnes, Healy and Melville. Similar to t-SNE, UMAP constructs similarity values in the high dimensional space and then searches for a low-dimensional embedding that reflects these similarities. The mathematical foundations of this method is heavily influenced by network theory and researching of topologies but parallels can be found between the two methods of UMAP and t-SNE with some caveats. In t-SNE the conditional probabilities of *all* the points are calculated. In UMAP however, the high dimensional similarities are computed for the k-nearest neighbours of each point. And non-zero similarity values are only assigned to points that are among the k-nearest neighbours of a given point.

The parameter k, which denotes the number of neighbours, plays a similar role to perplexity in t-SNE because it controls the effective neighbourhood size. Small values of k make UMAP focus more strongly on local structure, whereas larger values allow the method to take broader structure into account. Nevertheless, unlike perplexity, k also directly determines which pairs of points are considered when constructing the high-dimensional similarity graph.

For a point x_i, let x_j denote the j-th nearest neighbour of x_i. UMAP defines the directed similarity from x_i to x_{ij} as:

$$v_{j|i} = \frac{\exp(-\max(0, d(x_i, x_{ij}) - \rho_i)}{\tau_i}$$

where $d(x_i, x_ij)$ is a dissimilarity measure, usually Euclidean distance. The value ρ_i is the local connectivity , the distance from x_i to its nearest non-identical neighbour:

$$\rho_i = \min(d(x_i, x_{ij})$$

, with $d(x_i, x_{ij} > 0$. The purpose of ρ_i is to ensure that each point has at least one neighbour with similarity equal to 1. This means that the closest neighbour of each point is treated as fully connected in the graph.

The parameter τ_i is a local scale parameter. It is chosen separately for each point by a binary search so tha the similarities around x_i satisfy:

$$\sum_{j=1}^{k} \exp\left(\frac{(-\max(0, d(x_i, x_{ij}) - \rho_i)}{\tau_i}\right) = \log 2k$$

This makes the similarity values adaptive to the local density of the data. In dense regions, τ_i tends to be smaller, while in sparse regions it tends to be larger. Therefore, UMAP does not use one global distance scale for the whole dataset.

It is important to note that these high-dimensional similarities are only computed for the k nearest neighbours. If x_j is not one of the k nearest neighbours of x_i, then $v(i|j)$ is set to zero. Also, self similarities are ignored, so $v(i|i) = 0$. Since nearest-neighbour relations are directed, it is possible that x_j is a neighbour of x_i, while x_i is not a neighbour of x_j. Similar to t-SNE, UMAP symmetrises the directed similarities using a fuzzy union. If A is the matrix with entries $A_{ij} = v(j|i)$, then the symmetric high-dimensional similarity matrix V is defined as:

$$V = A + A^T - A \circ A^T$$

Where $\circ$ denotes entry-wise multiplication. Equivalent, for each pair of points:

$$v_{ij} = v_{(j|i)} + v_{(i|j)} - v_{(j|i)}v_{(i|j)}$$

This value can be interpreted as the strength of the connection between x_i and x_j in the high-dimensional fuzzy graph. If either point strongly consider the other to be a neighbour, then the final symmetric similarity v_{ij} will be large.

In the low-dimensional embedding, UMAP defines another similarity matrix W. The entry w_ij measures how similar the embedded points y_i and y_j are. Unlike the high-dimensional matrix V, the low-dimensional similarities are computed for all pair of points, not only for the nearest neighbours. There are two common ways to define w_{ij}. The first is the true low-dimensional similarity function:

$$\Psi(y_i, y_j) = \begin{cases} 1, & \text{if } \|y_i - y_j\| \leq \text{min\_dist}, \\ \exp\left(-(\|y_i - y_j\| - \text{min\_dist})\right), & \text{otherwise.} \end{cases}$$

Here, *min_dist* controls how close points are allowed to be in the embedding. If the distance between y_i and $y_j$is smaller than *min_dist*, their similarity is set to 1. For larger distances, the similarity decays exponentially. Smaller values of *min_dist* allow tighter clusters, while larger values produce more spread out embeddings.

The second option is a smooth approximation of this function, which is used in the UMAP optimization. This approximation is usually written as:

$$\Phi(y_i, y_j) = \left(1 + a(||y_i - y_j||_2^2)^b\right)^{-1}$$

where a and b are fitted parameters chosen to approximate the true curve determined by *min_dist* and spread. This smooth function is useful because it makes gradient-based optimization easier. Therefore, the low-dimensional similarity matrix W can be defined by the true function Ψ or by the approximation Φ. In both cases, w_{ij} is close to 1 when $y_i$and $y_j$are close together, and decreases as the embedded points move farther apart.

The final UMAP embedding is obtained by making the low-dimensional graph W resemble the high-dimensional graph V.This is usually formulated as minimising cross-entropy loss between the two graphs. Intuitively, pairs with high v_{ij} should have high w_{ij}, meaning that neighbours in the original space should remain close in the embedding. Conversely, pairs with low v_{ij} should not have high w_{ij}, because this would mean that the embedding has created false neighbours.

3 Diagnostics Framework and Package Design

In the previous section, we explained some of the prominent dimension reduction techniques both linear and non-linear. And for some, the internal mechanisms was also explained. In order to create a diagnostic tool for these techniques, it is important to come up with some framework to enhance the interpretability of the dimension reduction. Creating visualizations of the embeddings and the different characteristics is helpful but the great challenge lies in identifying meanings and relations in such visualizations.

3.1 Distance-fit Plots

One of the most natural quantities to inspect when diagnosing a dimensionality reduction embedding is the set of pairwise distances between observations. Distances are intuitive because we have the mental capacity to imagine how close or far two points are in space. If two points are close together in a two-dimensional plot are usually interpreted as similar, where as distant points are interpreted as dissimilar. This idea has originally stemmed from Multidimensional Scaling (MDS) which we reviewed earlier in the report. The goal of the method is to find a low-dimensional configuration whose pairwise distances correspond as closely as possible to the observed dissimilarities between the objects. Kruskal describes MDS as the problem of representing objects geometrically by points so that the pairwise distances match the original dissimilarities [Kru64].

A common diagnostics for this relationship is the Shephard diagram, which plots the original-space dissimilarities or distances against distances in the low-dimensional embedding. If the embedding preserves pairwise structure well, the points in the Shephard diagram should follow a clear increasing trend. In metric settings this trend may be approximately linear this is due to the transformation function that is used in metric MDS. This is also known as the ratio transformation [GvdV04], so that the fitted disparity is equal to the observed dissimilarity. To show an example of such a plot (see Figure 3), we have run metric MDS on Iris dataset and plotted the original distances in the x-axis and the embedding distances in the y-axis. This is important because we can use similar plots for non-linear methods to understand how distances are preserved. That includes plotting of the pairwise euclidean distances in the high-dimensional against the pairwise euclidean distances in the low-dimensional embedding. Furthermore, we can also take into account the geodesic distances in the high-dimensional space. To find the geodesic distance we use the method used in ISOMAP algorithm by first creating a neighbourhood graph by connecting each point using k-nearest neighbour and then find shortest path in the neighbourhood graph by Dijkstra's algorithm [TSL00]. Including both Euclidean and geodesic distance fit-plots allows the diagnostic framework to distinguish between ambient-space distance preservation and intrinsic manifold preservation. Euclidean distances measure straight-line relationships in the original feature space, while geodesic distances approximate distances along the underlying data manifold. For

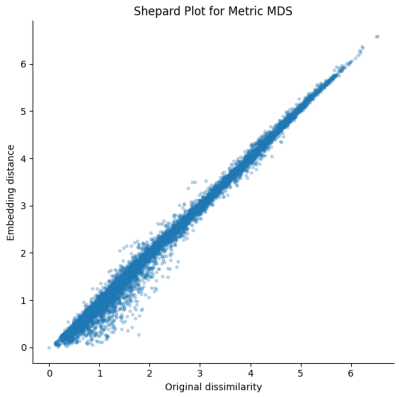


Figure 3: Shephard plot of metric MDS with ratio transformation on Iris dataset.

example, Figure 4 illustrates how distances are preserved, the disconnect between the two clusters is due to the inherent structure of the Iris structure.

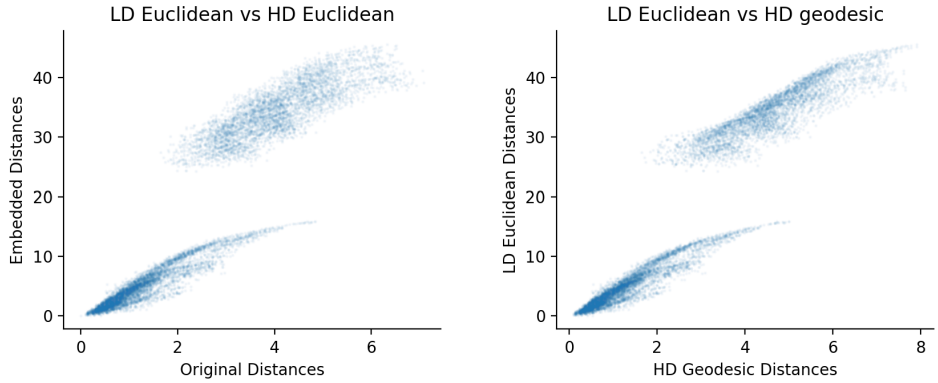


Figure 4: Left) Pairwise Euclidean distances in the high-dimensional space vs the Euclidean distances in the low dimensional space on Iris dataset using t-SNE Right) Pairwise geodesic distances in the high-dimensional space vs the pairwise Euclidean spaces in the low-dimensional space on Iris dataset using t-sne

It is important to note that these trends will not always be the case. The Iris dataset is a small data set and is linearly separable and MDS does a good job at preserving both the local and global distances as it is designed to do such a thing since it takes into account ALL of the pairwise distances and not only within the local neighbourhood of each point. However, this type of plots can still be used to see if at all the distances are preserved or not.

3.2 Similarity Plots

For the dimension reduction methods discussed for t-SNE and UMAP, we can use distance plots. Additionally, by looking deeper at the internal mechanisms of the methods we can extract more information and visualize them. We have reviewed that t-SNE and UMAP convert pairwise distances into similarity scores. We can then construct similarity matrices for P and Q for t-SNE and likewise

for UMAP, V and W. Even though they are both similarity matrices their interpretations are not similar to each other.

To further iterate the design choices of these mappings, they are mainly focused in preserving the *local neighbourhood* of points. We also discussed that the neighbourhood size in these algorithms are hyperparameters; perplexity for t-SNE and neighbourhood size k in UMAP. Thus based on the hyperparameter the algorithms are not interested to preserve distances outside the range of the neighbourhood size.

To visualize such plots, we ran t-SNE and UMAP on the Iris data set and plotted the P vs Q similarities. Figure 5 shows the t-SNE similarity matrices for the Iris dataset with perplexity of 30. Each point corresponds to a pair of observations. The x-axis gives the high-dimensional similarity p_{ij}, while the y-axis displays the low-dimensional similarity q_{ij}. The increasing trend indicates that pairs with larger similarity in the original space tend to remain more similar in the embedding. The large concentration of points near the origin occurs because the plot contains all pairwise relationships. In a dataset with n observations, there are $n(n-1)/2$ unique pairs, but in fact only a small subset of these pairs are close neighbours. In t-SNE, non-neighbouring pairs receive a very small similarity in high dimensions and if they are also far apart in the embedding, their low-dimensional similarities are also close to zero. Therefore, most pairs appear near the origin, while the more informative part of the plot is the upward trend away from the origin, which shows how well local similarities are preserved. Moreover, the spread around the trend also shows

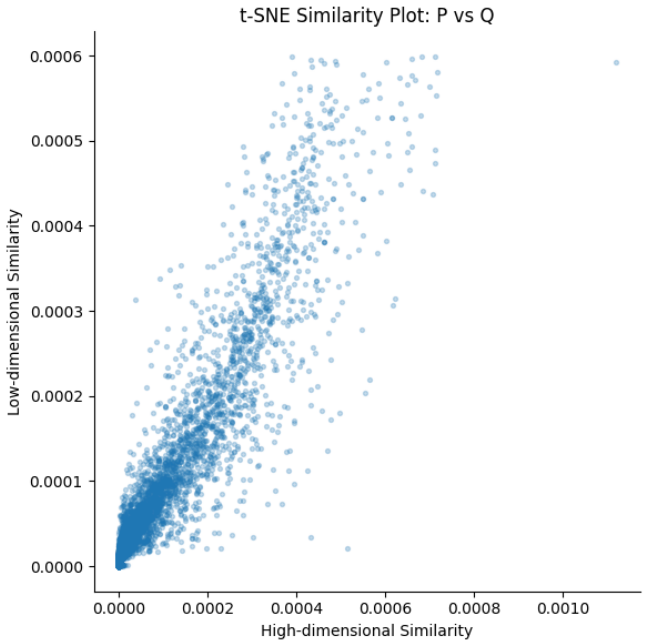


Figure 5: Similarity plot of the P and Q matrices on the Iris data set with perplexity of 30.

that the preservation is not exact. Pairs above the trend are represented as more similar in the embedding than in the original data, whereas pairs below the trend are represented as less similar. This reflects the fact that t-SNE optimizes a probabilistic neighbourhood objective rather than preserving euclidean distances directly.

