Universiteit Leiden
The Netherlands

Data Science & Artificial Intelligence

A Diagnostic Framework for t-SNE and UMAP:

Improving Interpretability

Dominika Haik

Supervisor:
Dr. S.M.H. Huisman

BACHELOR THESIS

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

01/07/2025

Abstract

Dimensionality reduction techniques like t-SNE and UMAP are widely used to visualize high-dimensional data, yet assessing the quality of the resulting embeddings remains a challenge. In this paper, we propose a set of diagnostic plots designed to evaluate how well such techniques preserve the underlying structure of the original dataset. We introduce and implement a distance fit plot, a matrix fit plot, and heatmaps as tools for analyzing the embeddings produced by t-SNE and UMAP. Additionally, we present a method for identifying potential outliers whose embedded positions may misrepresent their original relationships. These diagnostics are demonstrated on synthetic datasets with known structure, allowing us to illustrate the effects of hyperparameter choices and the benefits of using multiple plots for joint analysis. Our results highlight the importance of interpretability in evaluating dimensionality reduction outcomes and provide practical tools for deeper insight into embedding quality.

Contents

1 Introduction

When faced with growing datasets with hundreds or thousands of features, it is easy to feel lost. Dimension reduction techniques offer a tempting solution to make sense of the chaos. Nowadays, datasets used in fields like computer science[KHO25], biology[SWM93], or business[MTB22] often have many features, commonly referred to as dimensions, which can make them difficult to analyse. Spotting meaningful structures in such datasets is challenging, and even the application of machine learning techniques is limited by computational constraints. Dimensionality reduction solves the problem by creating a smaller set of distilled features that are intended to preserve the information in the original data. These techniques are widely used as both visualisation tools and pre-processing steps in machine learning as they are capable of reducing noise, improving interpretability and revealing patterns.

t-Distributed Stochastic Neighbour Embedding (t-SNE)[VdMH08] and Uniform Manifold Approximation and Projection (UMAP)[MHM18] are dimensionality reduction techniques that emerged relatively recently, but have already gained popularity. In contrast to other popular techniques like Principal Component Analysis (PCA), t-SNE and UMAP are nonlinear and are capable of projecting structures that cannot be captured by linear decomposition. On a conceptual level, both methods model pairwise similarities between points in the high-dimensional dataset and its corresponding low-dimensional embedding. The similarities can be interpreted as the probability that a data point would choose another point as its neighbour. The methods then use gradient descent to minimise a cost function that captures the difference between high- and low-dimensional similarities. Both t-SNE and UMAP focus on preserving the local structure of the data, but UMAP claims to accurately preserve the global structure as well.

However, the theoretical foundations behind these methods differ. t-SNE defines the pairwise similarities based on probabilities with Gaussian and Student's distributions for the high-dimensional and embedded data, respectively. An asymmetric cost function, the Kullback–Leibler divergence, is used to measure the difference between the two representations. On the other hand, UMAP uses an exponential distribution to calculate high-dimensional similarities. Theoretically, it is based on Riemann geometry and attempts to construct a topological representation of the high-dimensional data. Here, another intuition behind the similarities is that they can be understood as the probability that, in the high-dimensional graph, there exists an edge between two points. Afterwards, cross entropy is the cost function used to minimise the difference between the original dataset and the low-dimensional embedding.

While some evaluation methods for dimensionality reduction techniques exist[LV09a][Roy24], there are no widely accessible or standardised tools for assessing how well an embedding reflects the original data structure. Standardised diagnostic tools are necessary to judge how well information is preserved after dimensionality reduction. An evaluation method is required for deciding whether a dimensionality reduction technique is suitable for a given dataset and tuning hyperparameters. Considering how dimensionality reduction is often used to search for patterns in the data, the knowledge about its behaviour could prevent a researcher from drawing incorrect conclusions or misinterpreting noise for structure.

An example of a diagnostic plot already used in dimensionality reduction, in this case PCA, is

a Scree plot[Kan05] (also known as the elbow method). The plot shows the amount of variance preserved by each component and thus indicates how many principal components are necessary to preserve a required portion of information. Another diagnostic plot is the Shepard diagram[GvdV04], commonly used for the evaluation of Multidimensional Scaling[Mea92]. A Shepard diagram plots the pairwise distances in the high-dimensional space against the distances in the reduced space, along with the transformations of the distances. Points located close to a diagonal indicate that the relative distances are well-preserved. In this paper, a version of the Shepard diagram will be later introduced as a part of our diagnostic toolkit. Moreover, there are promising rank-based diagnostics for t-SNE and UMAP, which compare the ordering of nearest neighbours for each data point and thus assess how well the relative relationships among points are preserved after dimensionality reduction[LV09b][LMBH11].

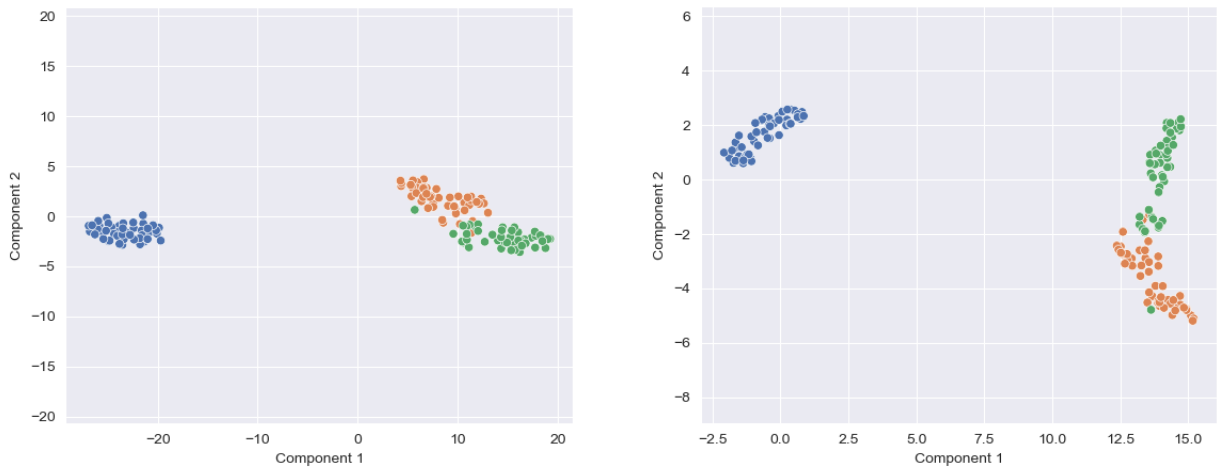
Although some diagnostics are available for t-SNE and UMAP, the need remains for more robust and interpretable measures. The approach suggested by the author of t-SNE[vdM], which recommends visually inspecting the result, relies on subjective interpretation rather than systematic evaluation. Importantly, the appearance of an embedding alone cannot be taken as a definitive measure of its quality[WVJ16]. Another limitation is that comparing the cost values for different runs of the algorithm provides only limited insight into how well the underlying structure of the data is preserved.

This paper presents a set of diagnostic tools for t-SNE and UMAP to fill this gap. Some methods have previously been proposed[Dui24][Wu23], but using them would often require retracing the researchers' steps and writing the code from scratch. We implemented these techniques into a user-friendly framework, hoping to improve research which relies on t-SNE and UMAP. The tools included in our Python package provide point-level insight into the quality of the embedding and show broader trends in the behaviour of the dimensionality reduction technique.

The remainder of the paper is structured as follows. In Section 2, we introduce the implemented diagnostic tools, including similarity matrix heatmaps, matrix fit plot, distance fit plot and an outlier detection measure. Section 3 presents a series of experiments on synthetic datasets with known structure being evaluated with the proposed diagnostics. Section 5 discusses key findings, limitations of the current approach, design decisions and implications for the evaluation of dimensionality reduction techniques. This work is a bachelor thesis completed at LIACS, Leiden University, under the supervision of Dr. Sjoerd Huisman.

2 Methods

In this section, we introduce a set of diagnostic plots designed to evaluate the quality of dimensionality reduction techniques. These include the distance fit plot, the matrix fit plot, and similarity matrix heatmaps. Together, they provide insights into how well local and global structures are preserved in the low-dimensional embedding. Examples of t-SNE and UMAP embeddings are presented in Figure 1.



(a) Embedding by t-SNE.

(b) Embedding by UMAP.

Figure 1: Embeddings of a high-dimensional Iris dataset (see Section 2.6) into two dimensions. Points are coloured according to the original data labels.

2.1 Similarity Matrices

Similarity matrices encode pairwise similarities between points in a dataset, reflecting how likely a point would choose another as its neighbour. The matrices are needed for all diagnostics that rely on the internal metric of the dimensionality reduction method. In this paper, we include several diagnostic measures that use the similarity matrices to draw conclusions about the dimensionality reduction. For instance, by comparing high-dimensional similarities and low-dimensional similarities, we can understand how well the dimensionality reduction algorithm preserves the positions of points in the original space. To aid the analysis of the similarity matrices, our package includes a function that generates a matrix heatmap so that one can visually analyse the plots to assess how well the patterns are preserved. Additionally, since the data points in the matrices are sorted based on their similarity, the heatmaps can be used to spot groups in the data and verify that they are reflected in the embedding. This section introduces the similarity matrices for t-SNE and UMAP, shows how the similarities are calculated, and presents examples of the resulting heatmaps.

2.1.1 t-SNE

In t-SNE, the conditional probability $p_{j|i}$, which is the similarity of data point x_j to x_i, is defined as follows:

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

where $||x_i - x_j||$ is the Euclidean distance between x_i and x_j, and σ_i is the standard deviation of a Gaussian distribution centred at the point x_i. The value of σ_i is found through binary search for the user-specified perplexity value, which we discuss in detail later in this section. We set $p_{i|i} = 0$ to ignore the similarity of a point x_i to itself. Given this, the probability $p_{j|i}$ that the point x_i

would pick x_j as its neighbour is proportional to a Gaussian distribution with standard deviation σ_i centred at the point x_i. For points that are close to x_i, the probability is high, and for points that are relatively far away, the probability is negligible. Resource-efficient implementations of t-SNE take advantage of this by skipping the computation of similarities for distant points.

A key hyperparameter in t-SNE is the perplexity, which is a smooth measure of the number of neighbours a data point has. It can be interpreted as the user's estimate of the number of close neighbours per point. It also controls the trade-off between the preservation of local and global structures in the data. As mentioned earlier, σ_i is computed through binary search for the desired value of perplexity:

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

where $H(P_i)$ is the Shannon entropy:

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

After the conditional probabilities are found, we can define the symmetric $n \times n$ P matrix, which entries are

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

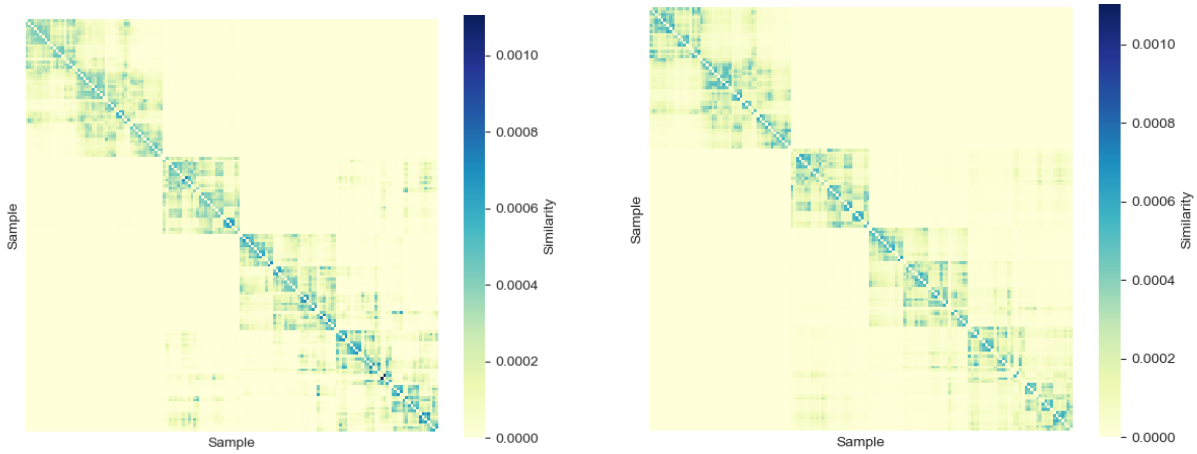
where n is the number of instances in the dataset. Since the conditional probabilities are asymmetric (usually, $p_{j|i} \neq p_{i|j}$), this step symmetrises and normalises the similarity matrix.

To mirror the P matrix in the low-dimensional space, the $n \times n$ Q matrix contains the similarities of the points in the embedded dataset, where point y_i represents the embedded point x_i. The low-dimensional similarities are defined as follows:

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

where the probability q_{ij} captures the similarity between points y_i and y_j and is proportional to a Student's t-distribution with one degree of freedom. The Student's t-distribution was introduced in t-SNE as an improvement over Stochastic Neighbour Embedding (SNE) to solve the crowding problem. The undesired crowding occurs when multiple equidistant points in high-dimensional space are mapped to a low-dimensional space, and, since it is impossible to accurately represent this, the points are crowded into one spot. The Student's t-distribution has much heavier tails than the Gaussian, so a moderate distance between points in the original data can be mapped to a greater distance in the low-dimensional embedding.

Figure 2 shows an example of P and Q heatmaps. The matrices are ordered according to the leaf sequence of hierarchical clustering of the high-dimensional similarity matrix to emphasize any clusters that may emerge.



(a) P matrix heatmap.

(b) Q matrix heatmap.

Figure 2: Heatmaps of the t-SNE's similarity matrices for the Iris dataset (see Section 2.6).

