UCLouvain

École polytechnique de Louvain

Directional quality assessment for nonlinear dimensionality reduction in data visualisation

Author: **Pierre REMACLE**
Supervisor: **John LEE**
Readers: **Pierre LAMBERT, Michel VERLEYSEN**
Academic year 2024–2025
Master [120] in Data Sciences Engineering

Contents

Chapter 1

Introduction

Dimensionality reduction (DR) is a fundamental technique in data analysis, enabling the simplification of high-dimensional datasets by projecting them into lower-dimensional spaces. This process addresses challenges such as the curse of dimensionality, where the sparsity of high-dimensional data complicates clustering, classification, and other analytical tasks. By reducing dimensionality, researchers can achieve more robust interpretations, reduce computational demands, and uncover patterns that might otherwise remain hidden.

One of the most powerful applications of dimensionality reduction lies in its ability to create interpretable visualizations. Reducing datasets to two or three dimensions allows researchers to identify relationships and structures that may not be apparent in the original data. These visualizations are particularly valuable in exploratory data analysis, serving as a bridge between raw, high-dimensional data and human intuition.

The utility of DR spans diverse domains, including biology, where it has become indispensable in analyzing high-dimensional datasets such as single-cell RNA sequencing (scRNA-seq). Techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) are widely used for such tasks. For example, t-SNE excels at uncovering local structures, visually separating cell types into distinct clusters that align with biological insights. However, these benefits are tempered by limitations. Specifically, t-SNE often fails to preserve global relationships between clusters, potentially leading to misleading interpretations in datasets where hierarchical or global structures carry biological significance.

To address these limitations, researchers have developed techniques and pipelines to improve the perceived quality of dimensionality reduction. For instance, Kobak and Berens [16] proposed modifications to t-SNE that better handle large-scale

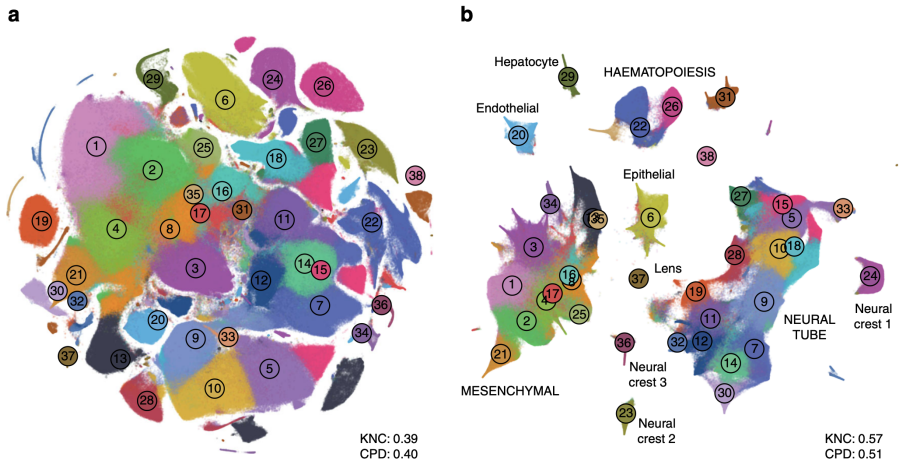


Figure 1.1: Improved t-SNE visualization for a large-scale scRNA-seq dataset (Cao et al.). The original t-SNE plot (a) lacks global structure, whereas the enhanced pipeline (b) preserves developmental trajectories and global relationships. [16]

datasets and global relationships, offering clearer and more reliable visualizations of scRNA-seq data (see Figure 1.1).

This example underscores a broader challenge: while DR methods offer valuable insights, they inherently distort some aspects of the data. Without robust metrics to assess the quality of dimensionality reduction, there is a risk of drawing inaccurate or incomplete conclusions. How can researchers ensure that patterns observed in lower-dimensional representations genuinely reflect the underlying structures of the original data?

The answer lies in the development of reliable quality metrics for dimensionality reduction. A good metric should evaluate how well a reduction method retains the meaningful features of the data while avoiding artifacts. These metrics are crucial not only for guiding the selection of DR techniques but also for validating the conclusions derived from them.

This thesis addresses this challenge by proposing a novel framework for assessing the quality of dimensionality reduction. Unlike traditional metrics that focus on local neighborhood preservation, this work introduces path-based metrics to evaluate global structures in the reduced space. By integrating both local and global perspectives, the proposed approach aims to enhance the trustworthiness and interpretability of dimensionality reduction across diverse datasets and applications.

1.1 Problem statement

Dimensionality reduction, by nature, abstracts and simplifies the data, but in doing so, it often sacrifices some level of detail and structure. This abstraction introduces the potential for misrepresentation, which can lead to inaccurate conclusions if the quality of the reduction is not properly quantified. To address this, it is essential to have effective and precise metrics for evaluating the quality of dimensionality reduction techniques.

Currently, one of the most commonly used approaches for evaluating the quality of dimensionality reduction is the $R_{\mathrm{NX}}(K)$ metric, which compares the neighborhoods of data points in high-dimensional space (HD) and their low-dimensional (LD) counterparts. This metric operates under the assumption that the relative ordering of neighbors should be preserved during the reduction. Specifically, it evaluates the preservation of neighborhood structure by creating expanding spheres around each data point and comparing the appearance of neighboring points between HD and LD representations.

While $R_{\mathrm{NX}}(K)$ serves as a useful tool, it presents a limitation: it assumes a purely isotropic approach to neighborhood comparison. In reality, when humans view low-dimensional data visualizations, they do not simply compare points by considering their proximity in a static manner. Instead, they follow the more intricate topological structures and paths that emerge in the data's representation, seeking patterns and relationships that might not be captured in a simple neighborhood comparison. This discrepancy between the metric's approach and human perception of data shapes presents a significant gap in the evaluation of dimensionality reduction.

To address this issue, the goal of this thesis is to propose a new quality assessment method that better aligns with how humans intuitively perceive data structures in low-dimensional spaces. By focusing on the logical paths between data points, rather than just their local neighborhoods, this new method aims to provide a more faithful evaluation of the embedding's quality. The method will compare paths between points in low-dimensional space with their corresponding structures in high-dimensional space, thus providing a more robust and human-centric approach to dimensionality reduction quality assessment.

In doing so, this approach seeks to improve the consistency and reliability of the evaluation process, making it more adaptable to the distortions often introduced by iterative methods such as t-SNE, UMAP, and MDS, which are commonly used in dimensionality reduction. These methods can produce variations in the representation due to their sensitivity to initialization and optimization steps, which, in turn, can affect traditional neighborhood-based quality measures like $R_{\mathrm{NX}}(K)$.

By accounting for the topological nature of the data in the low-dimensional embedding, the new method aims to ensure that the essential structural information is preserved and evaluated in a way that mirrors the human interpretation of the data. This could potentially improve the application of dimensionality reduction techniques in diverse fields where accurate and interpretable data visualization is crucial, such as in medical imaging, machine learning, and exploratory data analysis.

Chapter 2

Related work

This section reviews a variety of quality assessment metrics, arranged chronologically based on their introduction in the literature. Historically, many of these metrics have been closely associated with their respective dimensionality reduction techniques, often serving as the primary objectives for optimization during the reduction process. For many years, each technique had its own specialized metric to evaluate performance.