As for similarity plot in the UMAP embedding, presented in Figure 6 looks different to one

obtained via t-SNE and this is due to the internal mechanism of UMAP. Each point corresponds to a pair of observations similar to t-SNE. The x-axis is the high-dimensional similarity v_{ij} and the y-axis is the corresponding low-dimensional similarity w_{ij}. The concentration of points at $v_{ij} = 0$ occurs because UMAP constructs its high-dimensional representation from a nearest-neighbour graph, so most pairs are not connected in the original graph. Some of these pair nevertheless have moderate to high low-dimensional similarity scores which indicate false neighbours introduced by the embedding. The concentration near top right of the graph at v_{ij} and w_{ij} are consequences of UMAP's similarity definitions. In the high-dimensional space, the local-connectivity ρ_i ensures that each point has at least one neighbour with membership strength equal to one. This is due to:

$$v_{j|i} = \frac{\exp(-\max(0, d(x_i, x_{ij}) - \rho_i)}{\tau_i}$$

here, ρ_i is the distance from x_i to its nearest non-identical neighbour. Because of the term

$$\max(0, d(x_i, x_j) - \rho_i$$

the closest neighbour of each point gets distance adjustment 0. Thus:

$$v_{j|i} = \exp(0) = 1$$

After symmetrisation by the union formula:

$$v_{ij} = v_{(j|i)} + v_{(i|j)} - v_{(j|i)}v_{(i|j)}$$

Any pair tha is maximally connected in either direction also receives $v_{ij} = 1$. In embedding space, the low-dimensional similarity:

$$w_{ij} = \frac{1}{1 + a||y_i - y_j||^{2b}}$$

This means that if two embedded points are very close then the term $||y_i - y_j||$ is close to 0, thus:

$$w_{ij} \approx \frac{1}{1+0} = 1$$

So the concentration near $w_{ij} = 1$ and $v_{ij} = 1$ is an important part of the plot which correspond to pairs were strongly connected in the original and the embedded map. Although both diagnostics compare high-dimensional and low-dimensional similarity matrices, the t-SNE and UMAP plots have different shapes because the underlying similarities are defined differently. In t-SNE, P and Q are normalized probability distributions over pairs of points, so individual entries are very small and most points appear near the origin. In UMAP, V and W represent fuzzy graph membership strengths between 0 and 1 and are not normalized to sum to one. Moreover, UMAP constructs V from a nearest-neighbour graph, which gives many pairs a high-dimensional similarity of zero. Therefore, the UMAP plot contains vertical bands at values such as $v_{ij} = 1$ and $w_{ij} = 1$, while the t-SNE plot appears as a dense probabilistic similarity cloud. The two plots should therefore be interpreted within the internal similarity model of each method rather than compared directly by scale.

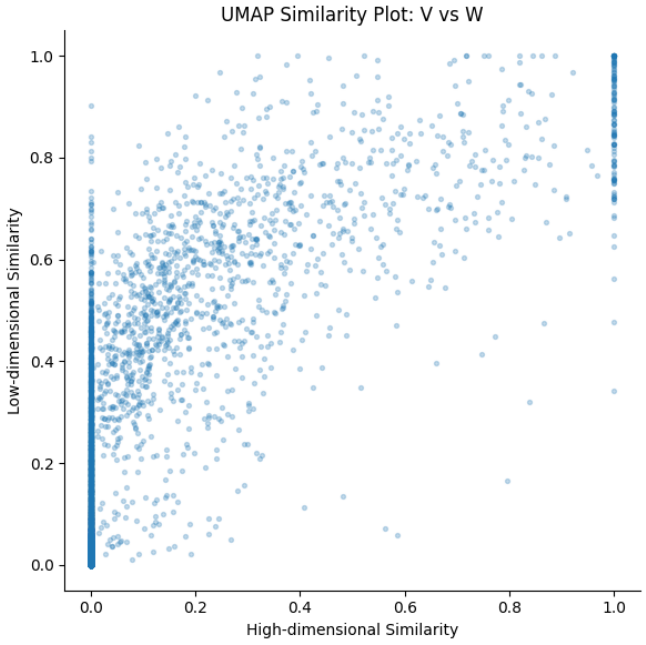


Figure 6: Similarity plot of the V and W matrices on the Iris data set with k=30.

3.3 Rank-Based Diagnostics

Other assessment criteria of dimensionality reduction methods that are also discussed in the literature involve inspecting the ranking of pairwise points before and after the embedding. This is an intuitive way to think about the quality of an embedding, because it does not require the exact pairwise distances to be preserved. Instead, it asks whether the relative neighbourhood structure of each point is preserved. An important question here is whether points that were close to each other in the original high-dimensional space stay close when they are mapped onto the lower-dimensional space. Another important question is whether the embedding has created false neighbours, meaning that points that were far apart in the original space have ended up as neighbours in the embedding.

This idea is related to the concepts of trustworthiness and continuity, which are commonly used as rank-based quality criteria for dimensionality reduction. These measures were discussed by Venna and Kaskl in the context of neighbourhood preservation and local multidimensional scaling, Where they explain the difference between preserving neighbours from the original space and avoiding false neighbours in the embedding [VK01, VK06]. The same idea is later developed more systematically by Lee and Verleysen, who describe dimensionality reduction quality assessment in terms of rank-based criteria and neighbourhood preservation [LV08, LV09]. trsutworthiness measures whether points that are close in the embedding were also close in the original high-dimensional space. Therefore, it penalizes false neighbours introduced by the embedding. Continuity measures the opposite effect by checking whether points that were close in the original space remain close in the embedding, and therefore penalizes original neighbours that are separated after dimensionality reduction.

It is important to note why such behaviour is possible. For non-linear methods such as t-SNE and UMAP, the neighbourhood preservation is influenced by local scaling. As mentioned in previous sections; σ_i for t-SNE and τ_i for UMAP are calculated separately for every point, This local scaling

factor changes based on the density of the points in different regions of the space. In dense regions, σ_i becomes smaller in order to match the desired perplexity, while in sparser regions it becomes larger. UMAP uses a similar idea, although the notation is different. Instead of σ_i the paper uses τ_i or ρ_i to adapt the size of local neighbourhood of each point. The local scaling factor can be thought of as a circle around a point, where the scaling factor controls the radius of the neighbourhood considered around that point. If the point is in a dense region, the raius of this circle will be smaller, while in less dense regions the radius will be larger. Conceptually, this design feature can also lead to misleading embeddings. For example, if some points are located in a sparse region, they may recieve a larger local scale. These points could simply be outliers, but because their local neighbourhood is expanded, they may still end up being treated as neighbours of other points. This can lead to false neighbours in the final embedding and may cause misinterpretations when analysing specific points or cluster of points.

Going back to concepts of trustworthiness and continuity, we can formalize them by the following: For a dataset with n number of points, let r_{ij}^X be the rank of point j with respect to point i in the original space, and let r_{ij}^Y be the corresponding rank in the low-dimensional space. Also let $N_i^Y(k)$ and $N_i^Y(k)$ denote the sets of the k-nearest neighbours of point i in the original and embedded spaces, respectively. Trustworthiness is then defined by using the set $U_i(k) = \frac{N_i^Y(k)}{N_i^X(k)}$, which contains points that are neighbours in the embedding but not in the original space. Conversely, continuity can be defined as the set $V_i(k) = \frac{N_i^X(k)}{N_i^Y(k)}$, which contains points that are neighbours in the original space but not in the embedding. The two measures are then calculated via: for trustworthiness

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_i(k)} \left(r_{ij}^X - k\right)$$

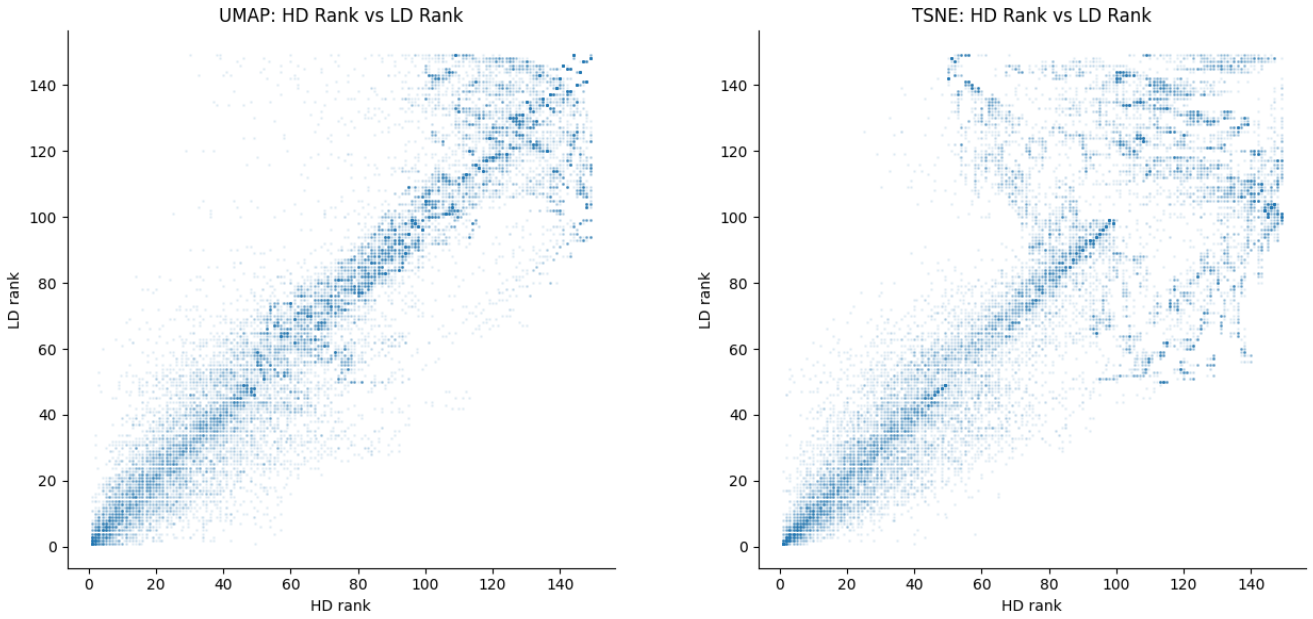
and for continuity:

$$C(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in V_i(k)} \left(r_{ij}^Y - k\right)$$

The term $r_{ij}^X - k$ measures the magnitude of the mistake. If the point was just outside the original k-neighbourhood, the penalty is smalll. If it was very far away the penalty will be large. The fraction in the front is a nomalization term that scales the score so it lies between 0 and 1. A trustworthiness score close to 1 means the embedding does not create many false neighbours. Additionally, we can take a look at "ALL" of the points if we set $k = n$.