2.1.2 UMAP

UMAP is similar to t-SNE in terms of the similarity matrices, but it only calculates the high-dimensional similarities for the k nearest neighbours of a point. The hyperparameter k is similar to perplexity in t-SNE, as it scales the distribution of the similarity values according to the local density; however, it also determines the pairs of points for which similarity is calculated. UMAP defines the conditional probability $v_{j|i}$, which captures the similarity of point x_j to x_i, as follows:

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

where $1 \leq j \leq k$ and x_{ij} is the j-th closest neighbour of x_i. The hyperparameter k balances the attention to local vs. global structure in the data and can be interpreted as a guess about the number of close neighbours per data point. Moreover, the map $d(a, b)$ is any dissimilarity measure, most often Euclidean distance between a and b. For each x_i, we define the distance to the nearest neighbour ρ_i as follows:

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

such that $d(x_i, x_{ij}) > 0$. The purpose of ρ_i is to ensure that point x_i has at least one neighbour with similarity 1. Finally, the scale parameter τ_i, defined for each point x_i, is found through a binary search for a value such that the following holds:

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_{ij}) - \rho_i)}{\tau_i}\right) = \log_2 k.$$

It is important to note here that, in UMAP, the similarities are calculated only for k nearest neighbours. For other points, the similarity is 0. Also, $v_{i|i} = 0$ because we are not interested in self-similarity. Additionally, the implementation of UMAP [MHM18] uses an approximation algorithm for finding the nearest neighbours, hence the $\max(0, d(x_i, x_{ij}) - \rho_i)$ term that ensures the similarity values stay below or equal to 1.

Once the conditional probabilities are calculated, we can define the symmetric similarity matrix V by

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

where A is a square matrix with entries $A_{ij} = v_{j|i}$ and $\circ$ denotes the entry-wise product. This step symmetrises the similarity matrix since $v_{j|i} \neq v_{i|j}$.

Reflecting the structure of the V matrix in the low-dimensional space, the square W matrix contains pair-wise similarities that can be defined in two ways. One can either use the true function or its smooth approximation, which is currently implemented in the UMAP algorithm to facilitate the gradient descent in later stages. In our package, the default option is the true function because it is loyal to the theoretical foundations of UMAP and thus, we believe, provides more accurate information about the embedding. However, the user can opt to use the approximation instead. The true function $\Psi : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ is defined as

$$\Psi(y_i, y_j) = \begin{cases} 1 & \text{if } ||y_i - y_j||_2 \leq min\_dist \\ \exp\left(-(||y_i - y_j||_2 - min\_dist)\right) & \text{otherwise} \end{cases}$$

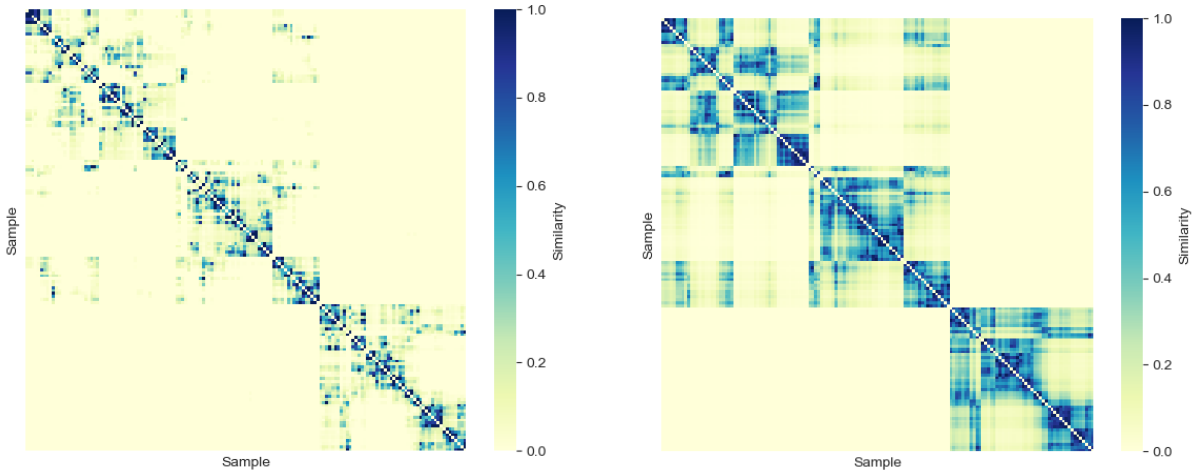
where $||\cdot||$ denotes the Euclidean distance and min_dist is a parameter that controls how tightly packed the data is in the embedding. Notably, min_dist is considered an aesthetic parameter by the authors of UMAP. If the true function is used, the (i, j) entry of the W matrix is given by $W_{ij} = \Psi(y_i, y_j)$.

The true function Ψ can be approximated by the function $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ defined as

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

where a and b are the result of the non-linear least squares fitting against the original function Ψ. If the approximation is used, the (i, j) entry of the W matrix is given by $W_{ij} = \Phi(y_i, y_j)$. It's important to note that whether the true function or the approximation is used, the similarities are computed for all pairs of points, not only the k nearest neighbours.

The Figure 3 shows an example of V and W heatmaps. Again, the matrices are ordered according to the leaf sequence of hierarchical clustering of the V matrix.



(a) V matrix heatmap.

(b) W matrix heatmap.

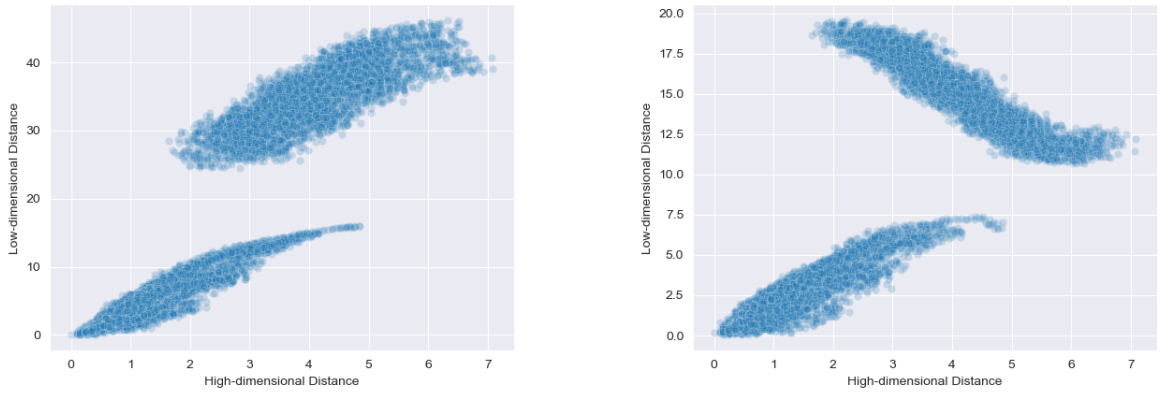
Figure 3: Heatmaps of the UMAP's similarity matrices for the Iris dataset (see Section 2.6).

2.2 Distance fit plot

In a distance fit plot, each point represents Euclidean distance between two points. The x-axis shows the distance between a pair of points in the original dataset, and the y-axis shows the distance between the same pair of points in the embedding. The scale of the axes is not meaningful because the distance in high- and low-dimensional spaces is not comparable, but the relative position of points on the plot is informative. Ideally, all points should be located on the diagonal, which indicates perfectly preserved distances. Conversely, if a point is located far from the diagonal, it indicates a disparity between the original and embedded distances: either two distant points in the original space are mapped close together in the embedding, or similar points have been separated. Furthermore, the most important region of the plot is the lower left quadrant. Those are pairs of points that are neighbours in the original dataset and should be preserved as such. Thus, if the quality of the dimensionality reduction is good, points in this region should be located on the diagonal. On the other hand, there should be no points in the upper left region of the plot because that would suggest that there are points which are originally close neighbours that have been mapped too far apart. Meanwhile, the right region of the plot is not as crucial because precise preservation of distance between dissimilar points is not prioritised by t-SNE and UMAP.

For example, Figure 4 shows the distance fit plots of the Iris dataset. The points lie approximately on the diagonal, especially in Figure 4a, which suggests good preservation of distances. It is important to note that t-SNE and UMAP are not fully deterministic, which means that the result of dimensionality reduction can change due to randomness. Here, the rotation of the upper cluster in Fig. 4b depends on the position of clusters in the embedding, and often closely resembles Fig. 4a.

The two clusters in the plot suggest the existence of clusters in the data, but, importantly, do not correspond to the clusters in the data themselves. The lower cluster in the plot reflects distances



(a) Distance fit plot for t-SNE.

(b) Distance fit plot for UMAP.

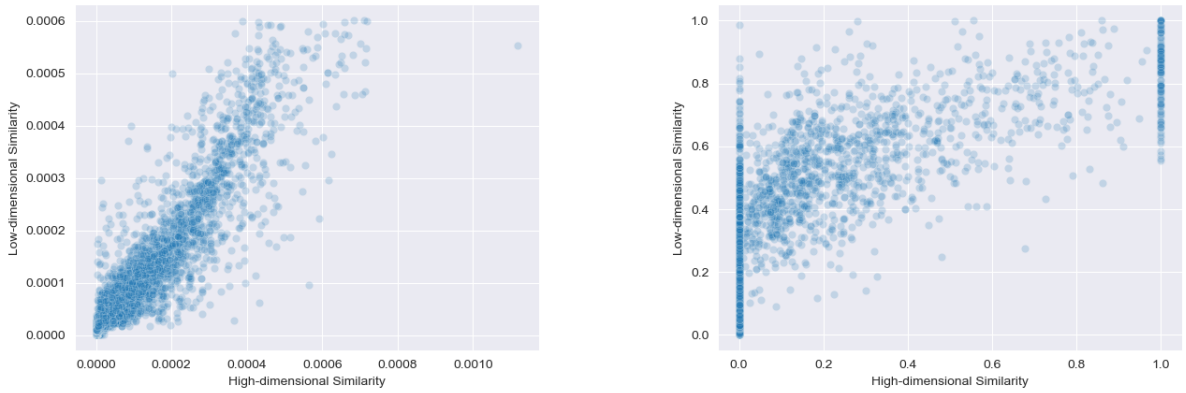
Figure 4: Distance fit plots of the Iris dataset (see Section 2.6) show how pair-wise distances are preserved in the mapping from the original space to the low-dimensional embedding.

between points that belong to the same cluster in the original data, while the upper cluster represents larger distances between points from two separate clusters. In this case, it is therefore the lower cluster that holds more significance in the assessment of the quality of dimensionality reduction.

2.3 Matrix fit plot

Similarly to the distance fit plot, we plot the high- and low-dimensional similarities. Each point corresponds to the similarity value between a pair of points. Depending on the dimensionality reduction methods, the x-axis shows the entries of either P or V matrix, and the y-axis shows the similarities from either Q or W matrix. Ideally, all points should be located on the diagonal, which indicates the perfect preservation of similarities. Practically, the most important region of the plot is the right side of the plot. Points in the upper right corner represent neighbours in the original dataset that were correctly mapped close to each other in the embedding. Moreover, the lower right quadrant of the plot should be empty because points there would correspond to neighbouring data points that were separated in the embedding. However, usually most points will be located in the lower left corner since they represent dissimilar points that remained distant after dimensionality reduction.

Figure 5 shows matrix fit plots of the Iris dataset. They exhibit similar trends, but differ visually due to distinct mathematical definitions of similarities in t-SNE and UMAP. The matrix fit plot for t-SNE (Fig. 5a) shows good preservation of similarities. The appearance of the plot is distorted because of a single pair of points with extremely high high-dimensional similarity (upper right corner), but the rest of points is located approximately on a diagonal. In UMAP, on the other hand, there are numerous points located on the left and right edges of the plot (Fig. 5b). This is due to the fact that most pairs of points in the original dataset, except for k neighbours of each point, have similarity 0, which corresponds to the left edge. Additionally, all points are connected to at least one point with similarity 1, which explains the crowded right edge. Despite that, the rest of



(a) Matrix fit plot for t-SNE.

(b) Matrix fit plot for UMAP.

Figure 5: Matrix fit plots of the Iris dataset (see Section 2.6) show how pair-wise similarities are preserved in the mapping from the original space to the low-dimensional embedding.

the points is located roughly on a diagonal. When assessing the quality of the embedding, the right side of the plot is again the most relevant. There the figure shows a few points that were nearest neighbours in the original dataset, but have only moderate similarity in the embedding. Fortunately, there are no points that were close neighbours originally, but are distant in the low-dimensional space. Overall, based on the diagnostic plots introduced so far, the embeddings produced by t-SNE and UMAP retain most of the relevant structure.

2.4 Outlier detection

A limitation of the dimensionality reduction performed by techniques such as t-SNE or UMAP is the tendency to draw outliers to nearby clusters. This happens because the neighbourhood of each points is scaled according to the local density. As a result, even though an outlier's nearest neighbours are distant in absolute terms, they are still considered similar within the scaled context. Although there is no universal definition of an outlier, the user should be aware of any points in the dataset that are likely candidates due to being significantly distant from the main distribution. In this section, we introduce a measure to identify such points by estimating how isolated a given point is relatively to the rest of the original data.

In the implementation of t-SNE and UMAP, all similarity matrices are symmetrised; however, the asymmetric versions of P and V matrices, which represent the conditional probabilities in high-dimensional data, encode valuable information about the relative isolation of points, which could correspond to outliers. Thus, we define the sum of incoming similarities of a point x_i as

$$\sum_j p_{i|j}$$

in the case of t-SNE and as

$$\sum_j v_{i|j}$$

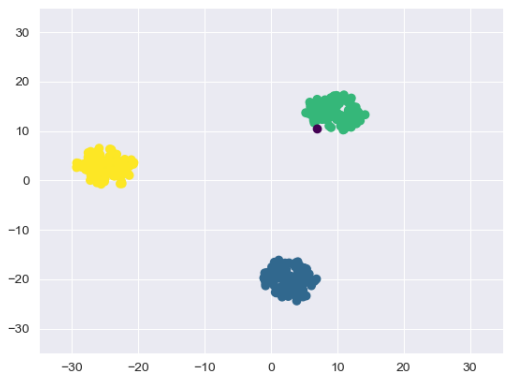
for UMAP. The conditional probabilities $p_{i|j}$ and $v_{i|j}$ can be interpreted as a measure of how much point x_j considers point x_i its neighbour. Therefore, the sum of incoming similarities indicates how much other points consider x_i a close neighbour. If the sum is relatively low, the point x_i is likely an outlier.