It was not until the early 2000s that the proliferation of different dimensionality reduction methods led to a recognition of the need for more comprehensive and universal metrics. As various techniques emerged and evolved, it became apparent that a global quantifier was necessary to provide a cohesive evaluation of the quality of the dimensionality reduction using different methods.

Table 2.1 summarizes the various quality assessment metrics that have been developed over the years. Each metric is identified by its name, the year of its introduction, the criterion it addresses and a references to it's original paper. It is important to note that not all methods presented in the table will be explored in this thesis; instead, the focus will be on a select few that are deemed most relevant to the current research.

Year	Name of the Measure	Criterion
1901	Explained variance [29]	Global
1962	Sheppard Diagram (SD) [32, 33]	Global
1964	Kruskal Stress Measure (S) [19, 20]	Global
1969	Sammon Stress (Ss) [30]	Global
1988	Spearman's Rho (SB) [34]	Local
1992	Topological Product (Tn-) [3]	Local
1997	Topological Function (Tf) [40]	Local
2000	Residual Variance (Rv) [35]	Global
2000	Konig's Measure (JCM) [18]	Local
2001	Trustworthiness & Continuity (T&C) [5]	Local
2003	Classification error rate [31] [38] [41]	classification error
2006	Local Continuity Meta-Criterion (Q_k) [6]	Local
2006	Agreement Rate (AB) / Corrected Agreement Rate (CAB) [10]	Local
2007	Mean Relative Rank Errors (MRRE) [21]	Local
2009	Procrustes Measure (PM) / Modified Procrustes Measure (PMC) [11]	Local
2009	Co-ranking Matrix (QNX and BNX) [24, 23]	Local
2011	Global Measure (G_y) [26]	Global
2011	The Relative Error (R_E) [14]	Local and Global
2012	Normalization independent embedding quality assessment (NIEQA) [37]	Local and Global
2013	Relative Quality $R_{\mathrm{NX}}(K)$ & AUC [25]	Local

Table 2.1: Summary of methods for evaluating the quality of DR algorithms, listed chronologically, this table is taken from [13].

2.1 Quality Assessment of Dimensionality Reduction Methods

As previously stated most of the early quality metric were associated with a specific dimensionality reduction method. this chapter gives a chronological rundown of a selection of these relevant metric and the evolution of the intuition about quality of dimensionality reduction.

2.1.1 Explained variance

The first intuition about the quality of dimensionality reduction comes from the metric maximized by the first algorithm developed for this purpose: Principal Component Analysis (PCA). Created in 1901 by Karl Pearson, PCA [29] aims to project data onto vectors, called principal components, that maximize the variance. By iterating multiple times and generating new principal components at each step, it becomes possible to represent the data using these components as unit vectors. To fully reconstruct the dataset, a sufficient number of these unit vectors is needed to explain the total variance of the data. Working backward from this concept, we can evaluate the quality of a reduction by the amount of variance explained by each of its components, which is closely related to the R^2 metric used in regression analysis.

In dimensionality reduction, R^2 represents the proportion of the total variance in the original data that is captured by the selected components. It indicates how well the lower-dimensional representation preserves the variability present in the full dataset. Mathematically, R^2 is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \ ,$$

where SSE (Sum of Squared Errors) is the sum of the squared differences between the observed values and the values predicted by the model, and SST (Total Sum of Squares) represents the total variability in the data. In other words, SST measures how much the observed data deviate from their mean, and SSE measures how much the predicted values deviate from the observed ones.

In Principal Component Analysis (PCA), this means that the higher the R^2 value, the more variance is captured by the selected components, indicating a more faithful lower-dimensional representation of the original data. This cumulative R^2 value provides insight into how well the principal components capture the total variance, offering a measure of the quality of the dimensionality reduction.

However, despite its widespread use, R^2 (or explained variance) has limitations when applied to dimensionality reduction. It only quantifies the fraction of variance captured, without considering the interpretability or practical significance of the components. A high R^2 value does not guarantee that the principal components are meaningful or that they capture relevant features for a particular application.

This can be demonstrated with the famous Anscombe's quartet. A set of 4 data sets with the same descriptive statistics, including r^2 with a value of 0.67, but that have a very distinct topological structure and a very different intuitive quality.

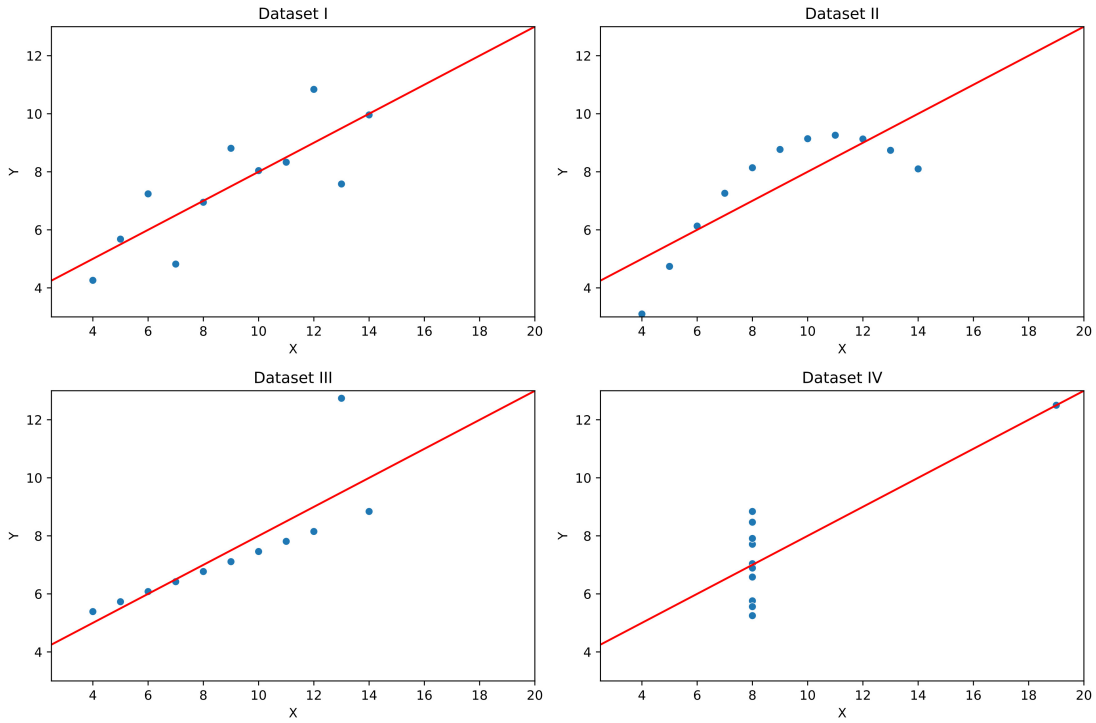


Figure 2.1: The four plots of Anscombe's Quartet.

C. H. Achen discusses this in his statement: "Thus, R^2 gives the 'percentage of variance explained' by the regression, an expression that, for most social scientists, is of doubtful meaning but great rhetorical value. If this number is large, the regression gives a good fit, and there is little point in searching for additional variables. Other regression equations on different data sets are said to be less satisfactory or less powerful if their R^2 is lower. Nothing about R^2 supports these claims" [1]. Next, after constructing an example where R^2 is enhanced just by jointly considering data from two different populations, Achen concludes: "Explained variance explains nothing" [2].