In order to show an example, we have taken the same Iris dataset and computed the t-SNE and UMAP embeddings, see Figure 7b for t-SNE and 7a for UMAP. And plotted the ranks in the original high-dimensional space against the low-dimensional space

The rank comparison plots show the relationship between the high-dimensional rank and the low-dimensional rank for UMAP and t-SNE. In an ideal embedding, all points would lie close to the diagonal, meaning that pairs of points have approximately the same neighbourhood ranking before and after dimension reduction. Both plots show a clear increasing diagonal trend, which indicates that both methods preserve a considerable amount of the neighbourhood structure. This is especially visible for small ranks, where points that are close in the original space often remain close in the embedding. However, the spread around the diagonal shows the preservation is not perfect. Points about the diagonal correspond to pairs that are ranked farther away in the embedding than

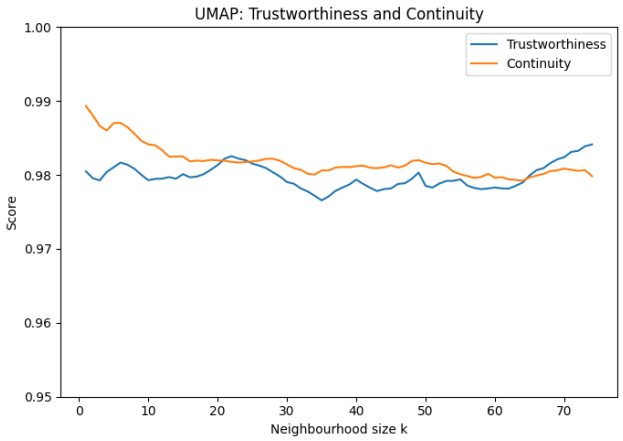


(a) Rank comparison for UMAP

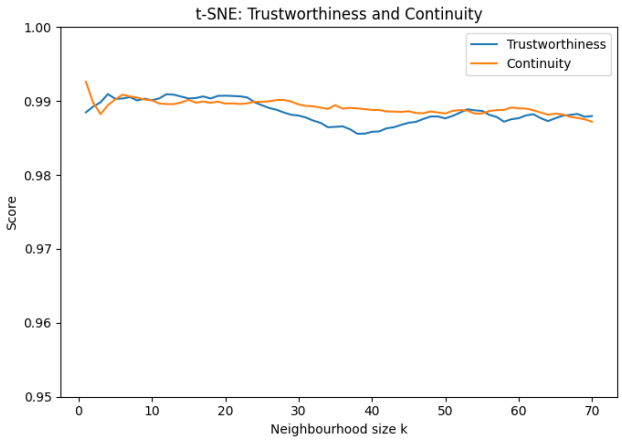
(b) Rank comparison for t-SNE

Figure 7: Comparison of high-dimensional and low-dimensioopnal pairwise ranks for UMAP and t-SNE

in the original space. These can be interpreted as original neighbours that were separated after embedding. Points below the diagonal correspond to pairs that became closer in the embedding than they were in the original space, which can indicate false neighbours. Nevertheless, both methods performed reasonably well at preserving neighbourhoods, but this is also due to the fact the iris dataset is a small dataset and is linearly separable, so it was easy for these methods to perform well. This can be further iterated by looking at the continuity and trustworthiness plots as illustrated in Figures 8a & 8b. The Trustworthiness and Continuity plots show that both UMAP and t-SNE preserve neighbourhood relationships well on the Iris dataset. For both methods, the scores remain close to 1 across all neighbourhood sizes k, which indicates that few false neighbours are introduced and few original neighbours are lost during the embedding. The t-SNE embedding obtains slightly higher and more stable scores than UMAP in this experiment, with both trustworthiness and continuity staying around 0.99. UMAP also performs well, although its scores are slightly lower and show a small decrease as k increases. Furthermore, The upper-left region of the rank scatter corresponds to trustworthiness violations meaning they pairs are close in low-dimensional space but far in the high-dimensional space; lower right regions of the rank scatter corresponds to continuity violations. The fan shape widening at higher ranks in both methods is expected as they try to preserve local structure and is not considered a failure. Overall, the figure indicates that both methods produce reliable neighbourhood-preserving embeddings, which is expected for the Iris dataset because it is small and relatively low-dimensional.



(a) UMAP



(b) t-SNE

Figure 8: Trustworthiness and Continuity scores for UMAP and t-SNE over different neighbourhood sizes.

3.4 Selected Point and Pair Neighbourhood Diagnostics

In addition to global diagnostic plots, the package includes a set of local diagnostic tools that allow the user to inspect individual observations or specific pair of points. Global plots show the overall behaviour of an embedding across all pairwise relationships, but they may mask local distortions that affect only a small number of observations. The local diagnostics described in this section are designed to make such distortions visible at the level of individual points and pairs.

3.4.1 Selected point neighbourhood

The first local diagnostic is the selected point neighbourhood. This feature allows the user to choose a single observation and highlight its k-nearest neighbours in the original high-dimensional space across all panels. Each panel corresponds to one of the diagnostics plots described in the previous sections. The top-left panel shows the original data in two-dimensional view obtained by PCA, where the selected point is marked in red and it k-nearest neighbours are shown in orange, while all remaining points are shown in grey. The top-middle panel shows the same information in the low-dimensional embedding. By inspecting these two scatter plots together, the user can immediately asses whether the neighbourhood of the selected point is preserved after the embedding, or whether the neighbours have been scattered or mixed with points that were originally distant.

The remaining four panels highlight the same selected point-neighbour pairs in red against the grey background of all other pairs. In the similarity plot, the highlighted points correspond to the pairs between the selected observation and each of its neighbours. If these pairs follow the global trend, the local similarities are well preserved. If the appear above the trend, the embedding has made the selected point and its neighbourhood more similar than they were in the original space, which corresponds to an overestimation of local similarity. In the Euclidean and geodesic distance fit plots, the highlighted pairs show where the selected neighbourhood sits relative to all pairwise distances. Pairs located in the lower-left region of both distance plots indicate that the selected point and its neighbours are close in both original and embedding which is evidence of good local

preservation. If instead the highlighted pairs deviate strongly from the global trend, this points to local distortion concentrated around that specific observation. Finally, in the rank comparison plot, the highlighted pairs show the HD and LD ranks for each selected point-neighbour relationship. In a well-preserved embedding, all highlighted pairs should appear close to the diagonal. As mentioned previously, points in the upper-left region correspond to continuity violations, meaning that the original neighbour has been pushed far away in the embedding. Points in the lower-right region correspond to trustworthiness violations, indicating a false neighbour has been introduced around this point. To illustrate this, Figure 9 shows the elected point neighbourhood dashboard on the Iris dataset using t-SNE, with point 140 selected and $k = 10$ neighbours. The neighbours remain close to the selected point in both the original and embedding scatter plots, suggesting that the local neighbourhood is well preserved for this observation. The distance fit plots also confirms this as the point and its neighbours have small distances.

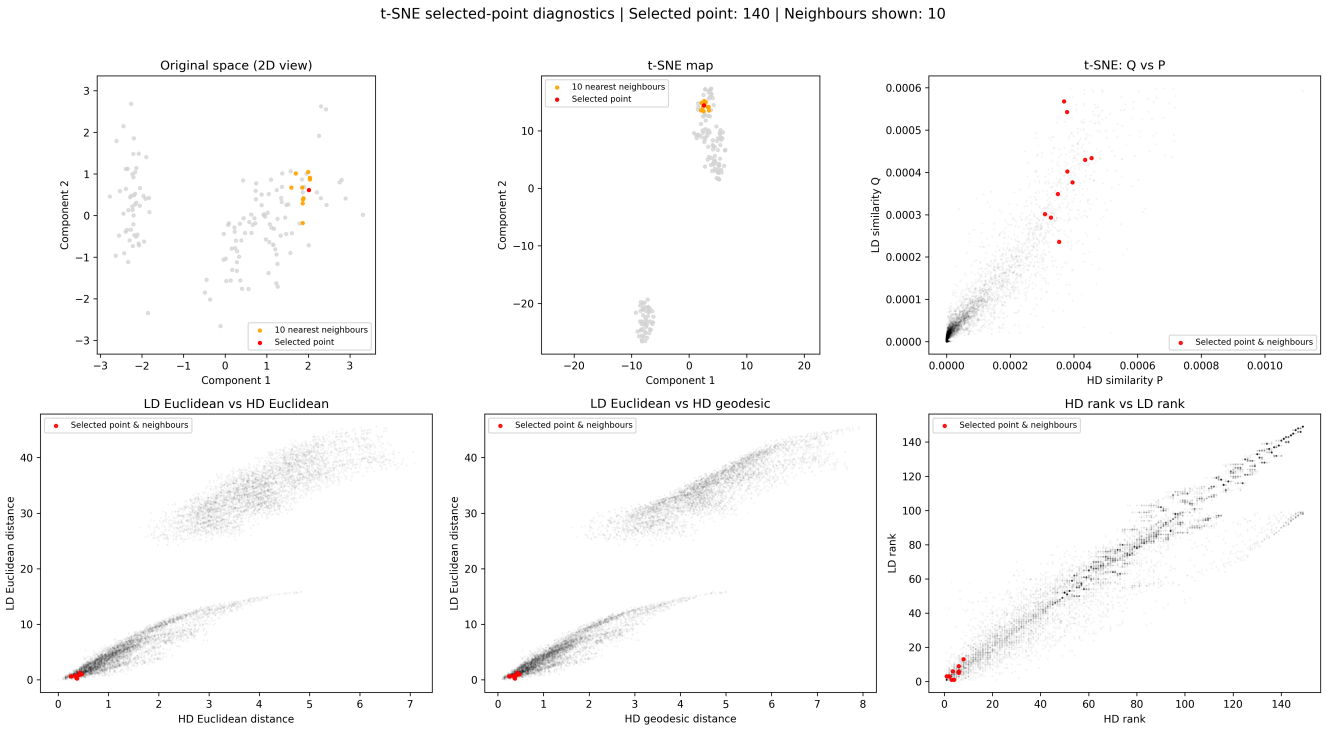


Figure 9: Complete dashboard of point 140 and its 10 nearest neighbours in the IRIS dataset using t-SNE

3.4.2 Selected pair diagnostics

While the selected point dashboard inspects the full neighbourhood of a single observation, It is sometimes more informative to examine one specific pair of points in detail. For example, the user may want to understand why two particular points appear close in the embedding even if they were far apart in the original space, or other way around. For this purpose, the package also includes a functionality to select a pair of points and instead of highlighting a neighbourhood, the user specifies two point indices, referred to point A and point B. The dashboard then uses the same six panel layout as the main diagnostic dashboard to keep the visualisations consistent. In the