Figure 6: First two features of a three-dimensional dataset created with the *make_blobs* function from scikit-learn. There are three clusters with 100 samples each. The outlier, marked in purple, was added manually.

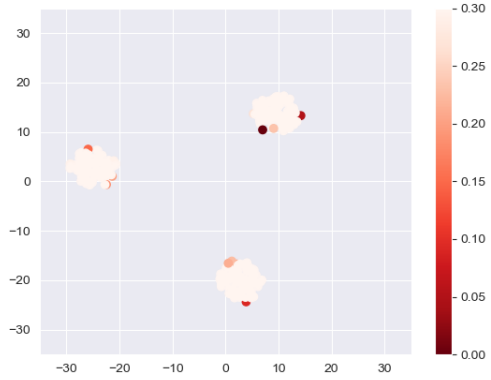
Figure 6 visualises the first two features of a three dimensional dataset with three clusters and an outlier. After we performed dimensionality reduction on this dataset (Figures 7a and 8a), the outlier was mapped close to the green cluster because it considered data points in that cluster to be its k close neighbours, even though they were far away. In this case, the attractive forces outweighed the repulsive forces from distant points and the outlier "gravitated" towards the cluster. The outlier wouldn't be visible in the embedding created by either t-SNE (Fig. 7a) or UMAP (Fig. 8a) if the points were not coloured according to the original clusters.

In Figures 7b and 8b we show the embedded data points coloured by their individual sums of incoming similarities. Darker points represent points with low values of the sum. The outlier stands out as the darkest in both graphs. Thus, even without knowing the original structure of the dataset, we could expect such point to be an outlier.

Notably, Fig. 7b shows several other points that appear darker than the rest. These are situated on the edges of very dense clusters in the original space. Relative to the local density, the darker points are further away from the cluster centres, resulting in lower sums of incoming similarities. Additionally, t-SNE assigns to all pairs of points a non-zero probability of being neighbours, even if the similarity is minute. As a result, even the outlier has a non-zero similarity sum. Since these values are all quite low, the peripheral points and the outlier appear similarly coloured in Fig. 7b. The lack of clear distinction between the outlier and the peripheral points encourages a cautious interpretation: the dark points likely lie further away from the original clusters than the visualisation

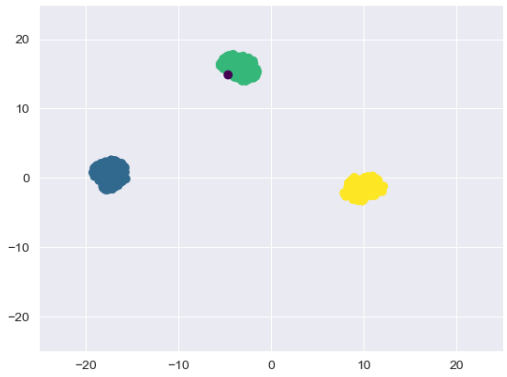


(a) t-SNE's result of dimensionality reduction on the dataset from Fig. 6 with perpexity = 30. The outlier was mapped close to the green cluster.

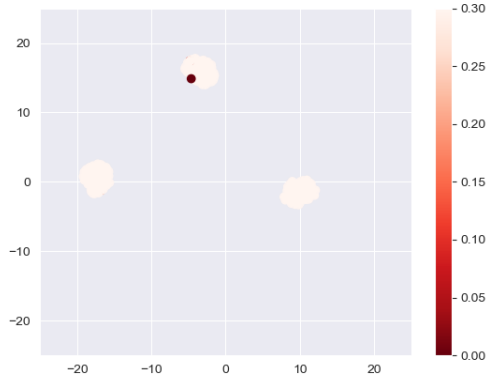


(b) The points in the embedding are coloured according to the individual sums of incoming similarities. The color range was capped at 0.3 to enhance the visibility of points with the lowest scores.

Figure 7: The behaviour of an outlier and the sum of incoming similarities in t-SNE.



(a) UMAP's result of dimensionality reduction on the dataset from Fig. 6 with $k = 30$. The outlier was mapped close to the green cluster.



(b) Embedding coloured by the sums of incoming similarities. The color range was capped at 0.3 to enhance the visibility of points with the lowest scores.

Figure 8: The behaviour of an outlier and the sum of incoming similarites in UMAP.

of the embedding suggests. On the other hand, in the UMAP graph in Fig. 8b, the outlier stands out as the only dark point. In contrast to t-SNE, UMAP assigns similarity 0 to all points which are not among the k closest neighbours. Since the outlier is not a neighbour for any data point, its similarity sum is 0, which makes it stand out. If we consider an outlier a point that is not considered a neighbor by any other, the proposed measure effectively identifies such outliers in the embedded dataset.

2.5 Individual cost

Kullback-Leibler (KL) divergence[KL51] is a statistical measure of the difference between two probability distributions. t-SNE uses it as the cost function to quantify and optimize the difference between high- and low-dimensional similarities. However, some information is lost because it is not known how much each point contributes to the final value. Also, before calculating the cost, t-SNE symmetrises the P matrix, which further obscures the details of the relationships between points. Therefore, to recover some of that information, we use the KL divergence with asymmetric high-dimensional similarities to calculate the cost for each data point. The goal is to show whether some points or neighbourhoods are preserved better than other. We define the individual cost of the data point y_i for t-SNE as

$$C_i^{t-SNE} = \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{ij}}.$$

On the other hand, UMAP uses cross-entropy as its loss function. Similarly to t-SNE, we use the asymmetric high-dimensional similarities. However, the cross-entropy cannot be applied directly. The V matrix is sparse and the W matrix is not, which leads to the problem of undefined values in mathematical functions. UMAP's solution is to use negative sampling to approximate the contribution of pairs of points for which the high-dimensional similarity is 0. Therefore, the exact contribution of each point to the loss value is not known. To remedy that, instead of using negative sampling, we added a small ϵ to all entries in the similarity matrices and normalized each row. These operations allow for the calculation of the cross-entropy for all pairs of points while preserving the relative similarity strengths. The advantage of this approach is the ability to quantify each point's contribution to the loss. Formally, we define the individual cost of the data point y_i for UMAP as

$$C_i^{UMAP} = \sum_j v'_{j|i} \log \frac{v'_{j|i}}{w'_{ij}} + (1 - v'_{j|i}) \log \frac{1 - v'_{j|i}}{1 - w'_{ij}}$$

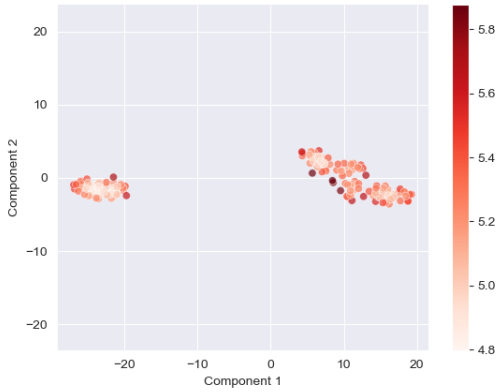
where

$$v'_{j|i} = \frac{v_{j|i} + \epsilon}{\sum_{k \neq i} v_{k|i}}$$

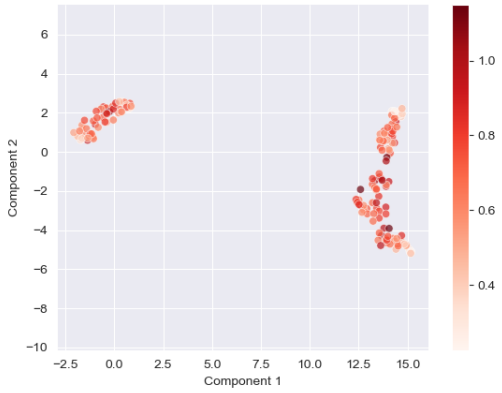
and

$$w'_{ij} = \frac{w_{ij} + \epsilon}{\sum_{k \neq i} w_{ik}}.$$

In the implementation, we used $\epsilon = 10^{-12}$ because, for a reasonable number of nearest neighbours, the similarities are expected to be much greater than ϵ. Notably, the individual cost is calculated differently depending on the dimensionality reduction technique and so the difference between the loss values is not informative. Thus, the individual cost, as defined in this paper, should not be used to choose one technique over the other, but to diagnose the behaviour of a given embedding. The range of the individual cost is comparable only between runs of the same diagnostic technique on the same dataset, as in this case the overall cost is indicative of the embedding's quality.



(a) Individual cost plot for t-SNE on the Iris dataset.



(b) Individual cost plot for UMAP on the Iris dataset.

Figure 9: Embeddings of the Iris dataset into two dimensions. Points are coloured according to their individual cost.

For example, Fig. 9 shows embeddings of the Iris dataset coloured according to the individual cost of each data point. The difference in the range of values on the graphs is not informative because the value of the individual cost is meaningful only within the context of an embedding. In Fig.9a, the points on the border of the left cluster are darker, which suggests that their relative position isn't preserved as well as the relative position of the lighter points inside the cluster. This might be due to the common issue with dimensionality reduction, where there is not as much space in lower dimensions, and some points that were near the centre of a high-dimensional cluster are now pushed to the outside. In Fig. 9b, there are no clear regions with either high or low individual costs. This suggests that the position of points is similarly well-preserved across the dataset, with a few exceptions of darker points, which, for example, might have been further away from the original clusters than it is indicated by their position in the embedding. An issue with the individual cost is that, just by looking at the result, it is not clear why a data point has a high cost. There are two possible explanations. First, the point's relative position might be much different from its position in the original dataset, but it can also be because the point was originally an outlier, which was drawn closer to the nearest cluster. We show in Section 4 how analysing the individual cost together with the outlier can help distinguish between the two situations.

2.6 Software and datasets

The source code for this project is available at https://github.com/dominika-haik/umap_tsne_diagnostics

All experiments were conducted using the following software:

- Python 3.12
- NumPy 1.26 [HMvdW+20] - numerical computations
- Scikit-learn 1.6 [PVG+11] - t-SNE implementation and datasets

- Matplotlib 3.10 [Hun07] - data visualisation
- Seaborn 0.13 [Was21] - data visualisation
- Scipy 1.15 [VGO+20] - scientific computation
- UMAP-learn 0.5 [MHM18] - UMAP implementation
- GitHub Copilot - assistant in creating the code documentation

Datasets:

- Iris dataset [Fis36], accessed via the scikit-learn library. The dataset contains 150 instances of three classes, 50 instances each. The classes represent species of iris plants. The dataset has four features: the length and the width of the sepals and petals.
- MNIST dataset [Den12], accessed via OpenML through the scikit-learn library. The dataset contains 70 000 28x28 images of handwritten digits.
- In addition to the above dataset, we used multiple synthetic datasets generated with the *make_blobs* function from the sklearn.datasets module or the *random.uniform* function from the numpy.random module. The 2D-in-5D dataset, where all points lie on a noisy 2D plane in a 5D space, was generated with the following code:

```
import numpy as np
base_data = np.random.uniform(low=0, high=1, size=(300, 2))
# Random projection matrix
projection_matrix = np.array([[ 1.912,  0.328,  0.838,  0.390, -1.309],
                              [ 0.650, -1.404,  0.849,  0.873, -1.175]])
# Project data into 5D
X = base_data @ projection_matrix  # shape (300, 5)
# Add small noise
X += np.random.normal(0, 0.1, size=X.shape)
```

- Another synthetic dataset is the Circle dataset, generated with the following code:

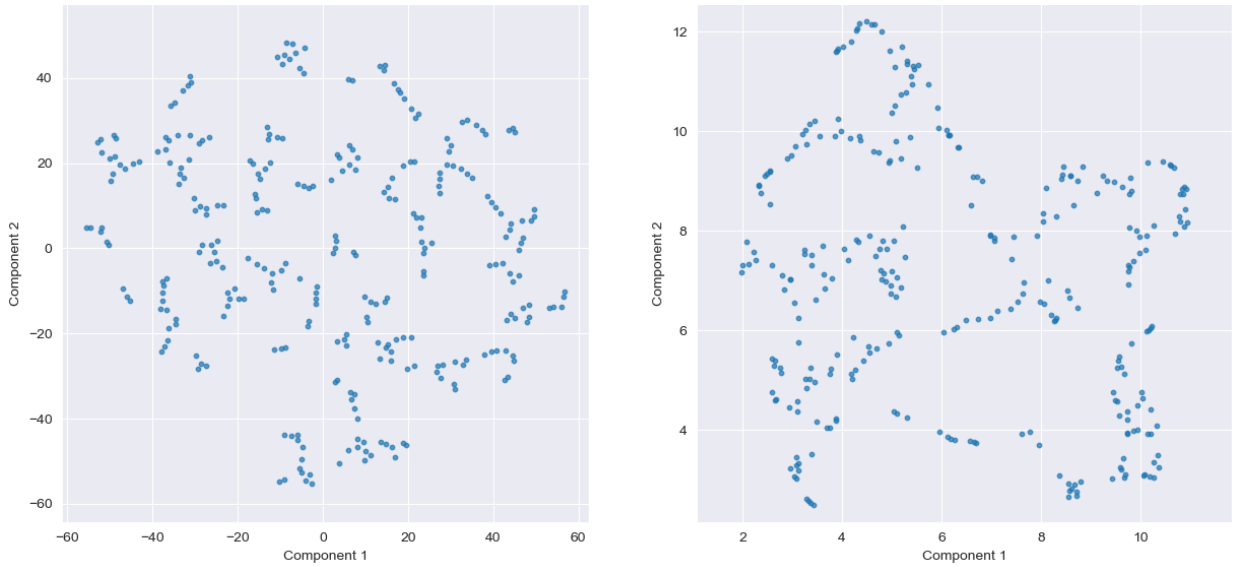
```
import numpy as np
r = 1
n = 10
A = np.arange(0, 2*np.pi, step=2*np.pi/n)
circle = np.column_stack((np.cos(A), np.sin(A)))
```

3 Results

To showcase the behaviour of the diagnostics we introduced in the previous section, we used them to investigate the outcomes of t-SNE and UMAP on two datasets. The first is a 5-dimensional

dataset with points uniformly distributed in space, created with the uniform function from the numpy.random module. The second dataset is also 5-dimensional, but the points are uniformly distributed on a noisy 2-dimensional plane (see Section 2.6 for details). Furthermore, we demonstrate a limitation of the outlier detection technique on a dataset in which three outliers are situated near each other.