2.1.2 Stress functions

In [36], Warren S. Torgerson introduced a novel approach to dimensionality reduction based on pairwise distances between points in the dataset. This method, known as Nonlinear Multidimensional Scaling (MDS), utilizes the Stress metric to optimize the quality of dimensionality reduction. This method got improved further by J. B. Kruskal in [20] by introducing the specific version of the stress metric referred to stress-1 that has become the standard for this method.

Stress provides notable improvements over the explained variance metric used in Principal Component Analysis (PCA). While explained variance focuses on maximizing the variance captured by principal components, it often fails to preserve the pairwise distances or dissimilarities between data points, an aspect that can be crucial for accurately representing the data structure in many applications.

Stress measures the discrepancy between distances in the original high-dimensional space and distances in the reduced low-dimensional space. It is calculated as the square root of the normalized sum of squared differences between original distances d_{ij} and reduced space distances $\hat{d}_{ij}$, expressed mathematically as:

$$\text{Stress} = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}$$

Strain, a related metric, provides a similar measure but is often applied in various formulations or scaling techniques. Both metrics aim to minimize the difference between original and reduced space distances, thereby preserving the geometric relationships among data points as accurately as possible.

The primary advantage of Stress and Strain over explained variance lies in their emphasis on preserving the relative distances between data points rather than merely capturing overall variance. This focus makes them particularly effective for capturing the intrinsic structure of the data and maintaining meaningful relationships, which is essential for nonlinear dimensionality reduction where linear methods like PCA may not adequately represent the underlying structure of data.

The Stress metric, while useful, has several limitations. It is particularly sensitive to noise and outliers, which can introduce large discrepancies and distort the evaluation of dimensionality reduction quality. Additionally, Stress is highly influenced by the scaling of the graph, making it less effective for non-distance-based methods. The optimization process involved in minimizing Stress often converges to local minima, leading to suboptimal results and inconsistencies across different runs. This metric also presents computational challenges, as calculating and optimizing Stress can be resource-intensive, particularly for large datasets, which affects the scalability of the method. Moreover, the choice of a distance metric impacts Stress, and for high-dimensional datasets, the curse of dimensionality reduces the informativeness of distances between points, diminishing the effectiveness of distance-based methods of dimensionality reduction. MDS solutions are not always unique, complicating their interpretation and validation. Moreover, Stress does not account for the interpretability of the reduced dimensions, making it difficult to relate them to the original data features.

The limitations of the Stress metric as a quality assessment tool become evident when applied to different dimensionality reduction techniques. In an experiment using a 3D S-curve dataset, embeddings were generated through MDS, t-SNE, and Hessian Eigenmaps. The results revealed a clear contradiction between the Stress values and the visual quality of the embeddings: MDS, which fails to untangle the curve, yields a low Stress value of 0.12, while t-SNE and Hessian Eigenmaps, which successfully untangle the structure, yield higher Stress values of 0.26 and 0.28, respectively. Despite the higher Stress values, both t-SNE and Hessian Eigenmaps better preserve local relationships between data points. This demonstrates that Stress, as a distance-based metric, does not always reflect the true quality of embeddings, especially for non-linear methods. Thus, it is essential to adopt a more nuanced approach when evaluating dimensionality reductions, as a low Stress value does not always indicate a faithful reduction, while a higher Stress value may still correspond to a useful and meaningful low-dimensional representation.

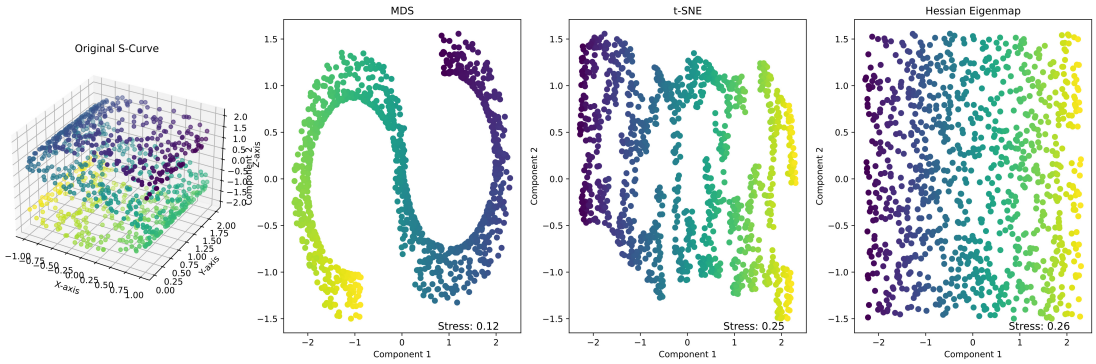


Figure 2.2: Stress for a S curve on different methods

2.1.3 Reconstruction Error

Reconstruction error is one of the most intuitive ways to evaluate the quality of dimensionality reduction. The principle involves calculating the sum of differences between the original data points and their reconstructions after applying the dimensionality reduction technique. Before 1991, the use of reconstruction error was limited because directly optimizing this metric was complex. Nonetheless, comparing the original data with its reconstruction—whether visually or through numerical measures—has long been a common method for assessing model performance.

In 1991, Kramer advanced this concept by introducing the autoencoder firstly as a nonlinear extension of Principal Component Analysis (PCA), which uses reconstruction error as its primary optimization metric. The neural network-based

structure of the autoencoders enabled the direct optimization of reconstruction error, treating it as the reward function to guide the model learning process.

Mathematically, if $X = \{x_1, \ldots, x_n\}$ represents the original data and $\hat{X} = \{\hat{x}_1, \ldots, \hat{x}_n\}$ denotes the reconstructed data, the reconstruction error can be computed using various norms, such as the Mean Squared Error (MSE):

$$\text{Reconstruction Error} = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2 \ .$$

The primary advantage of using the reconstruction error is that it provides a direct measure of how accurately the reduced-dimensional representation can reconstruct the original data. This metric is particularly useful for assessing the effectiveness of both linear and nonlinear dimensionality reduction algorithms, as it reflects the extent to which important information is preserved in the reduced space.

However, reconstruction error has its limitations. One of the most significant issues is that many dimensionality reduction methods lack an inverse transform function. For instance, methods that prioritize preserving relationships between points, such as neighborhood or distance (e.g., t-SNE, Isomap, LLE, MDS, SOM, ...), do not support reconstruction. As a result, reconstruction error cannot be calculated for these techniques. Even when it is possible, a lower reconstruction error merely indicates that the original data's details are better preserved, but this does not necessarily translate to the reduced representation being meaningful or useful for specific tasks. Furthermore, reconstruction error may fail to capture the interpretability or utility of the reduced dimensions. Additionally, it can be highly sensitive to noise and outliers, potentially distorting the assessment of the dimensionality reduction quality.