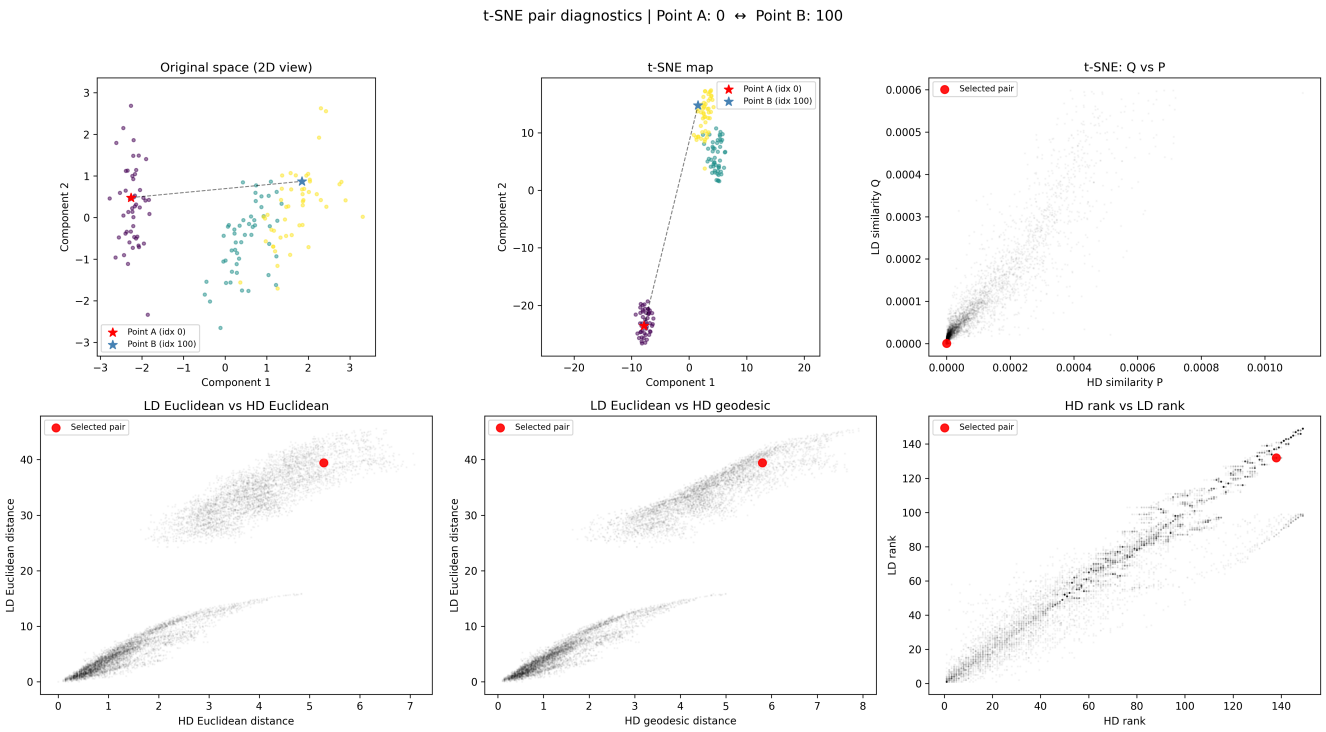


Figure 10: Complete dashboard of points 0 and 100 in the IRIS dataset using t-SNE

scatter plots, the selected pairs are marked with a star and connected to each other by a dashed line, making it easier to see how their relative positions change between the HD and LD spaces. Each of the four remaining plots highlights the single data point corresponding to the selected pair of points in colour red. Additionally, another feature that synergises with pair selection is a way to find the most distorted pair of points. The reason is that manual selection of pairs of interest for the user is not very straightforward, as it requires the user to know every point indices. As one of the methods to be able to find pairs of interest, it is natural for user the look for the most distorted pairs. The pairs can be found by fitting a linear regression through all pairwise HD and LD distances and compute the residual for each pair. A large positive residual means that the LD distance is much higher than the regression predicts. A large negative residual means the LD distance is much lower than expected, indicating compression. With such a function we can return the n most distorted pairs sorted by absolute residual and their point indices which then can be inserted by the user for further inspection. Figure 10 shows the dashboard for the selected pair of points with indices 0 and 100 respectively in the IRIS dataset using t-SNE.

4 Experiments

4.1 Experimental Setup

The experiments in this chapter are designed to demonstrate the diagnostic framework described in Chapter 3 on a set of concrete datasets. Rather than evaluating the performance of dimensionality reduction methods in isolation, the goal is to show how the different diagnostic plots and tool can

be used together to understand the quality of an embedding in practice. Each dataset has been selected due to their ability in highlighting different aspects of the diagnostics, and together they illustrate how the package can be applied to data with different structures and dimensionalities.

All experiments are run using two dimensionality reduction methods t-SNE and UMAP, so that the diagnostics can also be used to compare the behaviour of both methods on the same data. For t-SNE we use a *perplexity* of 30 and 1000 iterations unless stated otherwise. For UMAP, we use a neighbourhood size of $k = 15$ and a minimum distance of 0.1. These are commonly used default values for both methods and serve as a reasonable starting point for the experiments. Where the effect of hyperparameters is relevant, we discuss this separately. All experiments use a fixed random state seed to ensure reproducability. The geodesic distances are compute using the k-nearest neighbour graph approach described in Section 3.1, with the same neighbourhood size as used in the embedding method. All figures in this chapter are generated using the dr-diagnostics package.

4.1.1 Datasets

The first dataset is the Iris dataset first mentioned by R.A. Fisher [Fis36], which consists of 150 observations of four measurements of iris flower, belonging to three species. This dataset has been used throughout Chapter 3 to introduce the individual diagnostic plots. In this chapter we bring those results together into a unified comparison of the methods. This dataset is a good starting point since it is small, low-dimensional, and linearly separable meaning both methods are expected to perform well. This makes it a good baseline for interpreting what the diagnostics lok like when the embedding quality is high.

The second dataset is the Swiss Roll, which is a synthetic three dimensional dataset in which the data points lie on a curved manifold rolled up in a spiral shape. For this dataset we generate 1000 points with a small amount of Gaussian noise. This dataset is particularly relevant for the geodesic distance diagnostics because the Euclidean distance between two points on the opposite sides of the roll can be small even if the true distance along the manifold is large. This makes the Swiss roll an important test case for evaluating whether the embedding respects the intrinsic geometry of the data or only the Euclidean distrances.

The third dataset is the Digits dataset from scikit-learn [PVG+11], which contains 8x8 pixel grayscale images of handwritten images of handwritten digits from 0 to 9. Each image is represented as a 64-dimensional vector. We use a subset of 500 observations to keep the computation load low while still covering all ten digit classes. This dataset is chosen due to it high-dimensionality where the data has a meaningful class structure but the underlying geometrical structure is more complex than in the previous two datasets. We use this dataset to demonstrate the complete diagnostic workflow, including the automatic detection of distorted pairs using outlier detection function introduced in Section 3.4 and the subsequent inspection of those pairs.

The last dataset is another synthetic dataset which will consist of two clusters of different density and a point that that is an outlier. The dataset will have three different classes and we believe we can highlight many of the inherent traits of t-SNE and UMAP with this dataset.

Dataset	Samples	Features	Classes	Structure
Iris	150	4	3	Linear
Swiss Roll	1000	3	-	Non-linear manifold
Digits	500	64	10	High-dimensional
Blobs	1000	2	3	Different density clusters

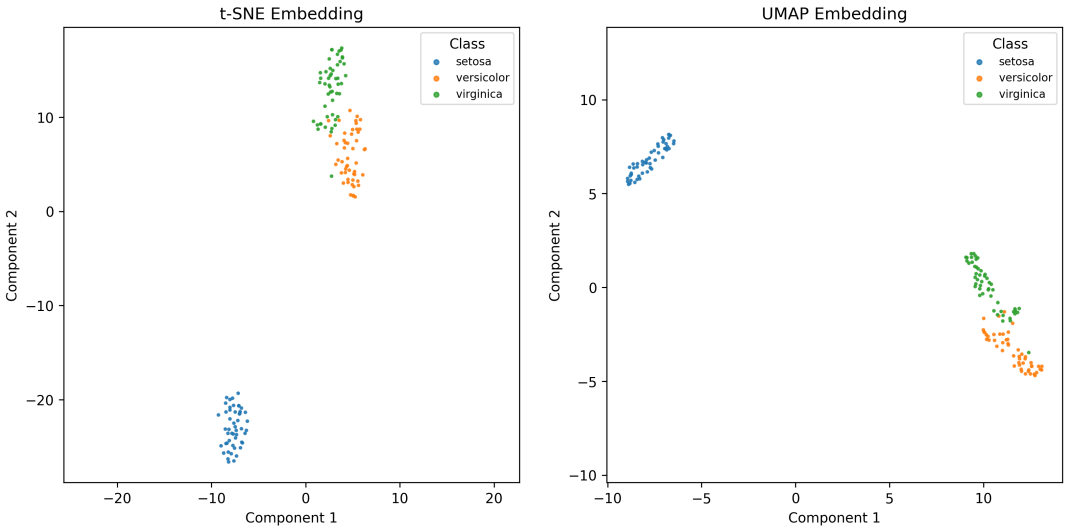
Table 1: Summary of the datasets and their properties

4.2 Iris Dataset

First, let us look at the original data, which is a PCA projection of the data into 2 dimensions, and look at the two different embeddings and compare them on a surface level. The PCA projection of the original data is depicted in Figure 11. The two different embeddings are illustrated in Figures 12a and 12b. In the original data, the setosa class is clearly separated from the other two species along the first principal component. Versicolor and virginica however, show more overlap between them with point from both classes occupying the same region of the PCA projection. This is due to the two species sharing similar morphological measurements in four-dimensional space and are not linearlt separable from each other.

In both the t-SNE and UMAP embeddings, we can see three visually distinct clusters. Setosa remains isolated in both embeddings, which is consistent with its clear separation in the original data. The more interesting observation is what happens to versicolor and virginica. In both embeddings, these two classes appears distinct and with only a few points overlapping. This means that both methods have introduced a separation between versicolor and virginica that does not fully exist in the original data. This is an important observation and is one of the known characteristics of non-linear dimensionality reduction methods such as t-SNE and UMAP. Since these methods optimize for local neighbourhood preservation, they tend to pull points of the same local neighbourhood closer together while pushing other points further apart, which can exaggerate the boundaries between classes.

Figure 11: PCA projection of the original data in the Iris dataset



(a) t-SNE (b) UMAP

Figure 12: t-sNE and UAMP mapping of the Iris dataset

Next, we investigate the distance plots of both embeddings in the Euclidean space as shown in Figure 13. Both plots exhibit a clear two band structure, which is a consequence of the structure of the original data. The lower band corresponds to within-class pairs, meaning pairs of observations from the same species. These pairs are close in the original high-dimensional space and remain close in the embedding, so they appear in the lower-left region of the plot. The upper band corresponds to between class-pairs, meaning the observations from different species. These pairs are farther apart in the original space and the embedding pushes them even further, placing them in the upper region of the plot. The gap between the two bands is the separation of the clusters, which we already observed in Figure 12. This two band structure will not appear in every dataset and is specific to data with clear separation of clusters.

However, the shape of the upper band differs in both embeddings. In the t-SNE plot, the upper band similar to the lower band has a increasing trend, meaning that pairs which are far apart in the original space tend to be farther apart in the embedding. This suggests that t-SNE retains some ordering of the between-class distances, even though the spread around the trend is wide. In the UMAP plot, the upper band is nearly flat with a slight downward trend in the end. This means that UMAP assigns approximately the same low-dimensional distance to all between-class pairs regardless of how far apart they actually are in the original space. This behaviour follows directly from UMAP's design; since between-class pairs are not connected in the high-dimensional neighbourhood graph, UMAP's objective function applies no pressure to preserve their relative ordering and simply pushes them to an equilibrium distance.