3.1 5D Uniform Dataset



(a) t-SNE's embedding of the uniform dataset with perplexity = 5.

(b) UMAP's embedding of the uniform dataset with $k = 5$.

Figure 10: Projections of a 5-dimensional dataset with uniformly distributed points, with low values of the key hyperparameters. The plots show structures that do not exist in the original dataset.

First, we show the results of dimensionality reduction on the 5-dimensional uniform dataset. High and low values of the key hyperparameters, perplexity and k, were used for comparison. Figure 10 shows the embeddings produced by t-SNE and UMAP with both perplexity and k equal to 5. Given that the dataset has 300 data points, a value this low means that the algorithm will focus on only the closest neighbourhood of the data points. Therefore, even though the distribution is globally uniform, t-SNE and UMAP may pick up on the accidental patterns and local variance in density, and mistake it for meaningful structures. This phenomenon is visible in the visualisations of the embeddings: they show structures that are not present in the original dataset.

Figures 11 and 12 show sets of diagnostic plots for t-SNE and UMAP, respectively. An initial inspection suggests that t-SNE retains the majority of information from the original dataset. The heatmaps look the same, even though the Q matrix could be expected to show the clusters visible in the embedding. Moreover, the upper right fit plot of similarities suggests the perfect preservation of similarities. Even though the distance fit plot is crowded, its left side, especially the lower left corner, also indicate a high quality of the embedding. However, it is known that the embedding (Fig. 10a) reveals patterns that do not exist in the original dataset. With perplexity = 5, the task of

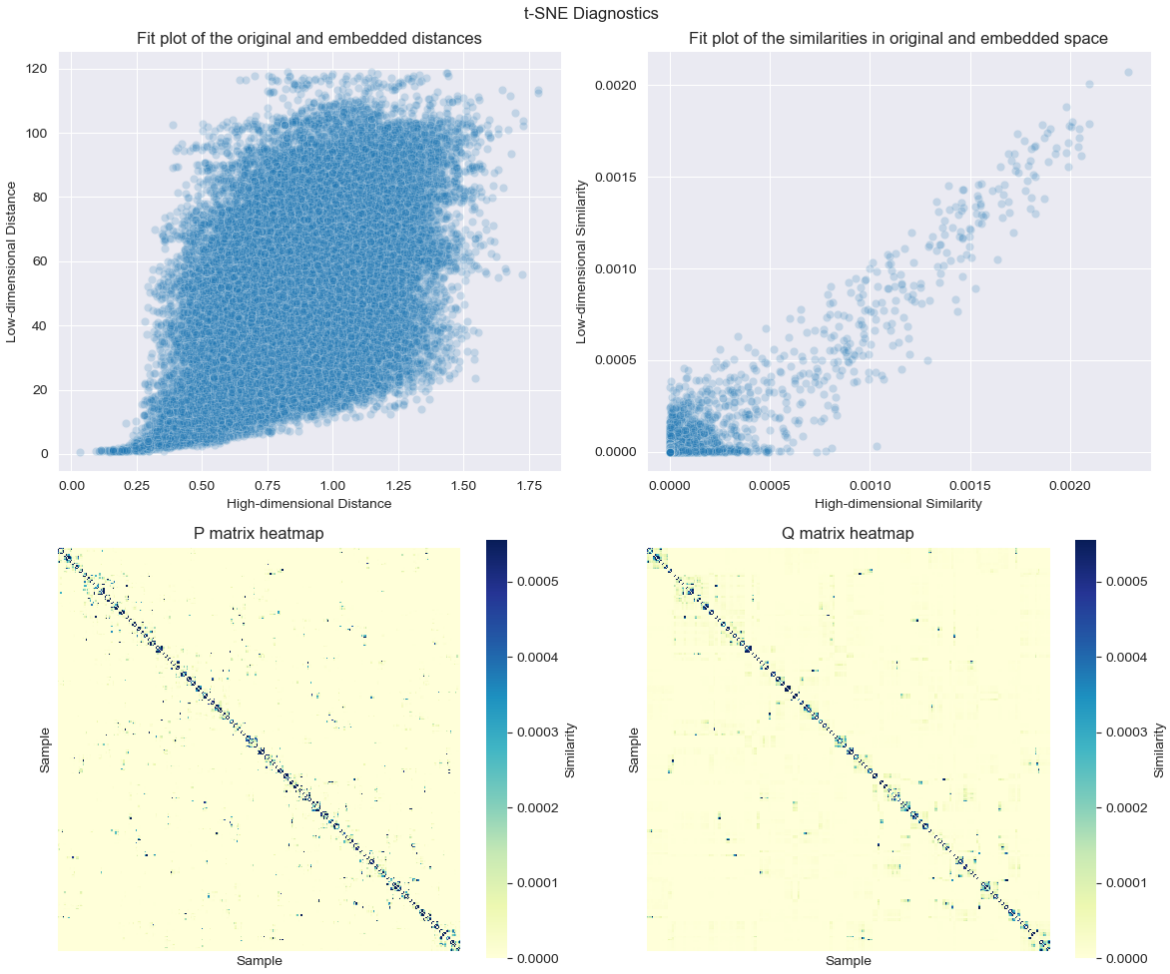


Figure 11: Diagnostic plots for the t-SNE's embedding of the 5-dimensional uniform dataset with perplexity = 5.

t-SNE is to preserve only the local "structures", and in that it is successful. The plot that stands out in Figure 11 is the distance fit plot, in which all points would be much closer to the diagonal if the embedding was as good as the other diagnostics suggest. Upon further inspection, the overwhelming majority of points in the matrix fit plot is crowded in the lower left corner, meaning that those pairs of data points had near-zero similarities in both high- and -low-dimensional spaces, which made the task of creating a projection relatively straightforward. Overall, the accuracy of the embedding, which showed nonexistent patterns, is not immediately revealed by the diagnostics because internal measures of the algorithm (the similarity matrices) were used as a basis.

On the other hand, based on Figure 12, UMAP's embedding is even less accurate. The matrix fit plot shows that many pairwise similarities were not preserved. There are numerous points which were not considered close neighbours in the original space (left edge of the plot), but have moderate to high similarity in the embedded space. The W matrix also appears, on average, much darker than the V matrix, which suggests that when applied to this dataset, UMAP suffers from the crowding problem. A possible explanation might be that fundamental assumptions of UMAP are

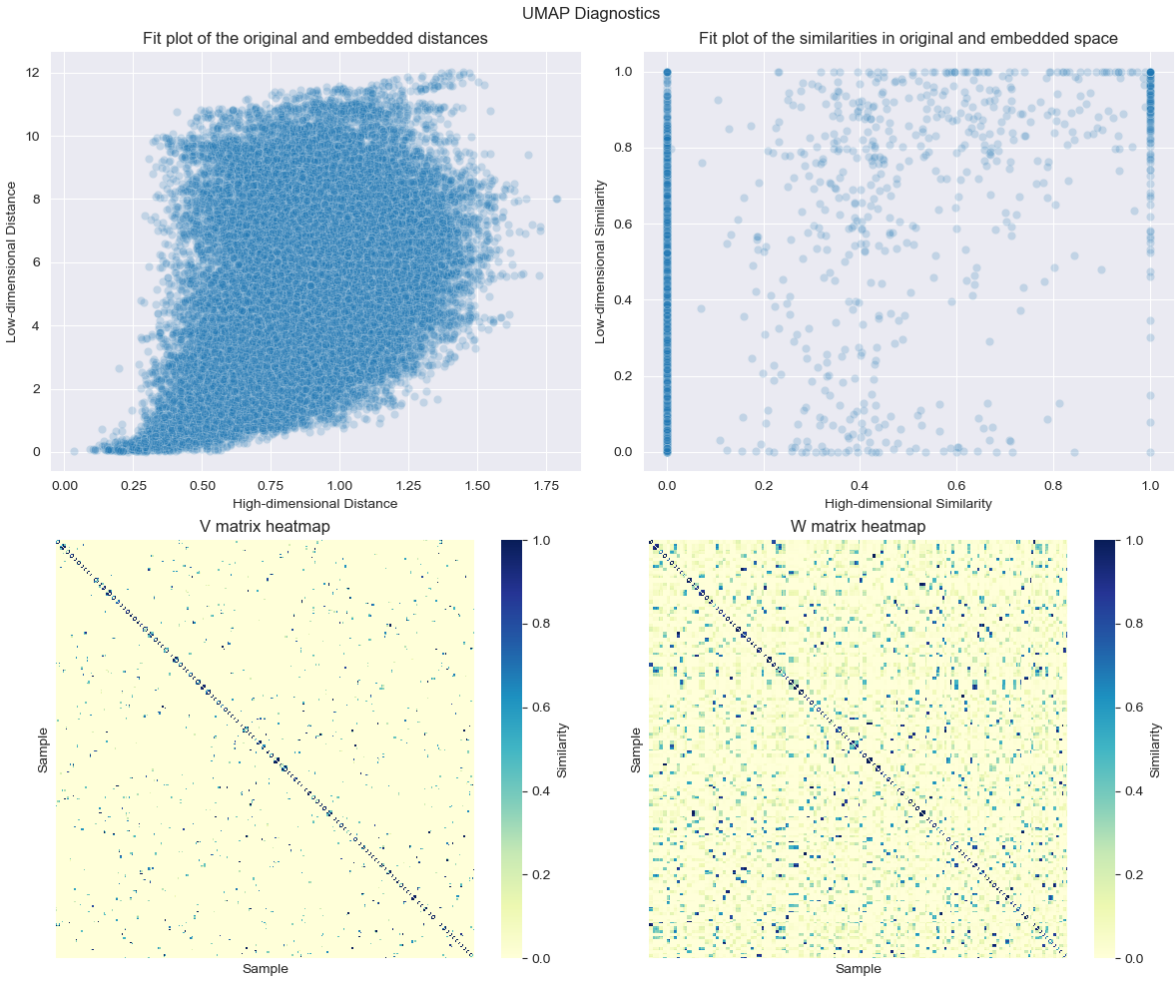
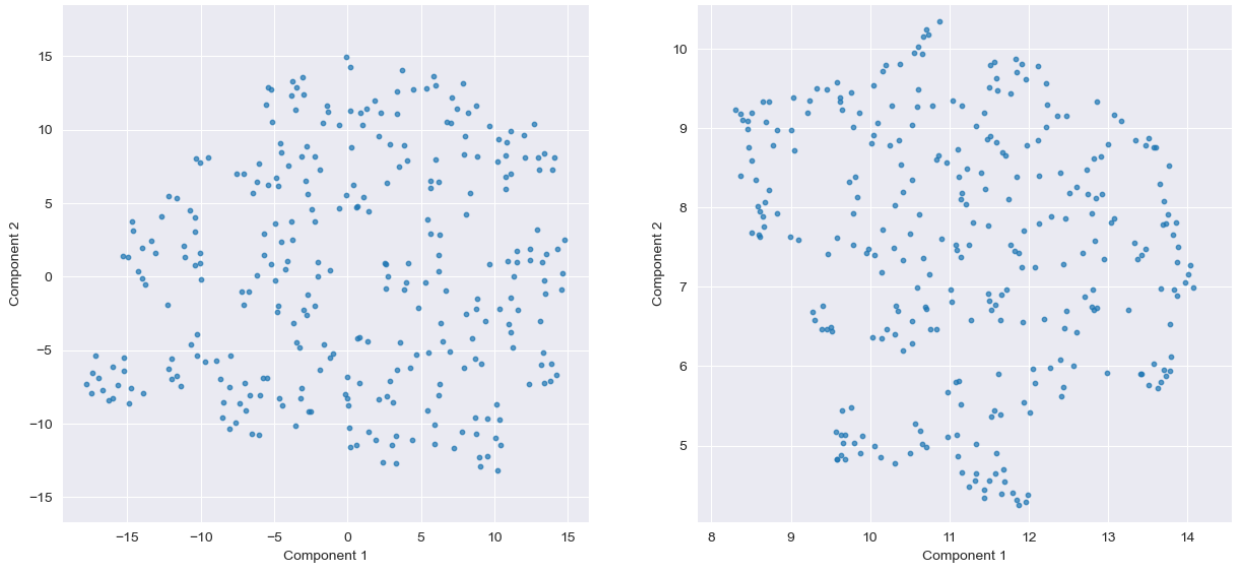


Figure 12: Diagnostic plots for the UMAP's embedding of the 5-dimensional uniform dataset with $k = 5$.

the cause of such behaviour. UMAP assumes that high-dimensional data lies on a locally connected manifold with lower intrinsic dimensionality. This assumption already suggests that UMAP expects to find some structure in the data. Since there is no structure in the uniformly distributed data, it struggles. However, these are only speculations and further investigation is needed to determine the cause.