2.1.4 Topographic Product

In parallel to all those models, Teuvo Kohonen introduced the concept of self-organizing maps (SOM, [17]). A self-organizing map is an unsupervised machine learning technique that uses an initial grid representation of the data and a weight vector for each of the points. The goal of the SOM algorithm is to find – for each HD point – the best candidate, called Best Matching Unit (BMU) and have it's position slightly adjusted to match the input vector. The BMU is choosen by comparing the distance in HD between all weight vector and a specific input vector and choosing the lowest one:

$$d^V(\mathbf{w}_i, \mathbf{v}) = \min_{j \in A} d^V(\mathbf{w}_j, \mathbf{v}) \ .$$

Here, $\mathbf{v}$ represents the HD input vector from the original space $\mathbf{V}$, $\mathbf{w}_i$ represents the weight vector of neuron i in the grid, and A represents the set of neurons in the LD space. The function d^V denotes the distance in HD space.

After the BMU is chosen, all the weight vectors are updated with decreasing magnitude based on a neighborhood function. This function defines how much each neuron is updated, depending on its neighborhood distance from the BMU on the 2D grid:

$$\mathbf{w}_i(t+1) = \mathbf{w}i(t) + \alpha(t) \cdot h_{\mathrm{BMU},i}(t) \cdot (\mathbf{x}(t) - \mathbf{w}_i(t)) \quad .$$

This iterative process allows the grid to adjust and approximate the topological structure of the original HD data.

As for the previous method a metric to evaluate the quality of this reduction has been proposed. In [4], Hans-Ulrich Bauer, Klaus Pawelzik, and Theo Geisel propose the following. For each point j we compute the ratio between the distance of the point k and its k-th nearest neighbor in both the HD space and the LD space:

$$Q_1(j,k) = \frac{d^{\mathbf{V}}\left(\mathbf{w}_j, \mathbf{w}_{n_k^A(j)}\right)}{d^{\mathbf{V}}\left(\mathbf{w}_j, \mathbf{w}_{n_k^{\mathbf{v}}(j)}\right)} \quad ,$$

$$Q_2(j,k) = \frac{d^{A}\left(j, n_k^A(j)\right)}{d^{A}\left(j, n_k^{\mathbf{V}}(j)\right)} \quad .$$

The next step is to compute the geometric mean of these ratios:

$$P_3(j,k) = \left(\Pi_{l=1}^{k} Q_1(j,k) Q_2(j,k)\right)^{\frac{1}{2k}} \quad .$$

Further averaging over all nodes and neighborhood orders finally yields the topographic product

$$P = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{k=1}^{N-1} \log\left(P_3(j,k)\right) \quad .$$

While the original intent of the Topographic Product (TP) metric was to assess whether the selected low-dimensional (LD) space had an appropriate number of dimensions—where $P \leq 0$ indicates insufficient dimensionality, $P = 0$ suggests the dimensionality is adequate, and $P \geq 0$ implies too many dimensions – it can still be repurposed as a qualitative measure of dimensionality reduction. In the following analysis, we explore this potential application across different methods.

First, we applied a Self-Organizing Map (SOM) to an S-curve dataset. The resulting grid aligns well with the structure of the original data but does not perfectly capture it. There are some stray points and links from the initial grid that span across the

topographic structure, indicating room for improvement is left. Nonetheless, the calculated Topographic Product of -0.00536 suggests that the 2D representation is suitable for this dataset, a conclusion that aligns with the known properties of the S-curve.

To further investigate the effectiveness of TP as a direct evaluation method, we applied it to Hessian Eigenmaps, which produces a highly accurate manifold of the data. Since Hessian Eigenmap does not produce a grid-based representation like a SOM, we averaged the points into a grid of the same size as the SOM for consistency. This yielded a Topographic Product of -0.00156, a better score, reinforcing the intuition that more accurate manifolds result in values closer to zero.

However, when applied to t-SNE, the limitations of the Topographic Product become evident. t-SNE clusters points in its representation, leaving empty regions in the manifold. This clustering negatively impacts the TP score, even though t-SNE arguably provides a representation that is at least as good as a SOM's. The Topographic Product, therefore, appears highly sensitive to empty areas, which can lead to lower scores even for well-formed manifolds.

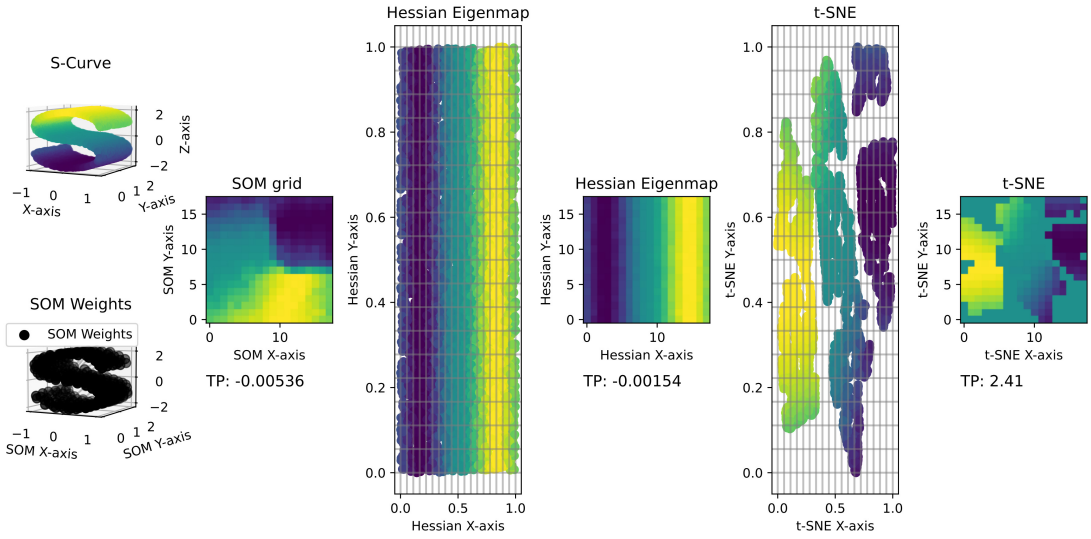


Figure 2.3: Topographic Product for an S-curve across different methods.

The Topographic Product is fundamentally tied to grid representations. Methods like t-SNE, which tend to cluster points in scatter plots, are disproportionately penalized, even when they preserve the overall structure of the data well. This suggests that while TP can be useful for assessing methods that output grid-like embeddings (such as a SOM), its applicability is more limited for more mainstream techniques like t-SNE.

Another major limitation of the Topographic Product is its reliance on distance computations between points. This dependence becomes especially problematic in high-dimensional spaces, where distances tend to lose their significance due to the curse of dimensionality and norm concentration. As a result, the method is restricted in its ability to accurately assess dimensionality reduction in such contexts.

However, the true innovation of the Topographic Product lies in its introduction of neighborhood relationships as a core component of the evaluation process. By shifting the focus from merely preserving pairwise distances to capturing local neighborhood structures, the Topographic Product offers a more refined approach to assessing the quality of dimensionality reduction. It enables the metric to account for both local and global topological features, which earlier distance-based metrics failed to capture. This evolution paved the way for future metrics that move beyond simple distance preservation, instead prioritizing topological coherence and neighborhood fidelity, ultimately providing a deeper understanding of how well a dimensionality reduction method preserves the intrinsic structure of the data.