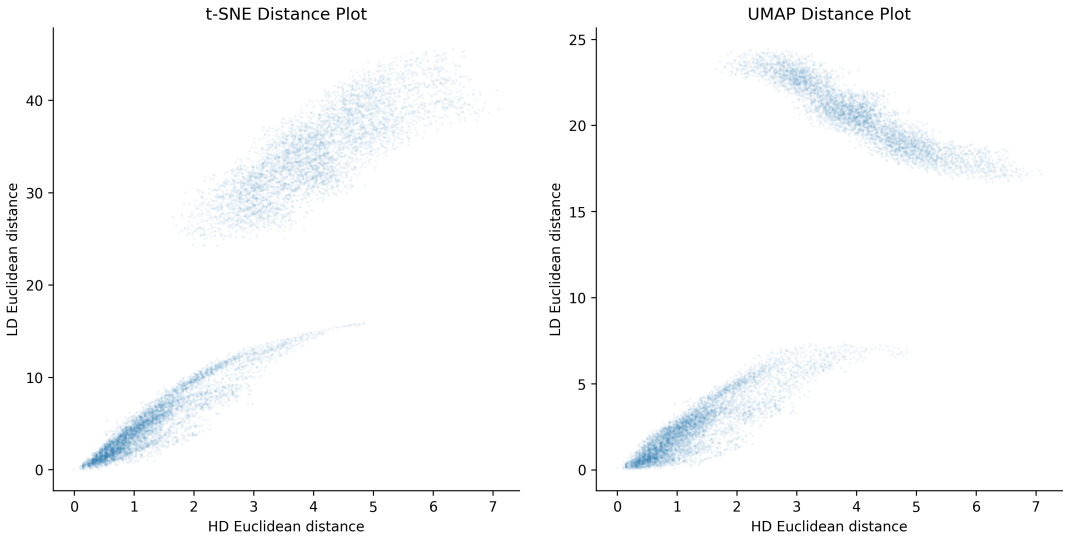
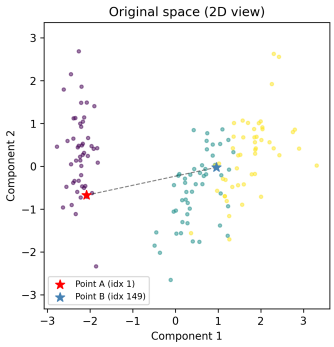
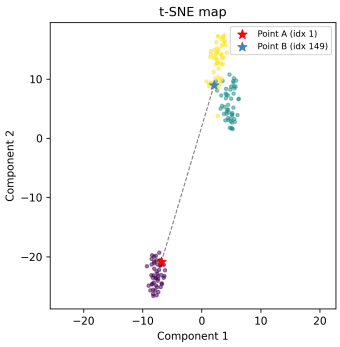


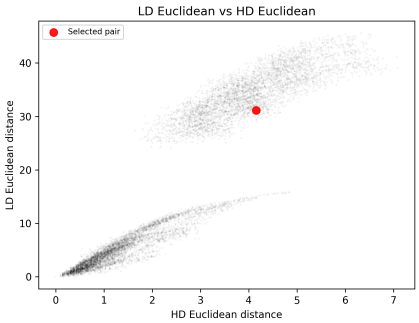
Figure 13: Distance plot of t-SNE and UMAP on the Iris dataset in Euclidean space

We can use the selected pair diagnostic tool to show the two band structure of the distance plots in Figure 14

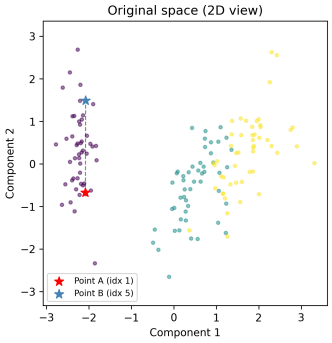


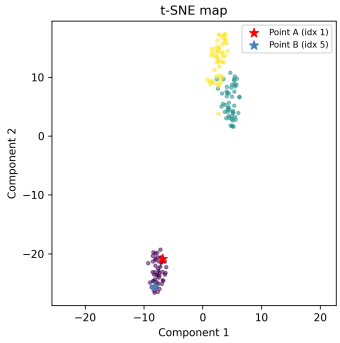


(a) Original space and t-SNE embedding, inter-class pair (idx 1 vs idx 149)

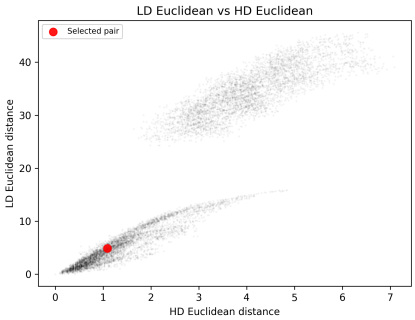


(b) Distance plot, pair lands in the upper band





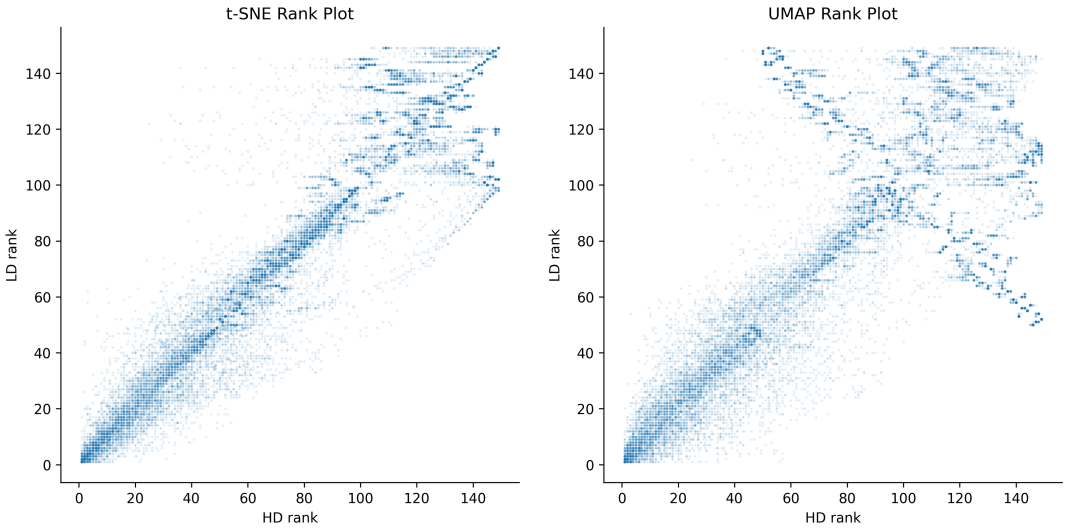
(c) Original space and t-SNE embedding, intra-class pair (idx 1 vs idx 5)



(d) Distance plot, pair lands in the lower band

Figure 14: Two selected pairs from the Iris dataset and their corresponding positions in the t-SNE distance plot. The top row shows an inter-class pair (setosa and virginica) which falls in the upper band, reflecting a large separation in both the original and embedded space. The bottom row shows an intra-class pair (both setosa) which falls in the lower band, confirming that within-class distances are small and well preserved by the embedding.

Furthermore, Figure 15 shows the rank comparison plots for both t-SNE and UMAP on the Iris dataset. Both plots display a clear diagonal trend, indicating that the neighbourhood rankings are mostly preserved after the embedding. Points that are close in the original high-dimensional space tend to remain close in the embedding, and points that are far apart tend to remain far apart. The spread around the diagonal is small for both methods, which is consistent with the distance plots shown earlier. The slight widening of the spread at higher ranks is expected behaviour for local methods such as t-SNE and UMAP, as both are designed to prioritise local neighbourhood structure over global rank ordering. Overall tank plots confirm that both methods produce well-preserved embeddings on the Iris dataset.



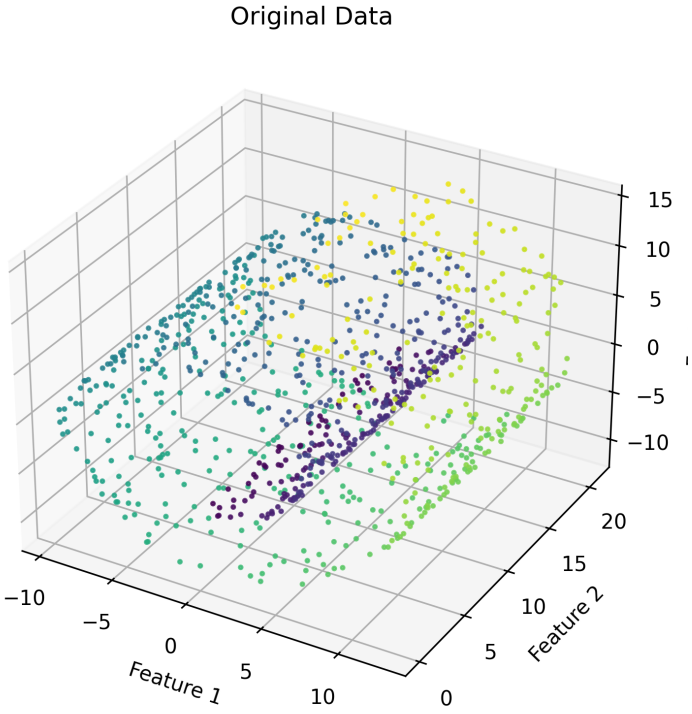
(a) Rank comparison for t-SNE (b) Rank comparison for UMAP

Figure 15: Comparison of high-dimensional and low-dimensional pairwise ranks for UMAP and t-SNE

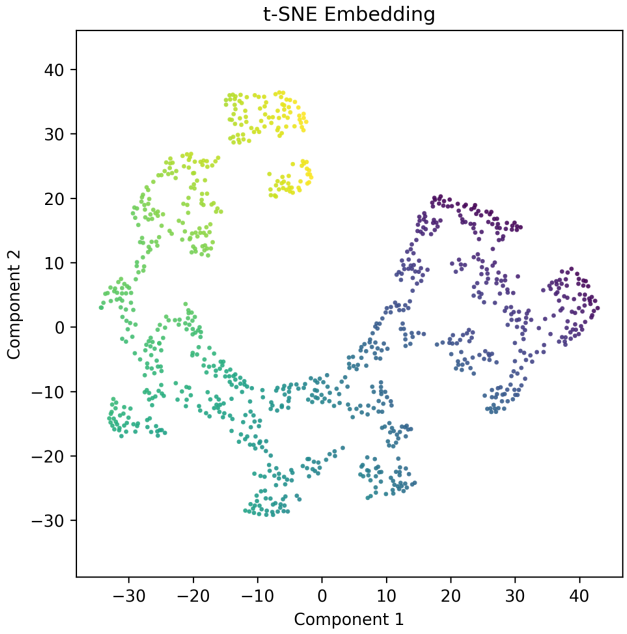
Collectively, diagnostics consistently show that both t-SNE and UMAP produce high-quality embeddings on the Iris dataset, making it a useful baseline for interpreting what well preserved plots look like before moving to more challenging datasets.

4.3 Swiss Roll Dataset

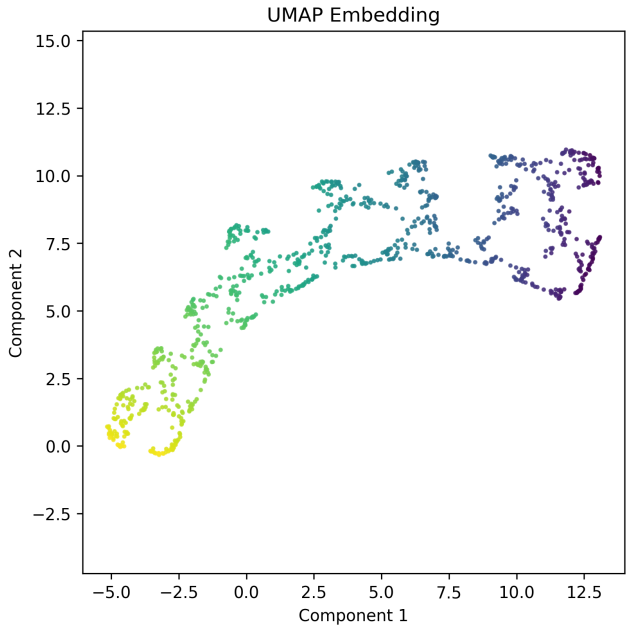
Unlike the Iris dataset, which is linearly separable and well-suited to standard Euclidean analysis, the Swiss Roll presents a fundamentally different challenge. As introduced in Section 2.3, the Swiss Roll is a three-dimensional data set in which the data points lie on a manifold shaped like a spiral. Two points on opposite layers of the roll can appear close in Euclidean Space while being fat apart when measured along the surface of the manifold. This makes the Swiss Roll an important test case for the geodesic distance diagnostic, since the Euclidean distance plot alone may suggest that distances are well preserved while the geodesic distance plot reveals a different picture. In this section we apply both t-SNE and UMAP to the Swiss Roll and use the diagnostics framework to assess how well each method respects the intrinsic geometry of the data.



(a) Original Swiss Roll data in three dimensions. Colour represents position along the manifold.



(b) t-SNE embedding. The roll is fragmented into disconnected clusters.