We now turn to the case where t-SNE and UMAP are applied to the same dataset, but the hyperparameter values are higher. The resulting embeddings in Figure 13 better reflect the true structure of the dataset. With higher values of perplexity and k, the algorithm considers a larger neighbourhood of the data points, which results in the distribution of points being more uniform. The t-SNE diagnostics in Figure 14 appear similar to the previous case from Figure 11 and, again, suggest a high quality of the dimensionality reduction. The most noticeable difference appears in the matrix fit plot, where the points deviate more significantly from the diagonal; however, this can be explained by the higher difficulty of the ask itself. The projection of a dataset from five dimensions to two can preserve the original structure only to a limited extent. Given that, the



(a) t-SNE's embedding of the uniform dataset with perplexity = 30.

(b) UMAP's embedding of the uniform dataset with $k = 30$.

Figure 13: Projections of a 5-dimensional dataset with uniformly distributed points, with high values of the key hyperparameters. The points are more evenly distributed compared to the embeddings with low hyperparameter values.

matrix fit plot still indicates an informative embedding.

Similarly to the previous example, UMAP struggles with a high-dimensional, uniformly distributed dataset, even with a higher hyperparameter value. The diagnostic plots in Figure 15 indicate issues similar to the case with a lower k value - the points are too crowded. The matrix fit plot shows that many pairs of points which are dissimilar in the original space, are mapped as close neighbours. This behaviour might again be attributed to the UMAP's underlying assumptions, but more investigative work should be done.

3.2 2D-in-5D Uniform Dataset

In this section, we examine the behaviour of t-SNE and UMAP on a 5-dimensional dataset whose data lies on a 2-dimensional plane. The reduced intrinsic dimensionality significantly simplifies the embedding task. To capture the global structure, we use high values of the hyperparameters. Figure 16 shows the embeddings. In t-SNE's embedding in Figure 16a, there are no clusters and the data appears to be uniformly distributed in most regions. However, the UMAP's visualisation (Fig. 16b) depicts a resemblance of a plane, but the points are located mostly on the "edges" and are not uniformly distributed.

The diagnostic plots for t-SNE in Fig. 17 suggest high quality of the embedding. Importantly, the difference between this result and the diagnostics of the t-SNE reduction with a low perplexity (Fig. 11) lies in the distance fit plot. Here, the points are located perfectly on the diagonal, indicating that both small and large distances were preserved. This plot is the only one that does not rely

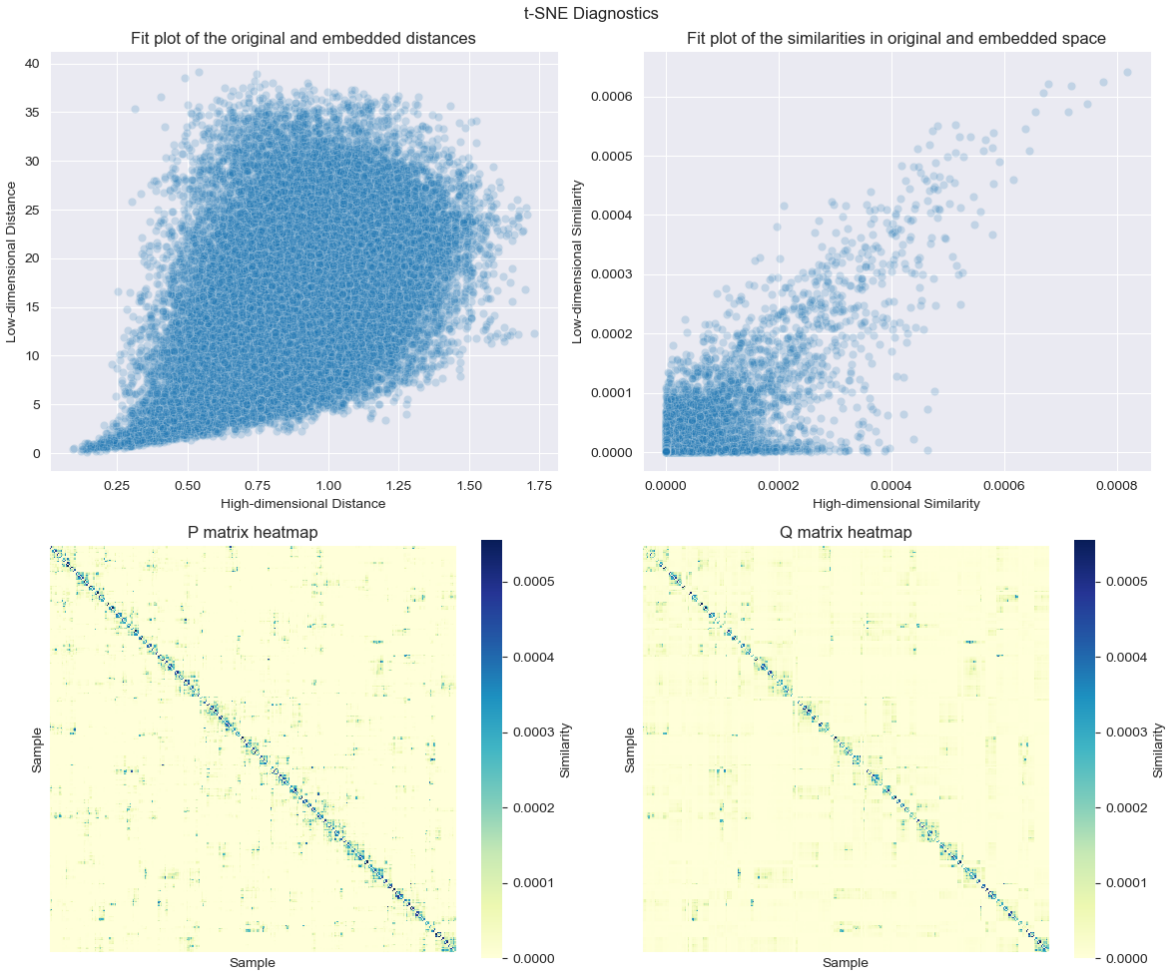


Figure 14: Diagnostic plots for t-SNE's embedding of the 5-dimensional uniform dataset with perplexity = 30.

on the similarity matrices and therefore can help diagnose issues that stem from an unsuitable hyperparameter. Another indicator of a genuinely high-quality embedding - rather than one resulting from a low perplexity value - is the matrix fit plot. In this case, more points deviate from the origin but remain aligned along the diagonal, indicating that non-negligible similarities are well preserved.

UMAP's embedding of the 2D-in-5D dataset is also much higher quality than in previous cases, but it seems to suffer from the same crowding problem where it clusters data points, even though they are dissimilar in the original space. This behaviour can be deduced from the diagnostic plots in Figure 18. The W matrix appears much darker than V, which suggests that more points are close neighbours in the embedded space. Moreover, there are numerous points in the upper left quadrant of the matrix fit plots, indicating that pairs of points with low similarity were mapped as close neighbours. It is important to note here that UMAP may produce better embeddings with higher values of the hyperparameter k. Nevertheless, the primary aim was to demonstrate that the quality of the dimensionality reduction can be assessed through a joint analysis of the proposed diagnostic plots.

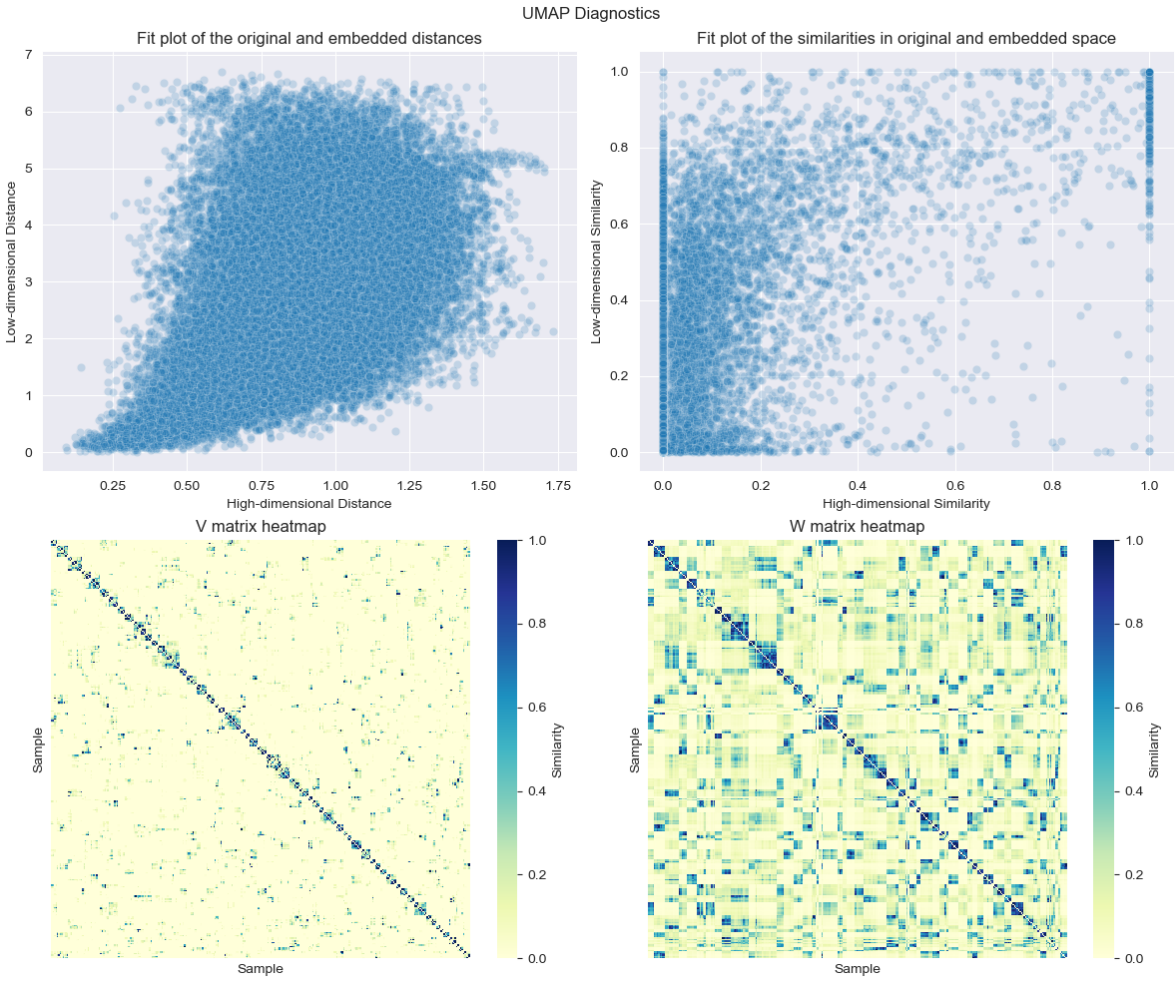


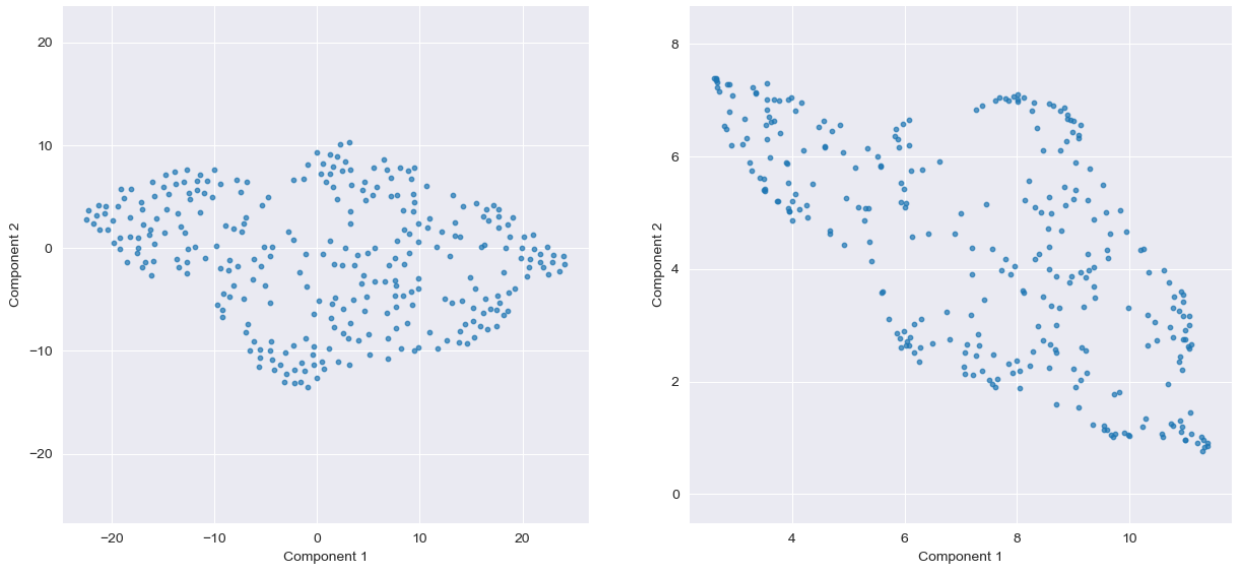
Figure 15: Diagnostic plots for UMAP's embedding of the 5-dimensional uniform dataset with $k = 30$.

3.3 Challenges in identifying outliers

In Section 2.4 we proposed a measure to indicate that given points are further apart from clusters in the original space than the embedding suggests. In this section, we present an inherent limitation of the method.

Figure 19 shows a similar three-dimensional dataset; however, in this example, it has three outliers which are located close to one another. The resulting embeddings, coloured by the sum of incoming similarities, are shown in Figure 20. The outliers are marked with dark outline. In this case, their sums of incoming similarities are relatively high and they are not coloured as potential outliers. This arises because, even though the outliers are further away from the clusters, they consider each other their closest neighbours.