2.2 Global Quality Assessment Metrics

The evaluation of dimensionality reduction techniques has evolved significantly over time, reflecting the growing complexity and scope of the field. Early approaches to quality assessment, such as explained variance, Stress, and reconstruction error, were tailored to specific methods. These metrics were designed to capture how well a particular algorithm retained important data characteristics, often focusing on localized aspects such as the preservation of variance or minimization of distortion within the reduced space. Although these metrics provided valuable information, they were inherently method-specific, limiting their broader applicability across different techniques.

As the field progressed, researchers recognized the need for more generalized, global approaches to quality assessment—methods that could evaluate the performance of dimensionality reduction algorithms more holistically, independently of the underlying method. This shift in thinking was driven by the realization that no single metric or algorithm could fully capture the complexity of high-dimensional data when projected into lower dimensions. Instead, researchers sought to develop metrics that assess not just the fidelity of individual methods but also the broader topological and structural preservation of the data across different dimensionality reduction techniques.

This period also coincided with a departure from distance-based comparisons, which became increasingly problematic in high-dimensional spaces due to the curse of

dimensionality. In such spaces, pairwise distances between points tend to lose carry less and less information as data points become more uniformly spaced, making it difficult to discern the true structure of the data. As a result, researchers began to favor metrics that focused on neighborhood preservation, which captures the local relationships between data points more effectively, even when global distances are distorted.

Global quality metrics, such as Trustworthiness and Continuity, the RAND Index, and the Local Continuity Meta-Criterion, reflect this paradigm shift. These metrics assess how well the reduced-dimensional representation retains both local and global data structures, offering a more comprehensive view of the quality of the dimensionality reduction process. They allow for meaningful comparisons across different algorithms by focusing on general properties of data embeddings, such as neighborhood preservation, clustering tendencies, and continuity in the mapping from high-dimensional to low-dimensional space.

In this section, we introduce these global quality assessment metrics, emphasizing their ability to transcend specific algorithms and provide a more holistic evaluation of dimensionality reduction methods. By focusing on metrics such as Quality $Q_{\mathrm{NX}}(K)$, Behavior $B_{\mathrm{NX}}(K)$, and Relative Quality $R_{\mathrm{NX}}(K)$ AUC, we aim to capture the broader aspects of dimensionality reduction quality, extending the analysis beyond individual method performance to a universal framework for quality assessment. This evolution marks a significant shift in the field, as the focus moves from assessing specific methods to evaluating the overall effectiveness of dimensionality reduction in preserving the intrinsic structure of the data.

2.2.1 Trustworthiness and Continuity

In [39], Jarkko Venna and Samuel Kaski introduced the concepts of *Trustworthiness* and *Continuity* to enhance Multi-Dimensional Scaling (MDS). Their work led to the creation of *Local MDS*, a novel dimensionality reduction technique aimed at optimizing the balance between these intrusion and extrusion. While the focus of their method was on algorithmic improvements, the metrics they introduced for evaluating the quality of dimensionality reduction remain central to this day.

The key insight in their work is that any dimensionality reduction method must strike a balance between two often competing goals:

- **Trustworthiness**: How well does the low-dimensional (LD) representation preserve local relationships found in the high-dimensional (HD) space?
- **Continuity**: To what extent does the low-dimensional space accurately reflect the original structure of the high-dimensional data?

To formalize these concepts, the authors introduced the following notations. Let N be the number of data samples, and let $r(i, j)$ represent the rank of data point j in the order of proximity to data point i in the original HD space. Define $U_k(i)$ as the set of data points that are within the k-nearest neighbors of point i in the LD space but not in the original HD space.

The concept of *Trustworthiness* is measured by ensuring that the k-nearest neighbors in the LD space also exist as neighbors in the HD space. If they do not, penalties are applied based on how far the misplaced points are ranked in the original space. This is expressed as:

$$T(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(i, j) - k) \quad .$$

Here, $r(i, j)$ denotes the rank of point j with respect to point i in the HD space, and $U_k(i)$ identifies the points incorrectly mapped as neighbors in the LD space. The sum quantifies the penalty based on how far these erroneous neighbors are from their true rank in HD, with $T(k)$ approaching 1 when the LD representation faithfully preserves the local structure of the HD space.

On the other hand, *Continuity* ensures that the structure of the HD space is adequately captured in the LD space. It measures whether the data points close to i in the HD space remain neighbors in the LD space. This metric is defined as:

$$C(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in V_i^{(k)}} (\hat{r}(i, j) - k) \quad ,$$

where $\hat{r}(i, j)$ denotes the rank of point j relative to i in the LD space, and $V_i^{(k)}$ refers to the set of points that are in the HD k-neighborhood but are not ranked correctly in the LD space. Like the trustworthiness measure, continuity also approaches 1 when the LD space perfectly mirrors the HD relationships.

To illustrate this trade-off let us consider the challenge of projecting a 3D spherical data structure into 2D space:

- When using **Principal Component Analysis (PCA)**, the 3D sphere is compressed along one of its axes, leading to a flattened representation. This method preserves continuity well, meaning points that were close in HD remain close in LD. However, it sacrifices trustworthiness by pulling together points that were originally distant in HD.
- On the other hand, **Curvilinear Component Analysis (CCA)** employs a cutting mechanism to flatten the sphere. This approach separates some points that were originally close in HD, thereby reducing continuity. However,

it improves trustworthiness by avoiding the clustering of points that were far apart in the original HD space.

These two strategies illustrate the inherent trade-off between preserving local neighborhood relationships (trustworthiness) and ensuring global continuity. An ideal dimensionality reduction method strives to balance both aspects, providing a faithful, yet informative representation of the original data.