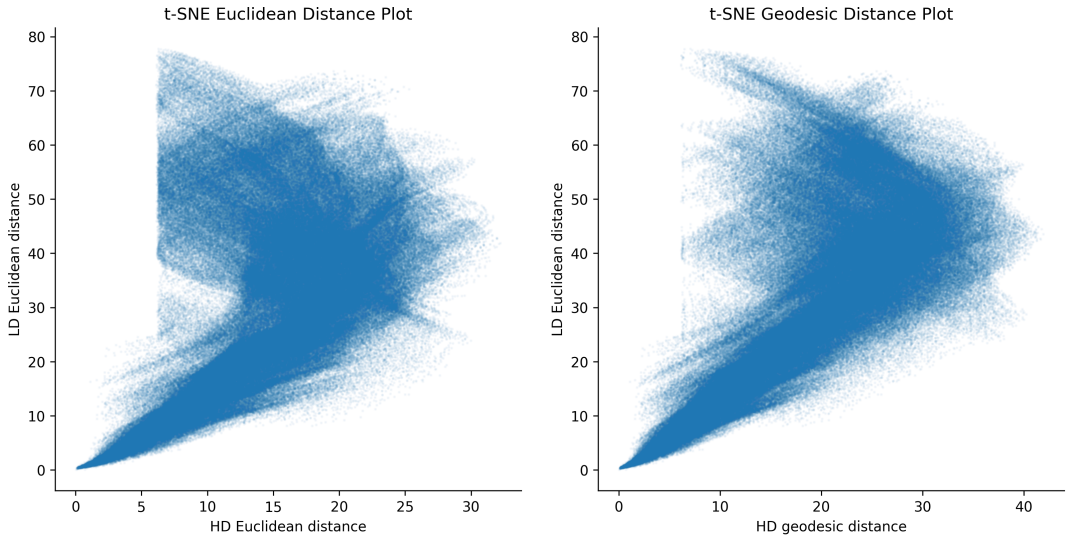
(c) UMAP embedding. The manifold is largely unrolled into a continuous curve.

Figure 16: Original data and embeddings for the Swiss Roll dataset. t-SNE fragments the manifold structure while UMAP preserves the global ordering along the roll.

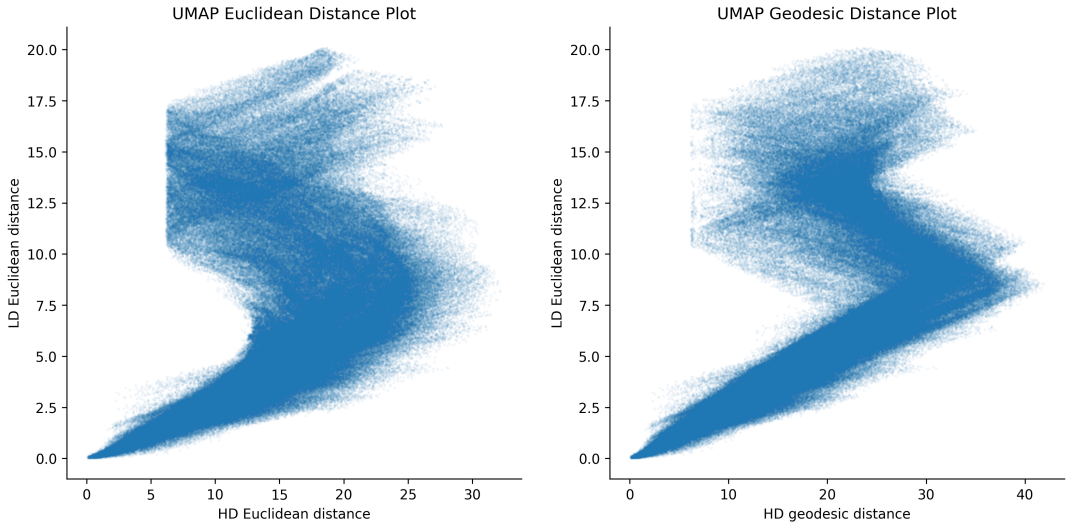
Figure 16a shows the original Swiss Roll data in three dimensions and Figures 16b and 16c show the t-SNE and UMAP embeddings respectively. After observation of the embeddings, both have unrolled the manifold but with some disconnections. Looking at the different colours of the points in different locations the embeddings have done good job at detecting the start and end points of the roll with yellow points being the outer layer of the roll and the dark purple points being the inner layer and the tail end of it. Some minor differences in the embeddings can be seen in the density of the points in the two different embeddings, with t-SNE having points more spaced and in UMAP the points are much closer to each other resulting in a more compact structure.

Figure 17 Shows the Euclidean and geodesic distance plots for both t-SNE and UMAP on the Swiss Roll. The Euclidean distance plots for both methods show a similar shape with there being an increasing trend in general. Meaning that pairs of points that are close in the original space remain close in the embedding. This is visible in the lower left region of both plots. However, the plots also show a large spread, especially for medium and large high dimensional Euclidean distances. This means that pairs with similar original Euclidean distances can be placed at very different distances in the low dimensional embedding. This behaviour is expected for the Swiss Roll, because the Euclidean distance does not fully describe the structure of the data. Some pairs that are close in Euclidean space may actually belong to different layers of the roll, and therefore should not remain close in the embedding.

The geodesic distance plot on the other hand are more informative for this dataset. Since Swiss Roll is a curved manifold, the geodesic distances better represent the distance along the structure. In the geodesic plots , the relationship between the high-dimensional and low-dimensional distances become more meaningful. The increasing trend can be seen in both methods but this trend is clearer that in the Euclidean plots. For UMAP, the geodesic distance plot shows a relatively clear increasing band. This means that as the geodesic distance between two points increases, their distance in the embedding also tends to increase. Therefore, UMAP appears to preserve the ordering of distances along the manifold better than it preserves raw Euclidean distances in the original space. However, the points do not form a thin diagonal line, so the embedding is still not distance-preserving in the strict sense. The spread around the trend shows that some pairs are compressed or expanded in the embedding.



(a) t-SNE: LD Euclidean vs HD Euclidean (b) t-SNE: LD Euclidean vs HD geodesic



(c) UMAP: LD Euclidean vs HD Euclidean (d) UMAP: LD Euclidean vs HD geodesic

Figure 17: Euclidean and geodesic distance plots for t-SNE and UMAP on the Swiss Roll dataset. The Euclidean plots show a distinctive two-regime shape caused by cross-layer pairs being pushed far apart in the embedding despite their small Euclidean distance. The geodesic plots show these same pairs spread to larger x-values, revealing that their true manifold distance was large all along.

4.4 Blobs Dataset

For the next dataset, we are going to showcase some of the limitations of t-SNE and UMAP. We will use a synthetic data set to create three distinct classes, two of which are clusters with different densities, and the third class is a singular point distant from the two clusters. Figure 18 shows the original data with a clear structure. Class 0 is a dense cluster, while class 1 is a much larger and more spread out cluster.

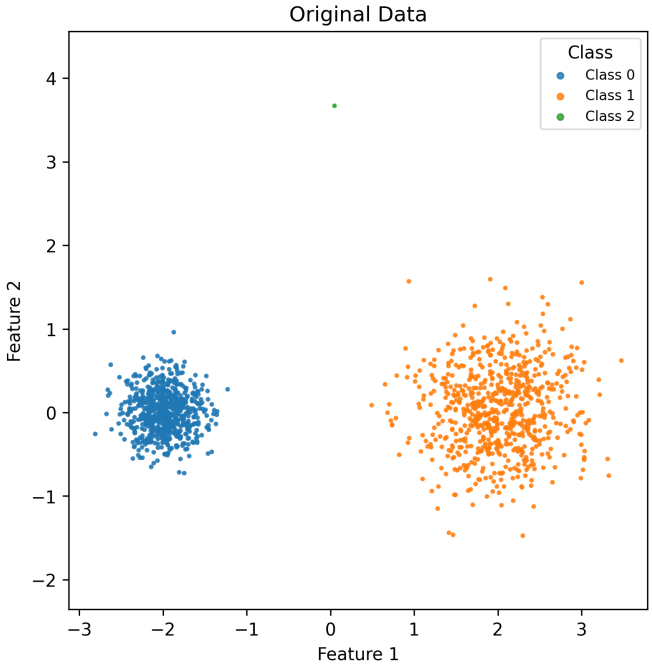
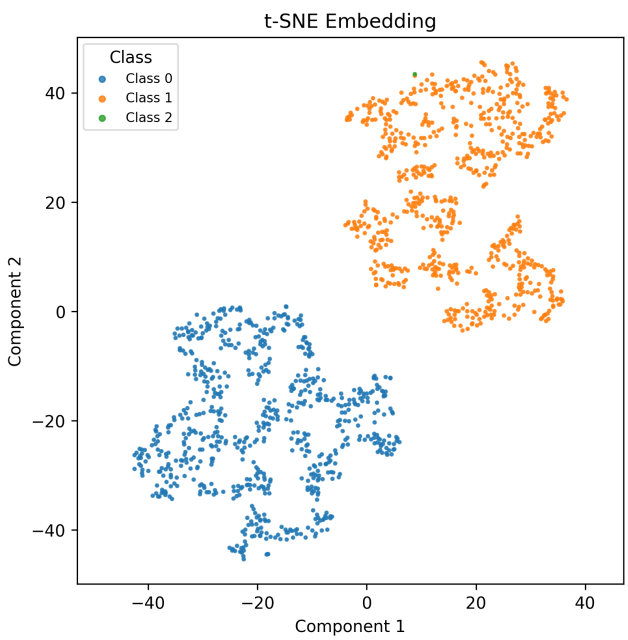
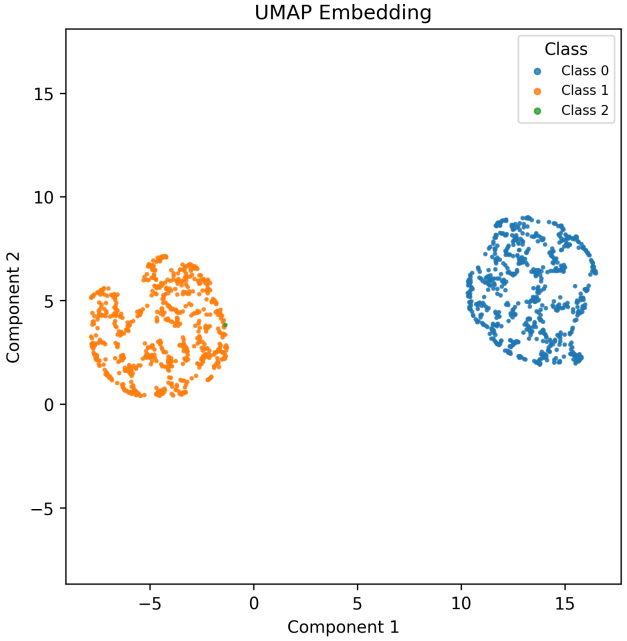


Figure 18: Original plot of the synthetic 3 class dataset

In both embeddings as shown in Figure 19, this difference in density is vanished. The two main clusters appear at a similar scale in the embeddings, and both have been fragmented into sub-clusters that do not correspond to any real structure in the original data. This behaviour is a direct consequence of the local scaling mechanism mentioned in Section 3.3. Because both t-SNE and UMAP normalise the neighbourhood size separately for each point, the points in the dense region have a much smaller kernel and sparser regions have a larger kernel which causes all parts of the data to appear at a similar local scale in the embedding regardless of their true density. The behaviour of the single outlier in Class 2 further illustrates this limitation. In the original data, the outlier is clearly isolated from both clusters. In both embeddings, the outlier is pulled into Class 1, the less dense cluster. This again can be attributed to the points in Class 1 having larger kernel size resulting in a stronger pull on the outlier. Nevertheless, the point has lost its characteristic completely. A user inspecting either embedding without the original data would have no reason to suspect that this point is an outlier.