A key limitation of any outlier detection method is the lack of the formal definition of an outlier. An important aspect to examine is whether the three points can be considered outliers. These points are not fully isolated on their own, but they are also not a part of any of the large, well-defined



(a) t-SNE's embedding of the 2D-in-5D dataset with perplexity = 30.

(b) UMAP's embedding of the 2D-in-5D dataset with $k = 30$.

Figure 16: Projections of the 2D-in-5D dataset (see Section 2.6), with high values of the key hyperparameters. The embeddings preserve the majority of the information due to the low intrinsic dimensionality of the dataset, which matches the embedded space.

clusters. Alternatively, since each of the three "outliers" appears among the k nearest neighbors of another data point, one could argue that they do not qualify as true outliers. Nevertheless, an effective outlier detection measure should be sensitive to such intermediate cases, as the embeddings misleadingly suggest that these points are typical, despite their distinct placement in the original space.

3.4 Individual cost on the Circle dataset

In Section 2.5, we introduced a metric for assessing how well the relative position of a data point is preserved. In this section, we illustrate the behaviour of individual cost on a simple Circle dataset (see Section 2.6). While it is unlikely to occur in the real world, it is intended to demonstrate the type of situations the diagnostic is meant to detect.

The Circle dataset has two dimensions, and its data points are arranged on a circle. A visualization of the dataset is shown in Figure 21. After performing dimensionality reduction to a single dimension, the points are arranged along a line. One way that t-SNE and UMAP achieve this is by opening the circle at some point and unraveling it. In doing so, the similarities between close neighbours are preserved for most points, except for the two points near where the circle was broken. It is impossible to perfectly preserve all distances when reducing to one dimension, so the structure of the data must be distorted in some way. Individual cost is meant to uncover this type of spatial disruption. The embeddings of the Circle dataset and the individual cost of each data point are shown in Figure 22. As expected, the cost for most points is relatively low, except for the two points at the ends. These points were close to each other in the original dataset, but due to the

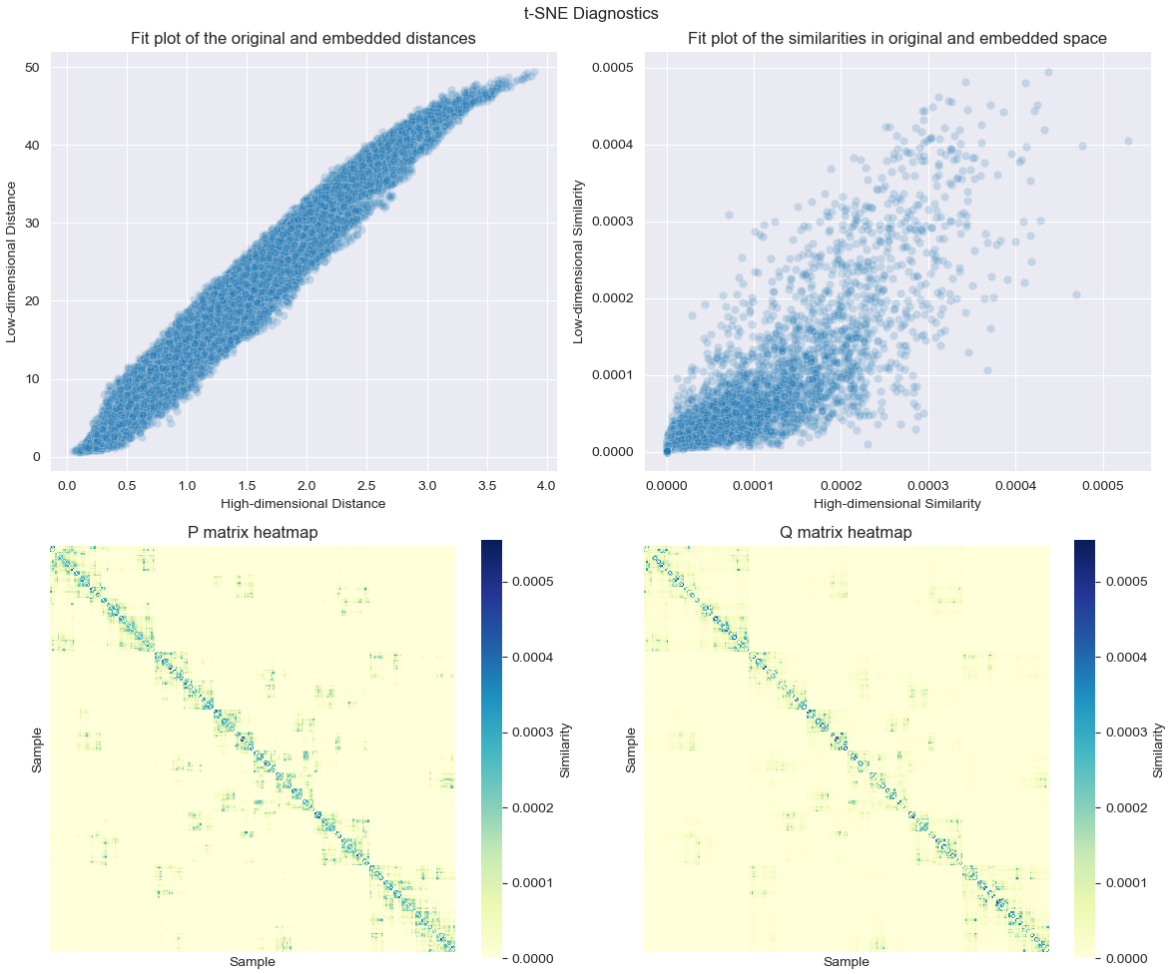


Figure 17: Diagnostic plots for t-SNE's embedding of the 2D-in-5D dataset (see Section 2.6) with perplexity = 30.

constraints of low-dimensional space, they were mapped farther apart.

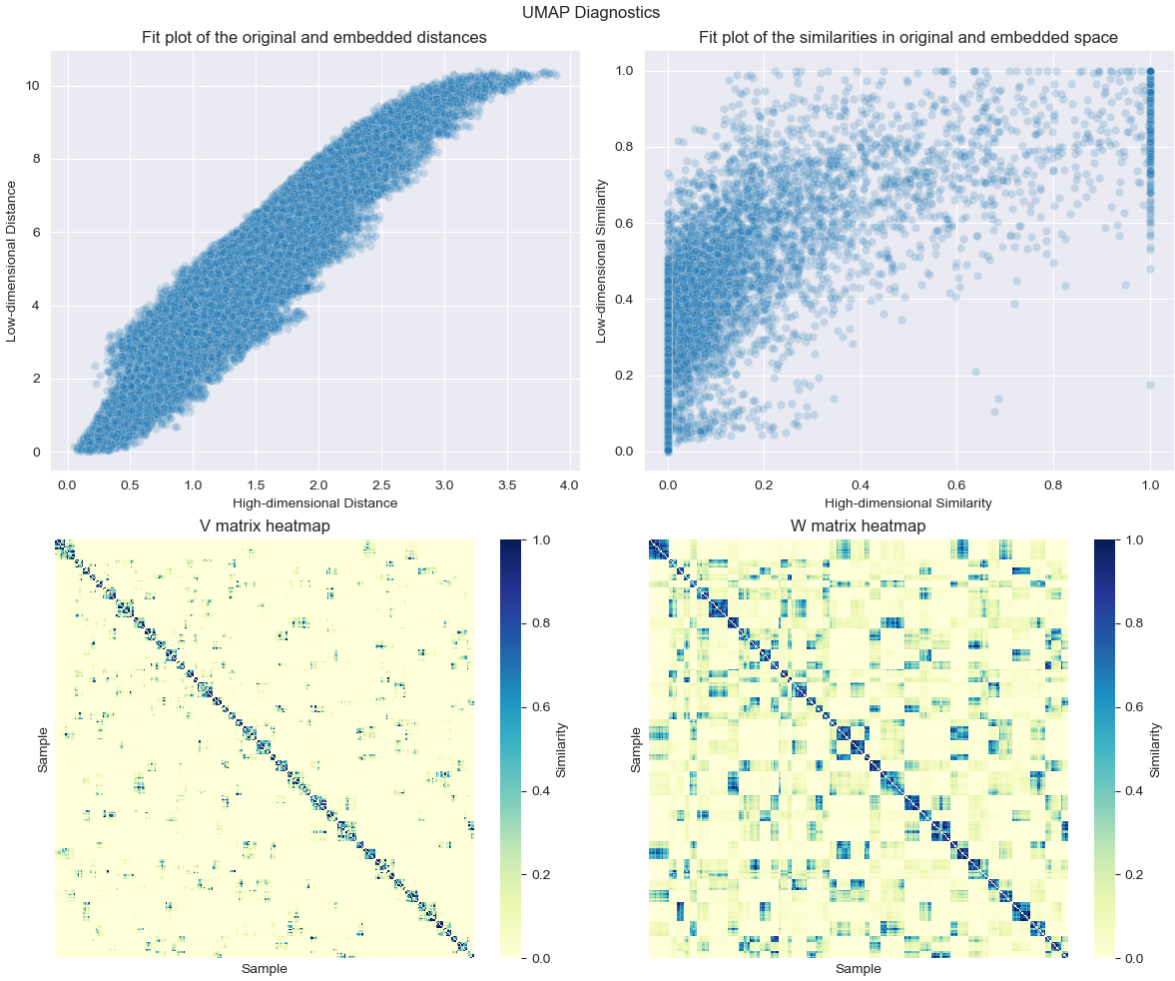


Figure 18: Diagnostic plots for UMAP's embedding of the 2D-in-5D dataset (see Section 2.6) with $k = 30$.

Figure 19: First two features of a three-dimensional dataset created with the make_blobs function from scikit-learn. Three outliers, marked in purple, were added manually.

(a) t-SNE's embedding.

(b) UMAP's embedding.

Figure 20: The embedded dataset coloured by the values of the sum of incoming similarities. The outliers are marked with dark edges. The color range was capped at 0.3 to enhance the visibility of points with the lowest scores; however, the outliers have relatively high scores.

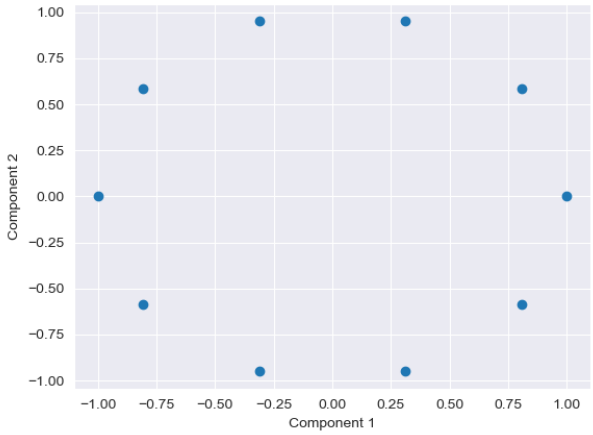
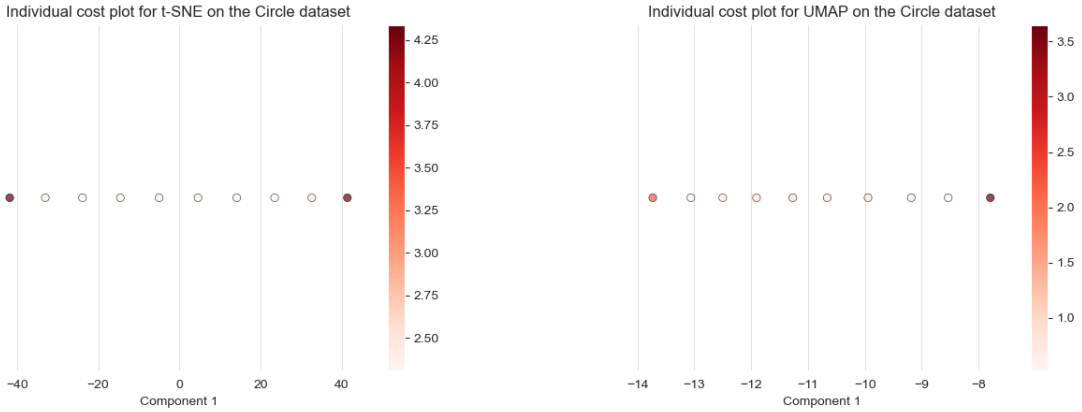


Figure 21: Visualisation of the two-dimensional Circle dataset (see Section 2.6).

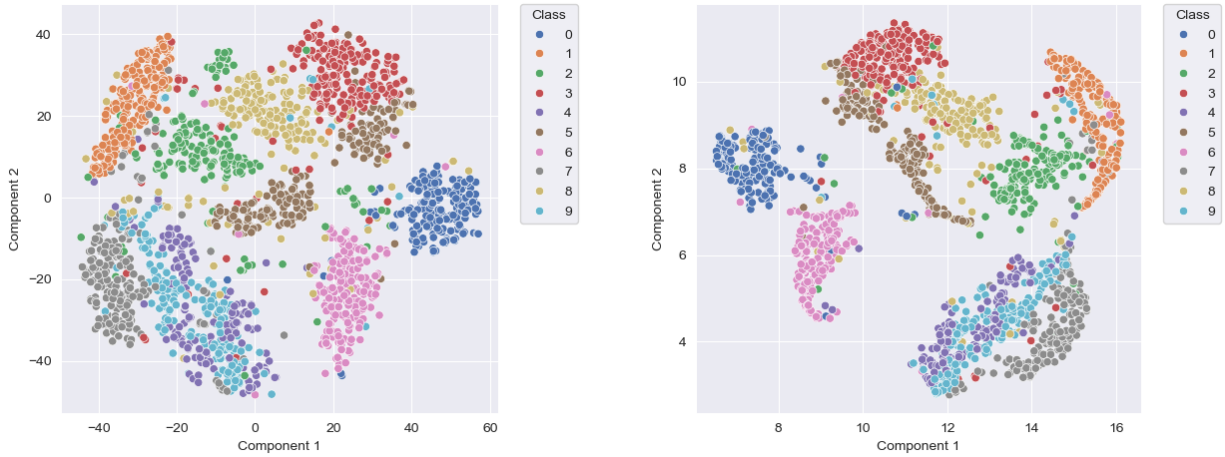


(a) Individual cost plot for t-SNE with perplexity = 3 on the Circle dataset.