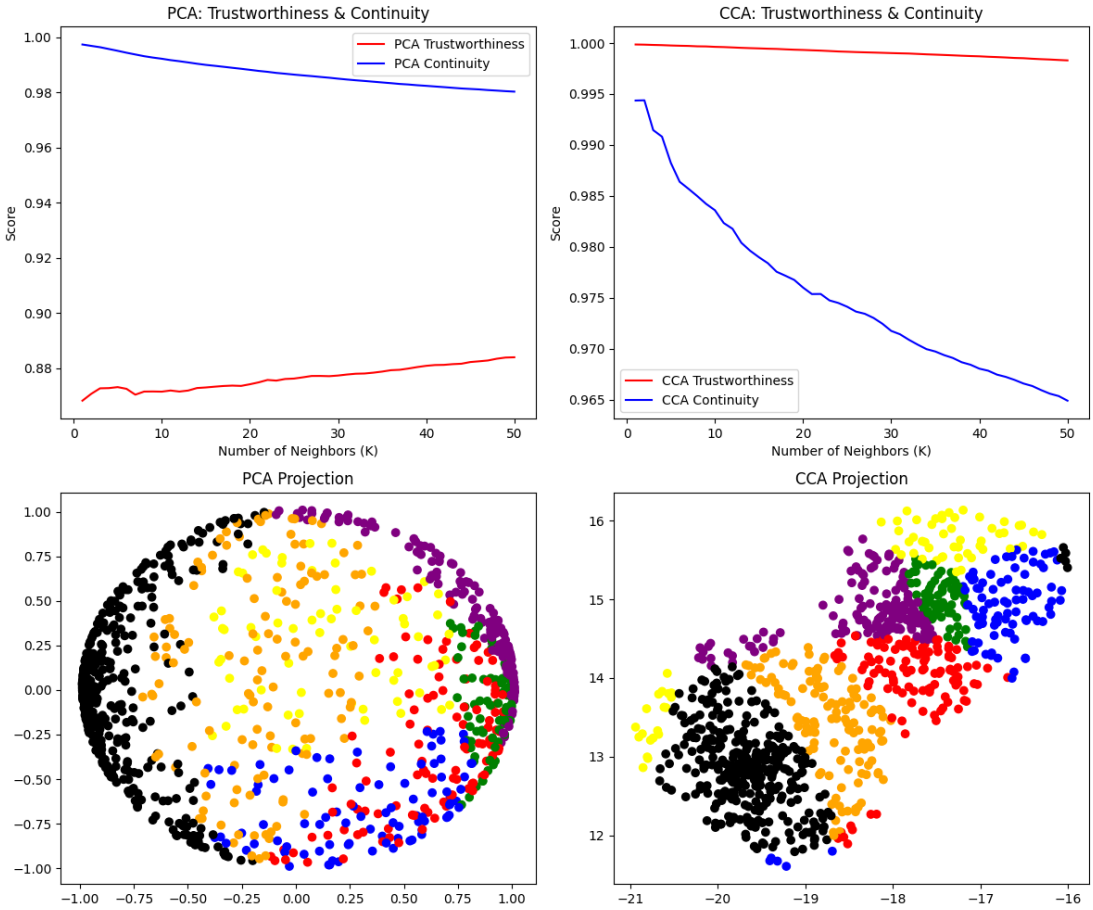


Figure 2.4: Trustworthiness and Continuity in the projection of a 3D sphere into 2D using PCA and CCA.

In the following methods, we will see a lot of these kind of plots with the score on the y axes and the neighboring range on the x axis. To output a finite score from these representations we simply compute the intergral of the score over the entire range.

2.2.2 Local Continuity Meta-Criterion

In [6], the authors Chen and Buja introduced the Local Continuity Meta-Criterion, which checks the degree of overlap between the neighboring sets of a data sample and their corresponding embeddings. This metric represents a pivotal advancement in the evaluation of dimensionality reduction (DR) techniques, as it is the first metric to directly utilize the degree of overlap between the neighboring sets of a data sample in high-dimensional (HD) space and their corresponding embeddings in low-dimensional (LD) space. By formalizing the concept of local continuity, Q_k provides a quantitative assessment of how faithfully the embedding captures local relationships found in the original data.

In the original paper [6], the Local Continuity Meta-Criterion is defined as $M_{K'}$, and it is calculated using the following equations:

$$N_{K'}(i) = \left|\mathcal{N}_{K'}^{D}(i) \cap \mathcal{N}_{K'}^{X}(i)\right|, \quad N_{K'} = \frac{1}{N}\sum_{i=1}^{N} N_{K'}(i) \ . \tag{2.1}$$

Here, $N_{K'}(i)$ denotes the number of neighbors of data point i that are shared between the high-dimensional space $\mathcal{N}_{K'}^{D}(i)$ and the low-dimensional space $\mathcal{N}_{K'}^{X}(i)$. Specifically, $\mathcal{N}_{K'}^{D}(i)$ represents the set of K' nearest neighbors of point i in the high-dimensional (HD) space, while $\mathcal{N}_{K'}^{X}(i)$ denotes the corresponding set of K' nearest neighbors in the low-dimensional (LD) embedding. The variable N is the total number of data points in the dataset.

The overall Local Continuity Meta-Criterion is defined as

$$M_{K'} = \frac{1}{K'} N_{K'} \ . \tag{2.2}$$

This formula calculates the average overlap of neighbors for each point in the dataset, normalized by the neighborhood size K'.

An adjusted version of this criterion is defined as follows,

$$M_{K'}^{\text{adj}} = M_{K'} - \frac{K'}{N-1} \ . \tag{2.3}$$

The adjustment accounts for the expected overlap due to chance. If there is no association between the data points and their embeddings, the overlap $N_{K'}(i)$ can be modeled as random, following a hypergeometric distribution. In this case, the expected number of overlaps, denoted as $E[N_{K'}]$, is given by

$$E[N_{K'}] = \frac{K'(K'-1)}{N-1} \ . \tag{2.4}$$

Here, K' represents the number of neighbors drawn from the original dataset, and N is the total number of data points. This expectation represents the number of overlaps that would occur by random chance when drawing K' neighbors from the entire dataset of $N-1$ items. By subtracting this expected value from $M_{K'}$, the adjusted criterion $M_{K'}^{\text{adj}}$ provides a more accurate reflection of the true local continuity by mitigating the influence of random overlap.

In contemporary literature, this criterion is often referred to as Q_k and is defined as follows,

$$Q_k = 1 - \frac{1}{nk} \sum_{i=1}^{n} |\Psi_k^{\chi}(i) \cap \Psi_k^{y}(i)| - \frac{k^2}{n-1} \quad . \tag{2.5}$$

In this expression, Q_k measures the local continuity by evaluating the overlap between the k-nearest neighbors in the original data space ($\Psi_k^{\chi}(i)$) and those in the low-dimensional embedding ($\Psi_k^{y}(i)$). The variables n and k represent the total number of data samples and the size of the neighborhood, respectively. The metric Q_k yields values in the interval $[0, 1]$, where values close to 1 indicate a high degree of neighborhood overlap between the two dimensional spaces, while values close to 0 signify the opposite.

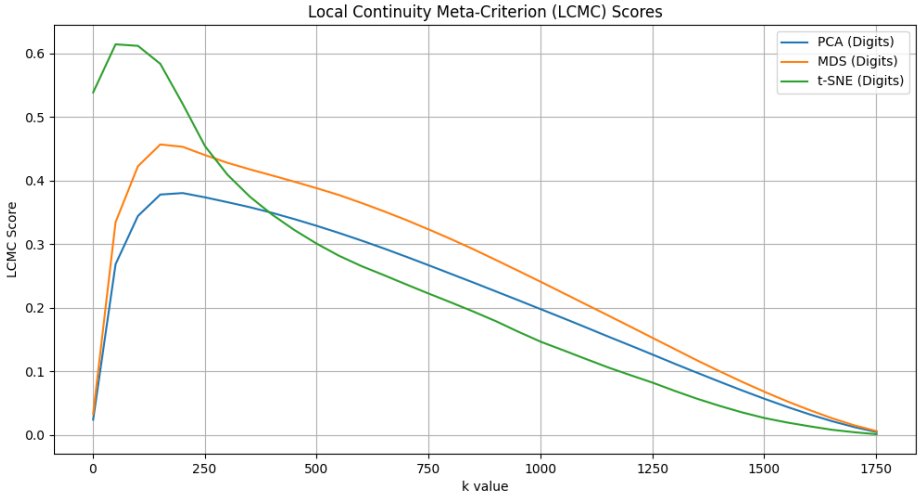


Figure 2.5: Example of LCMC on three different embeddings of the digits dataset.