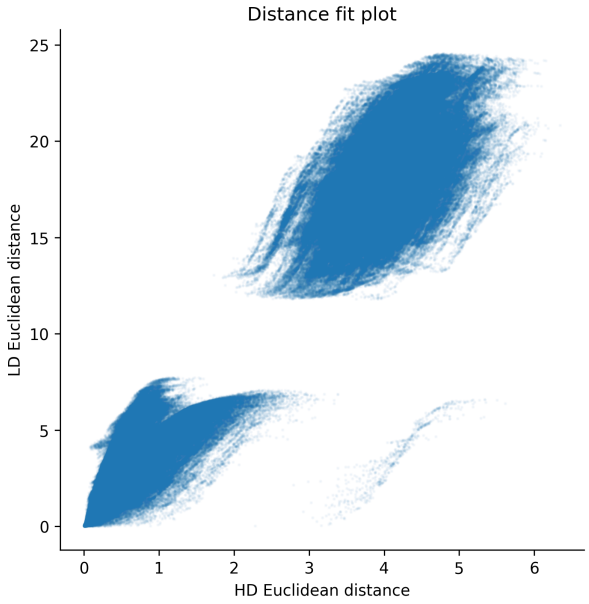
(a) t-SNE embedding. Both clusters are expanded to a similar size and the outlier is placed at the edge of Class 1.



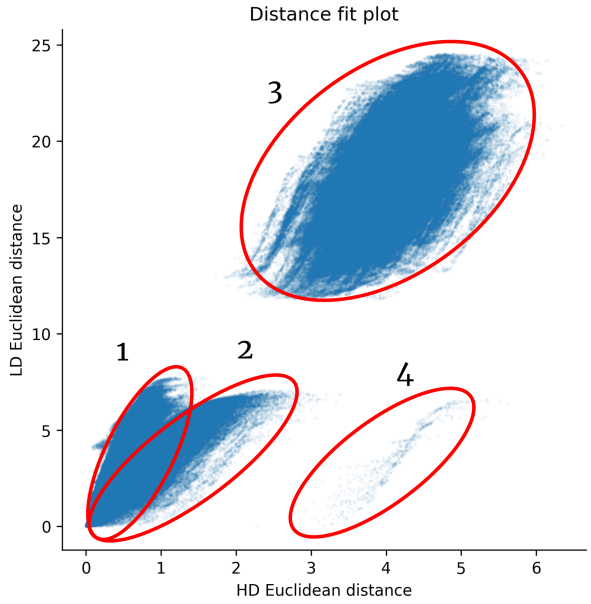
(b) UMAP embedding. The density relationship is inverted and the outlier is absorbed inside Class 1.

Figure 19: t-SNE and UMAP embeddings of the synthetic three-class dataset. Both methods equalise the cluster densities and fail to preserve the isolated character of the outlier.

The Iris dataset gave us an example of how disjoint clusters result in disjoint distance clusters in the distance plots. This result is also expected from the synthetic dataset that has been made here.In Figure 20 we have the distance fit plot for the dataset. Furthermore, We have highlighted 4 different regions in the distance plot of the UMAP embedding shown in Figure 20b in order to dissect the plot.



(a) Distance fit plot for the Blobs dataset.



(b) Highlighted version of the same plot. Region 1: dense cluster within-pairs. Region 2: sparse cluster within-pairs. Region 3: between-class pairs. Region 4: outlier-to-cluster pairs absorbed in the embedding.

Figure 20: Euclidean distance fit plot for the Blobs dataset. The four regions reflect the density structure of the data and expose the outlier absorption effect described in the text.

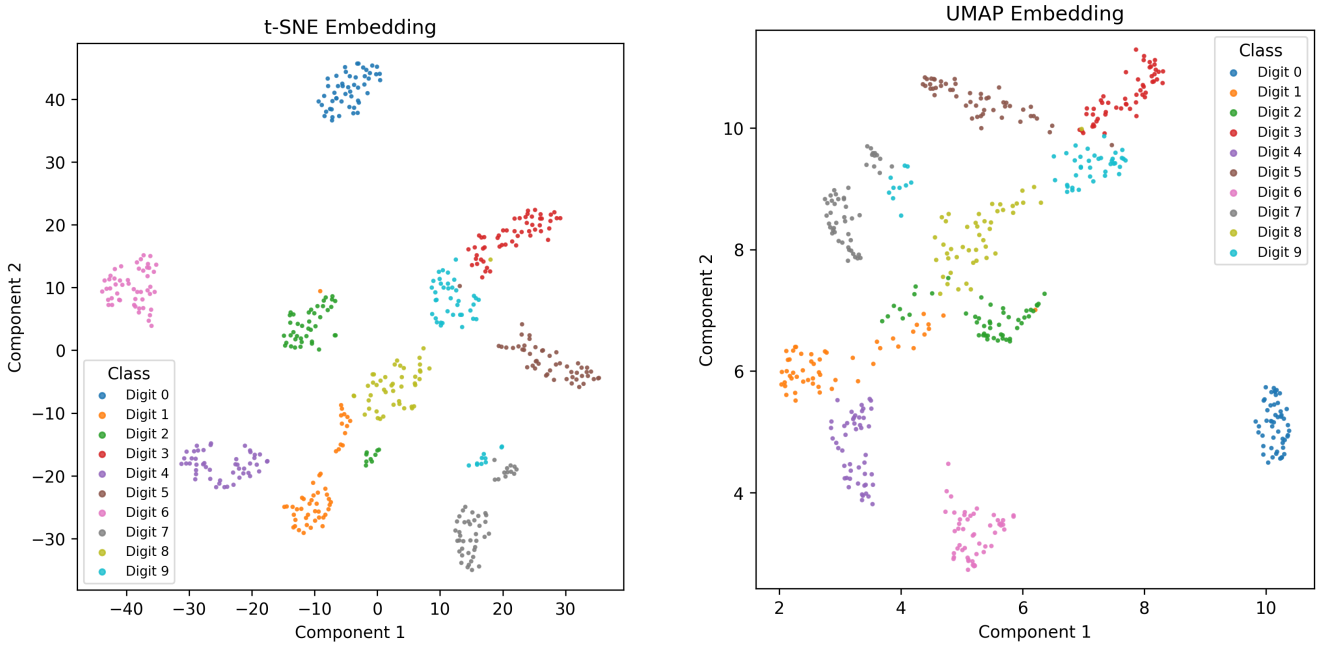
Region 1 showcases the within class pairs from the dense cluster (Class 0). These points have very small HD distances because the cluster is compact. However, their LD distances have a wider range (0-8) and this is again due to the density normalisation of t-SNE and UMAP. So if the distances in the HD is small they will be pushed apart in the embedding which gives shape to the S-curve shape of the cluster. In Region 2, we have the within class pairs from the sparse cluster (Class 1). These points are naturally more spread out, giving larger HD distances and because the scaling factor applies larger kernel to the sparse region, their LD distances end up in the same range as Region 1. Region 3 is the between class pairs (Class 0 vs Class 1). These pairs are far apart in both HD and LD space, forming the upper band. The two clusters are globally separated, and the embedding correctly reflects this. Lastly and more interestingly, Region 4 showcases the outlier to cluster pairs. Since it is being pulled in the sparse cluster (Class 1) we can infer that this region corresponds to pairs from Class 1 vs Class 2. The single outlier has large HD distances to the sparse cluster, but after being pulled in the LD distances are quite small. This region is the distance plot signature of the outlier absorption we observed in the embedding.

4.5 Digits Dataset

The previous experiments used either low-dimensional artificial data or small tabular datasets. For the next dataset we have chosen the Digits dataset which contain handwritten digits. This dataset contains 1797 samples of handwritten digits (0-9), each represented as an 8×8 greyscale pixel

image. Flattening each image produces a 64-dimensional feature vector, making this the highest-dimensional dataset in our experiments. With ten classes and a 64 dimensions to 2 dimensions reduction, the diagnostics face a considerable harder task than the previous datasets.

We chose this dataset for two reasons. First, it is a real-world dataset, the structure in the data comes from handwriting variation than from a designed formula, so any clustering or separation that emerges in the embedding reflects the actual patterns in the pixel intensities. Second reason is that we can visually inspect the images of the digits. This way if there are distortions in the embeddings, we can select the points and inspect the digits to see whether the results make sense. This makes the digits dataset a natural test case for the selected pair diagnostics. Initially, we take a look at the embeddings to see whether the methods can separate the classes.



(a) t-SNE embedding on Digits dataset.

(b) UMAP embedding on Digits dataset

Figure 21: t-SNE and UMAP embeddings of the 64-dimensional digits dataset. Both methods recover all ten digit classes as distinct clusters from pixel-space data, confirming that the class structure is captured in the low-dimensional representation.

Figures 21 show the embeddings for t-SNE and UMAP. Both methods have separated the digit 0 in a cluster far from other classes. t-SNE has created more distinct clusters with the digits 9 and 3 being quite close to each other. In the UMAP embeddings, however, the clusters are not as distinct and their boundaries are less clear compared to t-SNE's embedding. The digit 0 remains the most isolated cluster which is synonymous between both embeddings. The close boundaries of the clusters raise a question. Are these points similar in the 64-dimensional pixel space or are they an artefact of the embedding? In order to investigate this, we use the selected-pair diagnostics to inspect a pair of points. If we take a closer look at the t-SNE embedding, we see a point from class 5 (brown color) located in the cyan cluster of class 9. This observation gives us a good reason to inspect the pair closer and see whether the embedding has made an error or it is somehow justified.

Figure 22 shows the selected pair dashboard for point 5 (digit 5) and point 149 (digit 9). In the

t-SNE map, the digit labelled 5 is placed inside the cluster which belongs to digit 9. At first glance, this may appear to be a misplacement. However, the distance-fit plots shows the HD Euclidean distance between the two points is small relative to the rest of the dataset. The similarity plot also tells us that the both P and Q scores are non-negligible.

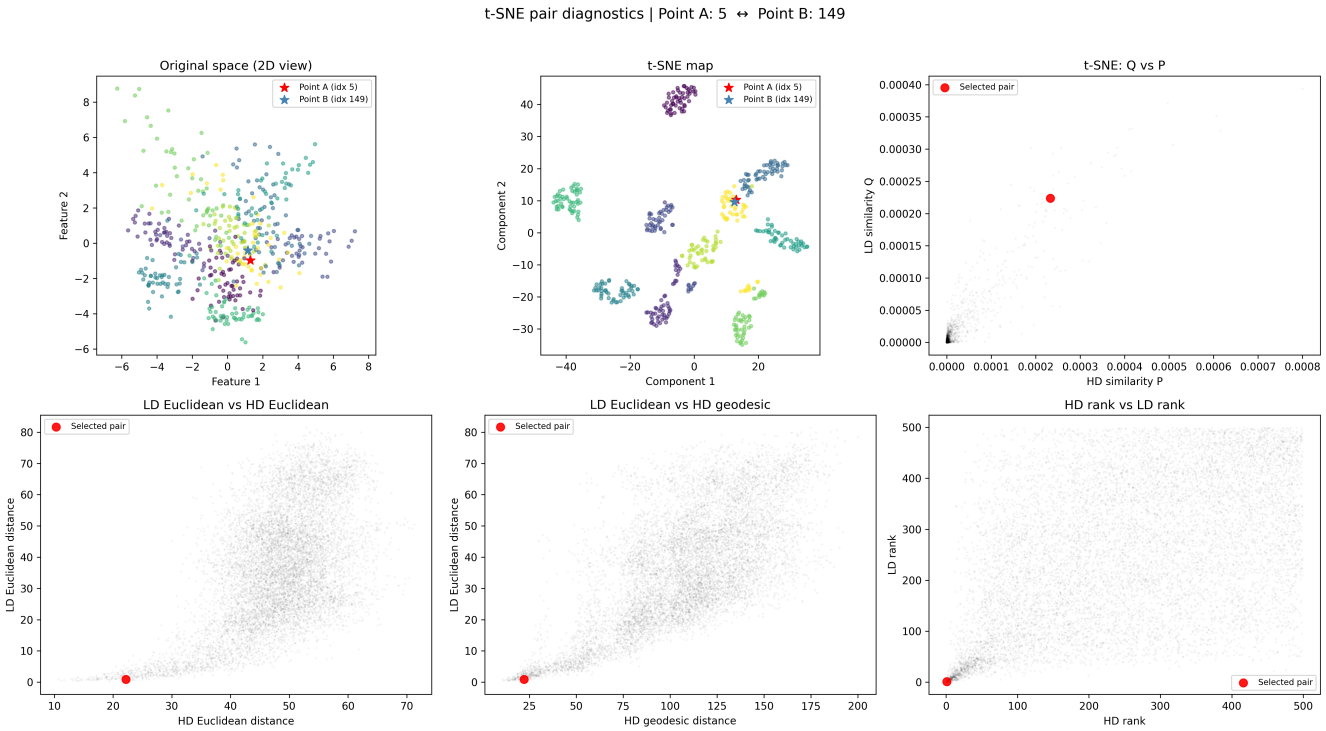


Figure 22: The selected pair dashboard for the two points that have different label but belonged in the same cluster after the embedding

Looking at the actual pixel images shown in Figure 23, the reason for this instance becomes clear; this particular handwritten 5 looks exactly the same as the handwritten 9. Thus, the embedding is not making an error, it is in fact reflecting a real ambiguity in the data. This demonstrates the importance of pair selection diagnostics.