(b) Individual cost plot for UMAP with $k = 3$ on the Circle dataset.

Figure 22: Embeddings of the Circle dataset into one dimension. Points are coloured according to their individual cost.

4 Applications to real-world data



(a) t-SNE's embedding of MNIST with perplexity = 30.

(b) UMAP's embedding of MNIST with $k = 30$.

Figure 23: Embeddings of the subset of the MNIST dataset into two dimensions. The plots show a random subset of 2000 samples. The points are coloured according to their original class.

In this section, we use a subset of the MNIST dataset (see Section 2.6) to demonstrate the behaviour of the diagnostics on real-world data. MNIST contains greyscale images of handwritten digits. Fig. 23 shows the result of applying dimensionality reduction to a subset of 2000 images. Some clusters are clearly separated, but others overlap. The imperfect embedding may be due to the small sample size which is not large enough to differentiate between similar classes, like 4 and 7, or due to t-SNE and UMAP not fully preserving the structure. The diagnostics allow us to evaluate the extend to which the latter is the case. Generally, t-SNE and UMAP work well on the full MNIST dataset, which can be seen in Appendix A, but the diagnostics cannot be directly applied to a dataset this big. We discuss this limitation later, along with a proposed solution for future work.

Fig. 24 shows four diagnostics of the t-SNE embedding: the distance fit plot, the matrix fit plot and the P and Q matrix heatmaps. Although the distance fit plot is crowded because of the size of the dataset, it still suggests that the distances are preserved well. The most important region is the left side of the plot, which shows how the distances between close neighbours are preserved. There, it is visible that small distances were preserved correctly, and none of the nearest neighbours were mapped far apart. However, many originally distant points were mapped close together. This observation is reinforced by the matrix fit plot, which shows that points that were similar in the original dataset have high similarity in the low-dimensional space. The majority of points are located above the plot's diagonal, indicating that pairs of points with non-negligible similarity were mapped even closer together. It is important to note that most points in the matrix fit plot are crowded in the lower left corner. These points are far from each other and therefore not similar, but the plot does not clearly show the proportion of such cases. Unlike the results for the Iris dataset in Fig. 2, the heatmaps for MNIST do not reveal clear cluster structure. This might be because the clusters in the embedding, even though they look distinct to the human eye, are often near each other, and data points from distinct classes are often close neighbours. Thus, even though the

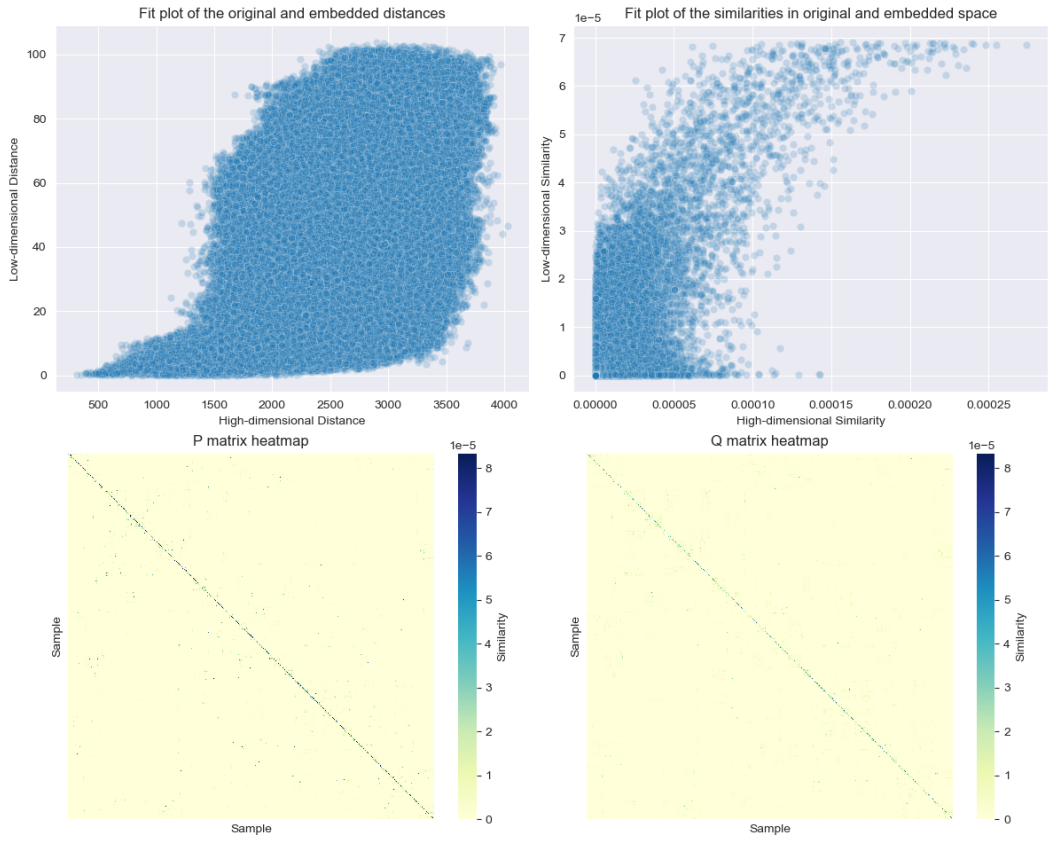


Figure 24: Diagnostics of the t-SNE's embedding of a subset of MNIST. Even though the distances and similarities are relatively well-preserved, the heatmaps do not reflect the clusters shown in the embedding.

matrices were sorted using hierarchical clustering, the structure has not emerged. Also, since the matrices are so large, it is easier to use the matrix fit plot to find any inconsistencies between them than the heatmaps.

The same set of diagnostics for UMAP is shown in Fig. The most notable difference between Fig. 24 and Fig. 25 is the matrix fit plot and the darker appearance of the W matrix heatmap. This is partially because, in UMAP, the V matrix is sparse and W is dense. When the heatmaps do not show visible clusters, visual comparison of V and W matrices becomes less informative. However, they still support the interpretation of the matrix fit plot. In UMAP's matrix fit plot, most points lie along the y-axis due to the sparsity of the V matrix. However, a limitation of this diagnostic plot is that it is unclear how many points are positioned there. Nevertheless, the matrix fit plot suggests that the similarities are not fully preserved. Some points that were close in the original dataset were mapped farther apart, as seen in the lower-right corner. This may explain the unusual shapes of the clusters in Fig., especially the crescent-shaped Clusters 1, 7 and, to some extent, 6.

Furthermore, it is possible to analyse the relative position of each data point individually. Fig. 26 shows the dataset embedded by t-SNE along with the individual cost and outlier score per point. The two measures complement each other to show whether a data point was suboptimally placed in

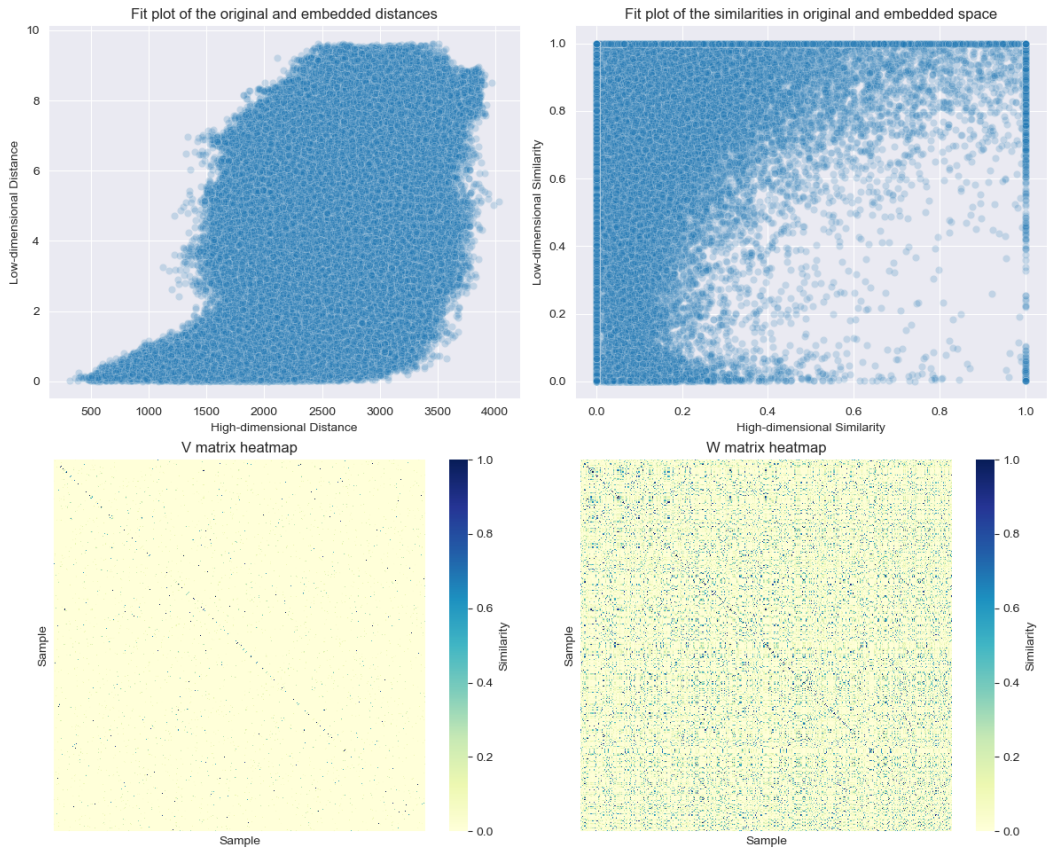
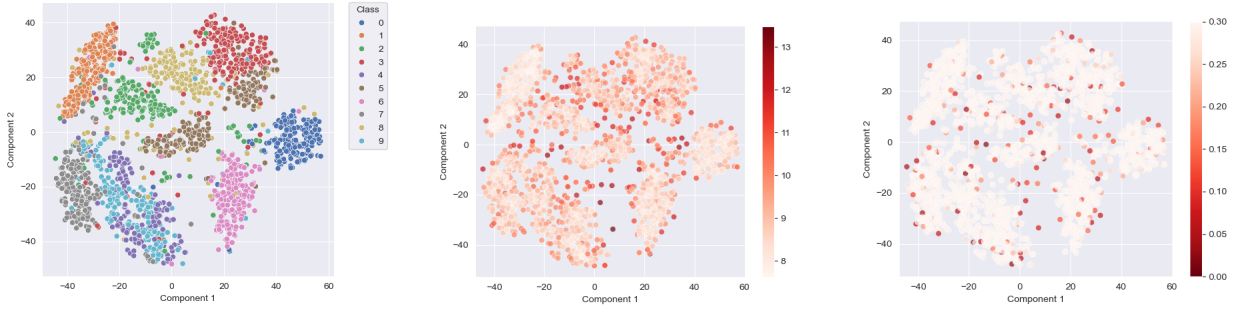


Figure 25: Diagnostics of the UMAP's embedding of a subset of MNIST. Although the distances are well-preserved, the similarities are not. V matrix heatmap appears much lighter due to the sparsity of the matrix.



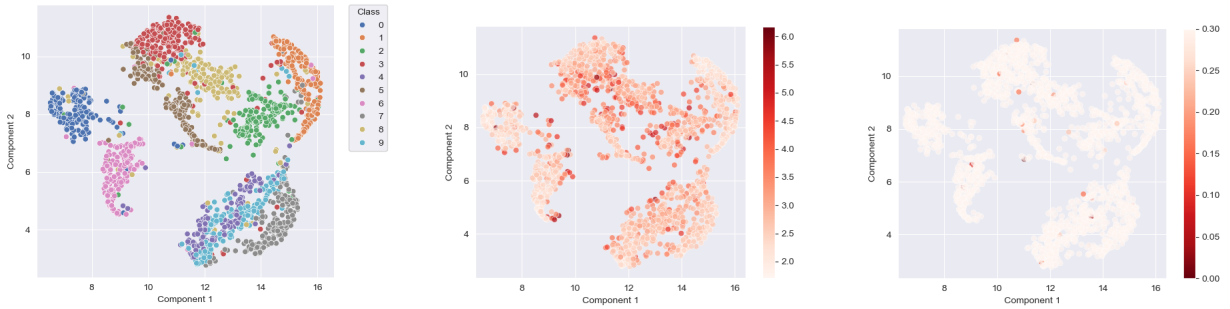
(a) t-SNE's embedding of a subset of MNIST.

(b) Individual cost.

(c) Outlier score.

Figure 26: Embedding of MNIST by t-SNE, coloured according to diagnostic metrics. The plots explain which points' relative position was not preserved in the embedding.

the embedding or it was an outlier grouped with a nearby cluster. For instance, Cluster 1 has the lowest overall cost, which suggest that there is a high confidence in grouping these points together. However, some boundary points in Cluster 1 belong to other classes and have a higher cost. Fig. 26c shows that many of them are possible outliers, which would explain why their misplacement.



(a) UMAP's embedding of a subset of MNIST.

(b) Individual cost.

(c) Outlier score.

Figure 27: Embedding of MNIST by UMAP, coloured according to diagnostic metrics. The plots explain which points' relative position was not preserved in the embedding.