2.2.3 Quality $Q_{\text{NX}}(K)$, Behavior $B_{\text{NX}}(K)$

In 2008, John A. Lee and Michel Verleysen [22] introduced the Co-ranking Matrix as a foundational tool for assessing quality metrics in dimensionality reduction

through rank-based methodologies. They demonstrated that earlier approaches could be redefined within the context of the Co-ranking Matrix, thereby proposing a unified framework for evaluating K-ary neighborhoods. This innovative framework leverages the Co-ranking Matrix representation of dimensionality reduction and introduces the metrics $Q_{NX}(K)$ and $B_{NX}(K)$.

The Co-ranking Matrix

This section is greatly inspired by the work of Lueks et al. [27]. The Co-ranking Framework is based on the Co-ranking matrix, which quantifies the preservation of rank order among neighbors during dimensionality reduction. In the Co-ranking matrix, the value at position (i, j) represents the number of times that the i-th neighbor in the high-dimensional (HD) space receives a different ranking j in the low-dimensional (LD) space. Specifically, if $i = j$, we have a perfect match; if $i > j$, we encounter an intrusion, indicating that a less relevant point is incorrectly moved closer; conversely, if $i < j$, we have an extrusion, where a relevant point is pushed further away.

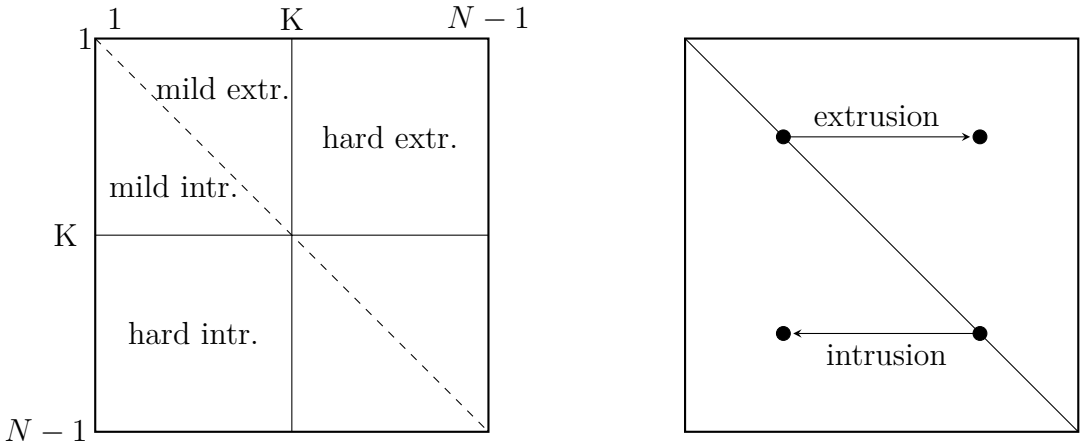


Figure 2.6: Large-scale structure of the co-ranking matrix. On the left, the matrix is split into blocks to show different types of intrusions and extrusions. In a perfect mapping, the co-ranking matrix will be a diagonal matrix. The image on the right shows how rank differences will alter the matrix. If a neighbor moves further away in the low-dimensional space, an extrusion, it will move mass to the right of the diagonal. Similarly, intrusions move mass to the left of the diagonal.

To define precicely what is co-ranking matrix let δ_{ij} be the distance from the i-th data point ξ_i to the j-th data point ξ_j in the high-dimensional space. Similarly, d_{ij} denotes the distance from x_i to x_j in the low-dimensional space. The ranks of the

neighbors for each point can be computed based on these distances. The rank of ξ_j with respect to ξ_i in the high-dimensional space is defined as

$$\rho_{ij} = |\{k \mid \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leq k < j \leq N)\}| \,.\ .$$

This results in ranks $\{1, \ldots, N-1\}$. Analogously, the rank of x_j with respect to x_i in the low-dimensional space is given by

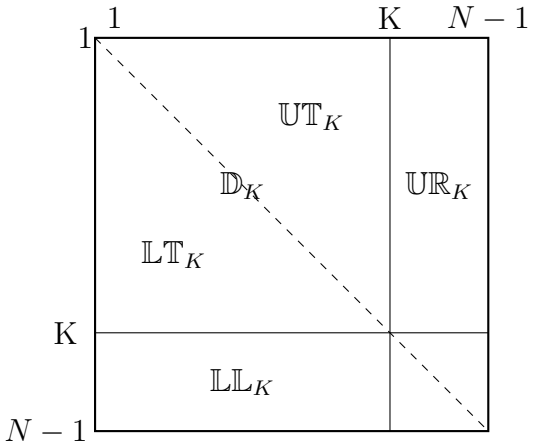
$$r_{ij} = |\{k \mid d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq N)\}| \,.$$

Using this notation, the co-ranking matrix Q is defined as

$$Q_{kl} = |\{(i,j) \mid \rho_{ij} = k \text{ and } r_{ij} = l\}| \,.\ .$$

The various types of intrusions and extrusions are associated with different blocks of the co-ranking matrix ,$\mathbb{D}_K$ defines the elements on the diagonal up to K, Hard K-intrusions and K-extrusions are found in the blocks $\mathbb{LL}_K$ and $\mathbb{UR}_K$, respectively. In a similar way, mild K-intrusions and K-extrusions are counted in the triangles $\mathbb{LT}_K$ and $\mathbb{UT}_K$, respectively.

In these different zones we define the following quantities. These quantities correspond to the fraction of points that keep their rank ($U_P(K)$) , the fraction of mild K-intrusions ($U_N(K)$) and the fraction of mild K-extrusions ($U_X(K)$).



$$U_N(K) = \frac{1}{KN} \sum_{(k,l)\in \mathbb{UT}_K} q_{kl}$$

$$U_X(K) = \frac{1}{KN} \sum_{(k,l)\in \mathbb{LT}_K} q_{kl}$$

$$U_P(K) = \frac{1}{KN} \sum_{(k,l)\in \mathbb{D}_K} q_{kl}$$

Existing quality metrics defined using the Co-ranking Matrix

It is possible to re define most of the rank based method using the Co-ranking Matrix.

The trustworthiness and continuity (T&C) measures [39] are defined as:

$$M_{\mathrm{T}}(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in n_i^K \setminus \nu_i^K} (\rho_{ij} - K) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathbb{LL}_K} (k - K) q_{kl} \ ,$$

$$M_{\mathrm{C}}(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K \setminus n_i^K} (r_{ij} - K) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathbb{U}_K} (l - K) q_{kl} \ ,$$

where the normalizing factor

$$G_K = \begin{cases} NK(2N - 3K - 1) & \text{if } K < N/2 \\ N(N-K)(N-K-1) & \text{if } K \geq N/2 \end{cases} \ .$$

The local continuity meta-criterion [6] (LCMC) is defined as

$$\mathrm{LCMC}(K) = \frac{1}{NK} \sum_{i=1}^{N} \left(\left| n_i^K \cap \nu_i^K \right| - \frac{K^2}{N-1} \right) = \frac{K}{1-N} + \frac{1}{NK} \sum_{(k,l) \in \mathbb{U}_K} q_{kl} \ .$$