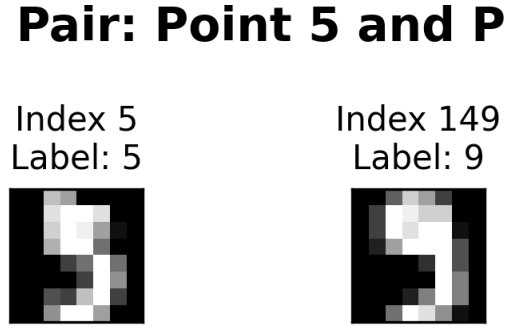


Figure 23: Actual pixel images of the selected pair. Both look quite similar to each other which makes their grouping in the embedding logical.

Conversely, we can find another pair of points that perhaps reflect an error of the embedding. We can find such a pair by looking at the ranking of the points. We need to find a pair of points that are in the same class and close in the HD rank and far in the LD rank. By setting these limits we can find a pair. Figure 24 shows the dashboard for the point 224 and 378 both labelled as digit 8. They have been placed far from each other in the embedding. The rank plot shows that their HD rank is very small meaning they are among each other's nearest neighbours in the 64-dimensional space, yet their LD rank is approximately 240, placing them far apart in the embedding. The distance fit plots show the pair above the point cloud, indicating the LD distance is larger than exp given the HD distance. Inspecting the pair's pixel images in Figure 25 they both look quite similar and the most differentiating point would be the brightness of the pixels. Maybe t-SNE has treated this brightness as a class boundary but it can not be said with utter confidence.

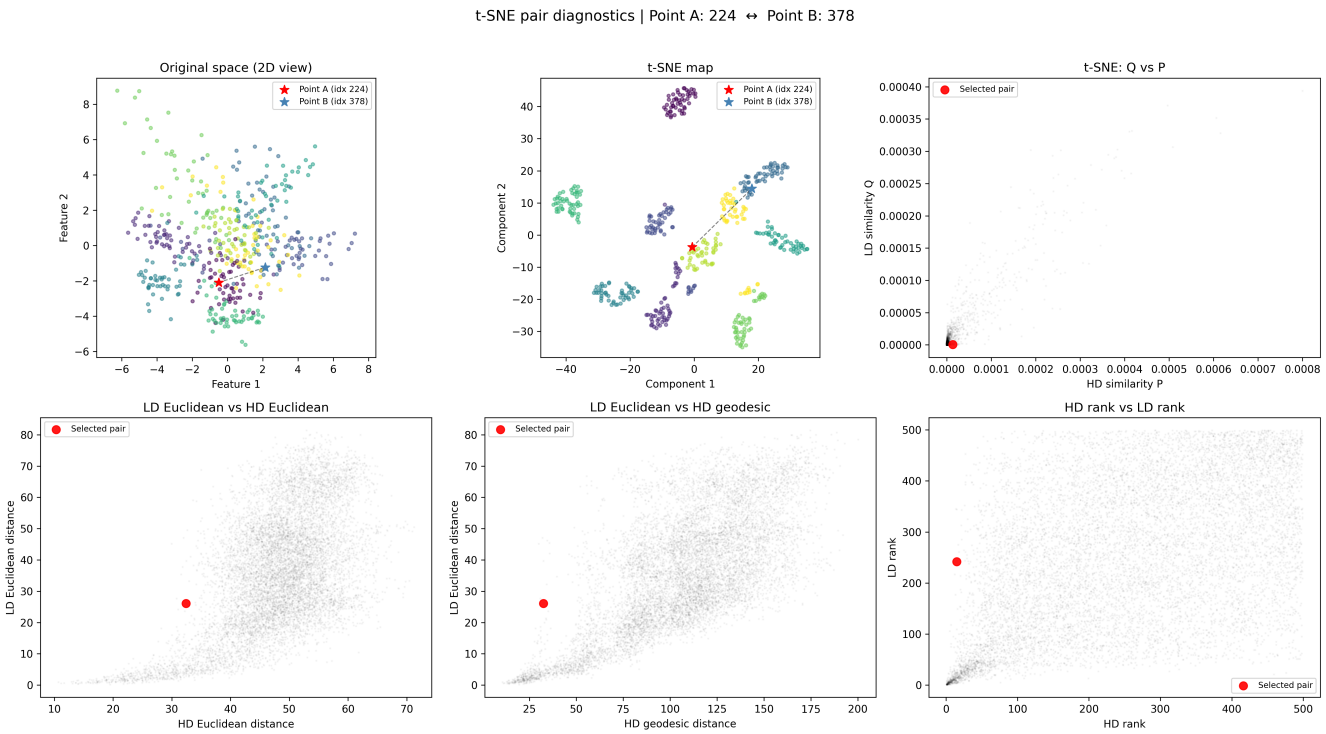


Figure 24: Original plot of the synthetic 3 class dataset

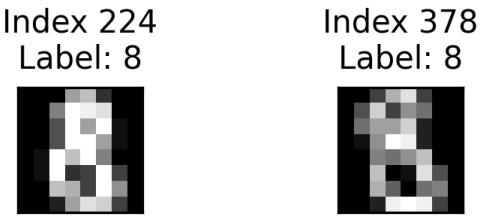


Figure 25: Actual pixel images of the selected pair. Both look quite similar to each other which makes their grouping in the embedding logical.

5 Conclusions and Further Research

This thesis set out to extend an existing diagnostics toolkit for both t-SNE and UMAP [Hai25] into a more accessible package, while adding new diagnostic tools that extend the scope of evaluation beyond what the internal objectives of the methods can reveal. The resulting framework, implemented in the `dr-diagnostics` package, combines global and local visualisations that together give a more thorough picture of embedding quality. The Package can be found at: https://pypi.org/project/dr-diagnostics/, it also includes a tutorial notebook which will guide the interested practitioners on how to use the package.

The global diagnostic framework includes distance-fit plots to visualise how pairwise distances in the original space relate to distances in the embedding. By offering both Euclidean and geodesic variants, the framework allows the user to distinguish between distortions caused by ignoring curved manifold structure and those that occur even when intrinsic geometry is accounted for. Similarity plots expose the internal mechanism each method uses: P-Q plot for t-SNE and V-W plot for UMAP. Because these methods define similarity differently, the plots have different shapes and should be interpreted within each method's own framework rather than compared directly. Rank-based diagnostics adds another layer for evaluation of the quality of the embedding which can be used to see whether false neighbours are introduced or not.

In addition to these global tools, the framework includes two local diagnostics. The selected-point diagnostic highlights the k-nearest neighbours of a chosen point across all six diagnostics panels, allowing the user to determine whether a specific point's neighbourhood is preserved. The selected-pair diagnostics examine a pair of points in detail, and with a supplementary function that ranks all pairs by their residual from a linear regression of low-dimensional vs high-dimensional distances, making it easier to spot the most distorted pairs for further inspection.

The experiments on four datasets demonstrate how these diagnostics can be used in practice. The goal of this is to perhaps help the practitioners to look at the data and the different plots and spot certain behaviours. For example the two band structure of the Iris data set with the bands corresponding to inter and intra cluster pairwise distances. The different datasets were chosen to highlight different concepts. The goal of the Iris dataset was to set a baseline as it is linear and fairly popular. The Swiss roll dataset highlighted the importance of incorporating geodesic distances beside Euclidean distances. This is especially useful if the structure of the manifold is

known. The synthetic blobs dataset was chosen to expose a fundamental limitation shared by both methods since both methods normalise neighbourhood sizes locally, they cannot preserve differences in cluster density. The dense and sparse clusters were equalised in scale, and the isolated outlier was absorbed into the less dense cluster. A practitioner inspecting either embedding without the original data would have no indication that the outlier exists. Finally, on the Digit dataset, the selected-pair diagnostics demonstrated two complementary uses. In the first case, a handwritten 5 that was placed inside the cluster of digit 9 turned out to be visually identical to a . The embedding did not make an error but was correctly reflecting the ambiguous nature of the original data. In the second case, two samples of digit 8 that were close neighbours in the original data were placed far apart in the embedding, suggesting a distortion caused by the embedding.

Taken together, the experiments confirm that internal loss values are not sufficient for evaluating embedding quality. The diagnostics implemented in this framework reveal structure that objective values hide, help differentiate justified placements from real distortions, and allow comparison between t-SNE and UMAP on equal terms.

5.1 Limitations

The package by no means is perfect has several limitations that should be acknowledged. The computational cost scales with the number of pairwise distances, which is $O(n^2)$. For larger datasets with more than a few thousand points will result high running times. The current implementation does not include any approximation or subsampling strategy. Moreover, geodesic distance estimation depends on the construction of a k-nearest neighbour graph. If k is too small, the graph may be disconnected and geodesic distances will be unreliable. If k is too large, shortcuts across the manifold are introduced. The choice of k is therefore important. Third, the most distorted pair utility fits a linear regression through all pairwise distances and ranks pairs by absolute residual. This is a simple and transparent criterion, but it may assign high priority to structurally uninteresting pairs, especially when the global distance-distance relationship is not well described by a line. Finally, the package only supports t-SNE and UMAP and their similarity plots. Extending these to other methods require adapting the similarity definitions accordingly.

5.2 Further Research

There are several directions in which this work can be extended. The most immediate one is scalability. The current implementation computes all pairwise distances, which becomes expensive as the dataset size grows. Using approximate nearest-neighbour methods or subsampling strategies would make the diagnostics applicable to larger data without changing the underlying approach.

Another direction concerns usability. At the moment, point and pair selection requires the user to supply numerical indices. This is not very practical, because it assumes the user already knows which index corresponds to which observation. An interactive interface, where the user clicks directly on the embedding to select a point or pair, would make the local diagnostics considerably more accessible.

Finally, the framework currently focuses on t-SNE and UMAP. Extending it to other dimensionality reduction methods such as ISOMAP or PaCMAP would require adapting the similarity diagnostics accordingly, but would make the package more generally applicable. A hyperparameter sensitivity panel that shows how the diagnostics change across different perplexity or neighbourhood

size settings would also be a useful addition, as practitioners often choose these values without a clear criterion.

References

[AW10] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[CG15] John P Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research*, 16(1):2859–2900, 2015.

[Fis36] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[GvdV04] Patrick Groenen and Michel van de Velden. Multidimensional scaling. Technical report, 2004.

[Hai25] Dominika Haik. A diagnostic framework for t-sne and umap: Improving interpretability, July 2025. Supervisor: Dr. S. M. H. Huisman.

[Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[Kru64] Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[KW78] Joseph B Kruskal and Myron Wish. *Multidimensional scaling*. Number 11. Sage, 1978.

[LV08] John A. Lee and Michel Verleysen. Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods. In *Proceedings of ESANN*, 2008.

[LV09] John A Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7-9):1431–1443, 2009.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[MHM18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[TSL00] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[VK01] Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks – ICANN 2001*, pages 485–491, Berlin, Heidelberg, 2001. Springer.

[VK06] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6–7):889–899, 2006.

[Wan25] Aasim Ayaz Wani. Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions. *PeerJ Computer Science*, 11:e3025, 2025.