Moreover, Cluster 2, marked in green, is split into two disconnected regions by Cluster 8. Points near the split have higher cost, suggesting that the their position is distorted. Additionally, a small group of green points next to Cluster 0 seems to have split Cluster 5. These points also have high individual cost, suggesting they may have become "stuck" during the embedding process. Interestingly, Clusters 4 and 9 are merged together, but it is not reflected in their cost. In this case, the data points were most likely highly similar, and the subset was not sufficiently large to distinguish between the two classes. Analysed together, the cost and outlier score plots provide valuable insight into position of certain points.

In the UMAP embedding shown in Fig. 27, Clusters 4 and 9 are again intertwined, but the consistently low cost suggests that this faithfully reflects the true structure of the subset. Although Cluster 2 is not split here, it is not as clearly defined as other clusters. This could indicate that either its structure is complex in the high-dimensional space and thus difficult to represent in two dimensions or no well-defined structure exists in the first place. The high average cost of Cluster 2 supports this interpretation.

Furthermore, Clusters 3, 5 and 8 are not separated, and may points near their boundaries have high costs, indicating poor preservation of their relative position. A key observation is that even though Fig. 27c identifies some of the high-cost points as potential outliers, there are fewer of them than in Fig. 26c This is due to the threshold set for the colour range of the plot. A default of 0.3 was based on empirical observations, but it might be valuable to establish concrete guidelines for choosing the threshold in the future. As previously discussed, there is no universally accepted definition of an outlier, so this metric only highlights isolated points. However, depending on the cap of the colour range, it may over- or under-identify potential outliers.

The diagnostics described above, in their current implementation, cannot be directly applied to a dataset the size of MNIST. For such high-volume data, the matrices become too dense to interpret, the fit plots are too crowded, and the visualisations may fail to render due to memory limitations. A potential workaround is to apply diagnostics only to a subset of the dataset, after performing the embedding of the full dataset. However, this approach does not give reliable results. Currently, the high-dimensional similarities are calculated based on the input dataset, with hyperparameters that,

by design, mirror the hyperparameters used during dimensionality reduction. If only a subset is provided as input, the similarities are distorted. This is because a subset of a large dataset is usually more sparse, so, given that perplexity and k are constant, the neighbourhoods of data points shift. The probability distribution centred at a data point, which determines the similarity between that point and its neighbours, will be scaled differently and tend to spread out more. Consequently, the neighbourhood of any given point will increase, effectively shifting focus towards the global structure of the incomplete dataset. As a result, the diagnostic output no longer accurately reflects the quality of the original dimensionality reduction.

A solution we propose is to modify the implementation so that it is possible to put the desired subset size as a parameter. The similarity matrices would be calculated for the entire dataset, but the diagnostic plots could be constructed only for a random subset of the given size. In UMAP, high-dimensional similarities are already calculated only for the k nearest neighbours, so it should be computationally feasible with the current implementation. For t-SNE, however, it would be beneficial to change the current implementation, which computes high-dimensional similarities for all pairs of points, to a method such as Barnes-Hut t-SNE for an efficient approximation. After calculating similarities for the original dataset, new, smaller similarity matrices could be constructed to include only similarities between points in the subset. This way, the similarity matrices can be computed based on the original dataset without losing information, and the diagnostic plots could show the results on a representative subset of data. Although these modifications are out of the scope of the current work, it may be beneficial to consider them in the future.

5 Discussions

In this paper, we introduced a set of diagnostic plots designed to assess the quality of dimensionality reduction techniques and demonstrated our implementation of these tools. We based multiple visual diagnostics on similarity matrices, which store pairwise similarities between data points in both the high-dimensional and low-dimensional spaces. We first visualised the similarity structure using heatmaps, enabling a qualitative comparison between the original and embedded spaces. To assess whether the similarities were preserved, we introduced the matrix fit plot, which highlights the relation between pairwise similarities before and after embedding. In addition, we proposed the distance fit plot to evaluate how well relative pairwise distances are preserved. Furthermore, we presented a method to identify points that appear embedded closer within clusters than their original positions justify by quantifying how many points consider them their close neighbours. Lastly, we introduced a measure to show how much each point contributes to the final loss value after the embedding is optimised.

To demonstrate the utility of these diagnostics, we applied them to two synthetic datasets: a 5-dimensional dataset with uniformly distributed points, and a 5-dimensional dataset in which points lie on a noisy 2-dimensional plane. The plots revealed that embeddings with suboptimal hyperparameters can appear deceptively good. By jointly interpreting all diagnostics, we could distinguish such cases from genuinely well-structured embeddings. We further illustrated the behaviour of t-SNE and UMAP when applied to data with low intrinsic dimensionality, showing how the reduced complexity led to cleaner diagnostic outputs. Furthermore, we highlighted a limitation of the outlier detection method: when multiple potential outliers are located in close proximity, their

mutual influence may suppress their individual outlier scores. Lastly, we showed how individual cost may uncover situations in which dimensionality reduction "breaks" the high-dimensional space.

Additionally, in Section 4, we demonstrated the use of the proposed diagnostics on real-world data. While the diagnostics proved effective on a reduced subset of the MNIST dataset, we uncovered a significant constraint. A key limitation of the presented diagnostics lies in their reliance on visualising large quantities of data, which becomes increasingly challenging as the dataset size grows. As the number of points scales exponentially, fit plots may become too dense to interpret, heatmaps may lose detail, and some plots may even fail to render. A partial workaround is to visualise only a subset of the data, but, as discussed in Section 4, this prevents comprehensive analysis. In the same section, we proposed a potential solution. An interesting direction for future research would be the development of scalable diagnostics, comparable in simplicity and interpretability to the Scree plot used in PCA.

A further constraint that stems from the need for visualisations is that the methods that plot the individual cost and the outlier score show only one- and two-dimensional embeddings. It is possible to visualise three-dimensional embeddings by directly calculating the outlier score or individual cost through methods provided in the package, and then manually plotting them on a three-dimensional scatter plot. For embeddings into higher dimensions, the possibilities for a visual analysis are limited, and other measures for interpreting the cost should be employed.

Furthermore, colour ranges in plots like the individual cost are informative only in the context of applying the same technique on the same dataset. Since there are no definitive thresholds for which individual cost and outlier cost values are considered low or high, the colour ranges are adjusted to the values of a particular embedding. Thus, even the comparison between consecutive runs of the same algorithm is challenging, as the viewer has to translate the colour to values in order to make a meaningful comparison. Future research could include an attempt to establish guidelines to interpret absolute values of the individual cost and the outlier score. This could, for example, help avoid situations in which many points are incorrectly placed in an embedding and have a high cost, but it is obscured by one point with an unusually high cost. Such a point would skew the colour range and make other points appear to have a low cost.

This issue is also related to the question of visual emphasis in a diagnostic plot. For instance, the distance fit plot is usually crowded, and the attention of a viewer is drawn to the large cloud of points in the middle, while the actual important region is the lower-left corner and the empty upper-left area. Future research and improvements could focus on highlighting the regions of plots that are the most informative and thus simplifying the analysis of the embeddings.

As previously discussed, another limitation concerns the dependency on internal measures of the dimensionality reduction algorithms themselves, which constrains the diagnostic plots' utility for hyperparameter optimisation. Ideally, a more method-agnostic approach would provide deeper insight, especially for comparisons between dimensionality reduction methods. Nevertheless, we have shown that the current set of diagnostics, when interpreted together, can uncover cases where satisfactory appearance of some tools masks structural distortion caused by poor hyperparameter choices. In particular, while similarity-matrix-based diagnostics often gave seemingly excellent results, the distance fit plot revealed that the relative distances of the original data were not

faithfully preserved.

Another complication stems from the difference between the theoretical and implemented similarity functions in UMAP. Although the true function defines the low-dimensional similarities according to the underlying mathematical foundations, the actual implementation uses a smooth approximation. While the true function is arguably more suitable for assessing how well pairwise relationships are preserved, the approximated version reflects the behaviour of the algorithm. To accommodate both perspectives, our implementation allows users to choose which function to use, with the true function being the default.

A broader question is whether similarity measures derived from the dimensionality reduction technique itself provide a reliable basis for evaluating embedding quality. Our results indicate that these diagnostics, particularly when used alongside the distance fit plot, can offer meaningful insight into the embedding's structure. However, the distance fit plot frequently proved essential in revealing inconsistencies that similarity-based diagnostics failed to detect. Similarity-based plots showing great results can be misleading, particularly for users unfamiliar with their interpretation, as they may suggest that the embedding is of higher quality than it truly is. This highlights the importance of incorporating diagnostics that are independent of internal similarity measures, which could offer a more balanced assessment. That being said, it can be argued that both t-SNE and UMAP do not claim to preserve the distances between data points, but produce a representation of the high-dimensional structure. Thus, a perfect distance fit plot is not the goal. The diagnostics offer distinct perspectives on the embeddings, which can be used to asses embeddings on a case-by-case basis with consideration for individual priorities.

Lastly, there is a potential for the individual cost and the outlier score to be used in the clustering of unlabeled datasets. The measures could be combined to create a trustworthiness score, which would quantify to what extent a point's position in the embedding should influence the boundaries of clusters. Data points that are likely misplaced or are potential outliers could be ignored. Future research could propose an improved technique to cluster datasets preprocessed with t-SNE and UMAP.

In conclusion, we presented a practical set of diagnostic tools for evaluating dimensionality reduction results. Our method enables users to assess whether an embedding meaningfully reflects the structure of the original data and to verify the preservation of pairwise relationships. Embeddings often present compelling visual patterns that may be overinterpreted, especially when users are primed to find structure. Our plots help to distinguish genuine structure from meaningless patterns, offering a foundation for robust analysis.

References

[Den12] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[Dui24] Nicolette Emily Duijn. *Diagnostic plots for the dimension reduction techniques t-SNE and UMAP*. Master thesis, Universiteit Leiden, August 2024.

[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, May 2004.

[HMvdW+20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[Hun07] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[Kan05] Gibbs Y Kanyongo. Determining the correct number of components to extract from a principal components analysis: a monte carlo study of the accuracy of the scree plot. *Journal of modern applied statistical methods*, 4(1):13, 2005.

[KHO25] Takuya Kataiwa, Cho Hakaze, and Tetsushi Ohki. Measuring intrinsic dimension of token embeddings. *arXiv preprint arXiv:2503.02142*, 2025.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[LMBH11] Wouter Lueks, Bassam Mokbel, Michael Biehl, and Barbara Hammer. How to evaluate dimensionality reduction? - improving the co-ranking matrix, 2011.

[LV09a] John A. Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7):1431–1443, 2009. Advances in Machine Learning and Computational Intelligence.

[LV09b] John A. Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7):1431–1443, 2009. Advances in Machine Learning and Computational Intelligence.

[Mea92] A. Mead. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(1):27–39, 1992.

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

[MTB22] Fatemeh Mostofi, Vedat Toğan, and Hasan Basri Başağa. Real-estate price prediction with deep neural network and principal component analysis. *Organization, Technology & Management in Construction*, 14(1):2741–2759, 2022.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blon-

del, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Roy24] Subhrajyoty Roy. Trustworthy dimensionality reduction. *arXiv preprint arXiv:2405.05868*, 2024.

[SWM93] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. SPIE, 1993.

[vdM] Laurens van der Maaten. t-sne. https://lvdmaaten.github.io/tsne/ [Accessed: March 1 2025].

[VdMH08] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

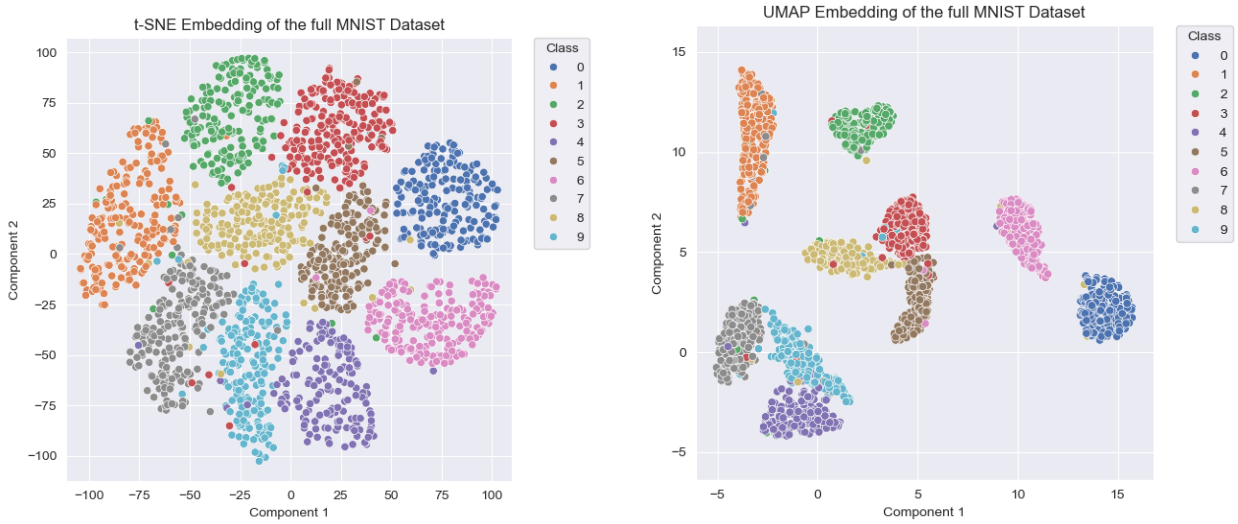
[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[Was21] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[Wu23] Yueheng Wu. *Uncovering Limitations in Dimensionality Reduction: A Tool for Diagnosing Individual Information Distortion in t-SNE*. Master thesis, Universiteit Leiden, August 2023.

[WVJ16] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.

A Appendix: Full MNIST Embedding



(a) t-SNE's embedding of the full MNIST.

(b) UMAP's embedding of the full MNIST.

Figure 28: The result of running t-SNE and UMAP on the full MNIST dataset. The classes are separated well and, compared to Fig. 23, there are no overlaps or split clusters.