$\mathbf{Q_{NX}(K)}$

Based on this setting, a simple quality measure can be defined: it counts the number of points that remain inside the K-neighborhood while projecting. In other words, all points in $\mathbb{UT}_K$, $\mathbb{LT}_K$ and $\mathbb{D}_K$

$$Q_{\mathrm{NX}}(K) = U_N(K) + U_X(K) + U_P(K) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{l=1}^{K} Q_{kl} \ .$$

The quality criterion is very similar to the local continuity meta-criterion (LCMC) that was proposed by Chen and Buja [6]. In fact, it coincides up to the linear term that accounts for the quality of a random mapping, namely,

$$\mathrm{LCMC}(K) = Q_{\mathrm{NX}}(K) - \frac{K}{N-1} \ .$$

BNX(K)

Using the unified framework, we can define a new quantity that represents the tendency of the reduction to favorize either extrusive or intrusion by taking the difference between the fraction of mild K-intrusions and the fraction of mild K-extrusions

$$B_{\mathrm{NX}}(K) = U_N(K) - U_X(K) \ .$$

$\mathbf{R_{NX}(K)}$

In 2013, Lee and Verleysen introduced a renormalized version of the quality measure $Q_{\mathrm{NX}}(K)$. They noted that for a random embedding, the expected value of $Q_{\mathrm{NX}}(K)$ is approximately $\frac{K}{N-1}$. The goal of this normalization is to adjust the value of $Q_{\mathrm{NX}}(K)$ so that a random embedding would yield a value of 0. This led to the formula

$$R_{NX}(K) = \frac{(N-1)Q_{\mathrm{NX}}(K) - K}{(N-1-K)} \quad .$$

2.2.4 Limitations of neighborhood based approach

The principal limitations of the metrics that are based on neighborhood is in the creation of the neighborhood itself. As the dimensionality reduction method is applied, it creates a manifold that follows the HD structure of the data and tries to recreate it in an LD space. But when constructing the neighborhood for the evaluation of $R_{\mathrm{NX}}(K)$ we select the neighborhood in an isentropic manner around each point, completely neglecting the embedding of the data.

To avoid this limitation we try to create a new metric that takes into account the shape of the embedding by checking if low-dimensional shortest path in the data between two points create a coherent path in the HD space, meaning a path where the order of the neighbor for the subset of point is the same for HD and LD.

2.3 Supporting Methods and Techniques

This section presents a selection of established methods that serve as the foundation for the new approach introduced in this thesis. These techniques are well-documented and extensively studied in the literature, providing a solid theoretical framework for the development and evaluation of the proposed method.

2.3.1 Delaunay triangulation

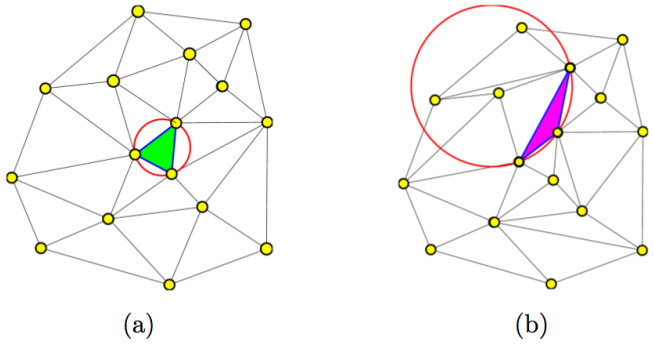


Figure 2.7: Delaunay triangulation of a point set (a) and a triangulation that is not Delaunay (b). In (b), the colored triangle is not empty.

The notion of triangulating a set of points $P = \{p_1, \ldots, p_n\}$ in dimension d is perfectly defined. Triangulating consists of covering the convex hull $H(S)$ of S with simplices – a simplex is the simplest possible polyhedron in dimension d, a triangles for $d = 2$, a tetrahedron for $d = 3$... There is a combinatorial number of possible triangulations for a given S. A simplex t of a triangulation $T(S)$ is said to be locally Delaunay if its circumsphere is empty, that is, if it does not contain points of S. A triangulation T is said to be a Delaunay triangulation if every simplex is empty (see Figure 2.7). If we consider points in a general position – $d + 1$ points of the set are not in a hyperplane and $d + 2$ points of the set are not on the same hypersphere – the Delaunay triangulation is unique. The Delaunay triangulation $\mathrm{DT}(S)$, introduced by Boris Delaunay in 1934 [7], is a cornerstone technique in computational geometry.

In dimension 2, $\mathrm{DT}(S)$ is endowed with the min-max property: among all triangulations, the minimum angle of all triangles of $\mathrm{DT}(S)$ is larger than any other triangulations of S. In higher dimensions, Delaunay triangulations do not generally have any specific geometric optimality properties. In general, Delaunay triangulation $\mathrm{DT}(S)$ tends to connect points of S that are close together, leading to the

possibility of computing the "shape" of a point cloud (see the forthcoming section §2.3.2 on α-shapes).

The computational complexity of constructing a Delaunay triangulation is significantly dependent on the dimension d. In small dimensions ($d < 4$) the time complexity can be as small as $\mathbb{O}(n \log n)$. For higher dimensions, the complexity grows to $\mathbb{O}(n^{\lceil d/2 \rceil})$ [12], reflecting an exponential increase with respect to the dimension. This rapid growth in complexity renders the approach infeasible for high-dimensional data.

Due to this limitation, we restrict our use of Delaunay triangulation to low-dimensional spaces in this thesis. For high-dimensional spaces, alternative methods will be explored to compare paths, circumventing the computational challenges posed by the exponential scaling of Delaunay triangulation.

2.3.2 Alpha Shapes

Alpha shapes, introduced by Edelsbrunner et al. [9], are a generalization of the convex hull that provide a flexible way to capture the shape of a point set in d -dimensional space. By varying a parameter α, alpha shapes interpolate between the convex hull of the set of points (for large α) and a set of discrete points (for small α). This adaptability makes alpha shapes particularly useful in applications that require a detailed representation of the underlying structure of the data, such as shape analysis and topological data analysis.

An alpha shape is derived from the Delaunay triangulation of the point set. Specifically, each simplex (triangle in 2D, tetrahedron in 3D, etc.) in the Delaunay triangulation is included in the alpha shape if its circumsphere has a radius less than or equal to α. The parameter α thus controls the level of detail in the representation: smaller values of α lead to more detailed shapes, while larger values simplify the shape by removing smaller features.

The computational complexity of constructing alpha shapes is closely tied to that of Delaunay triangulation, as the latter serves as the foundation for identifying valid simplices. In d -dimensional space, the complexity of computing the Delaunay triangulation is $\mathbb{O}(n^{\lceil d/2 \rceil})$, making alpha shapes computationally expensive in high-dimensional spaces [12].

In this thesis, we employ alpha shapes as a tool to define more finely the convex hull of the embedding. By adjusting the α parameter, we can capture geometric structures within the embedding that are not discernible from the convex hull alone.

Figure 2.8 shows the example of α-shapes. Figure 2.8(a) shows the set of points S that correspond to $\alpha = 0$. Figure 2.8(b) shows the Delaunay triangulation that

corresponds to $\alpha = \infty$. Figure 2.8(b) and 2.8(b) and shows α-shapes for $\alpha = 50$ and $\alpha = 200$. A given choice of α thus gives a given "shape" of the point cloud.

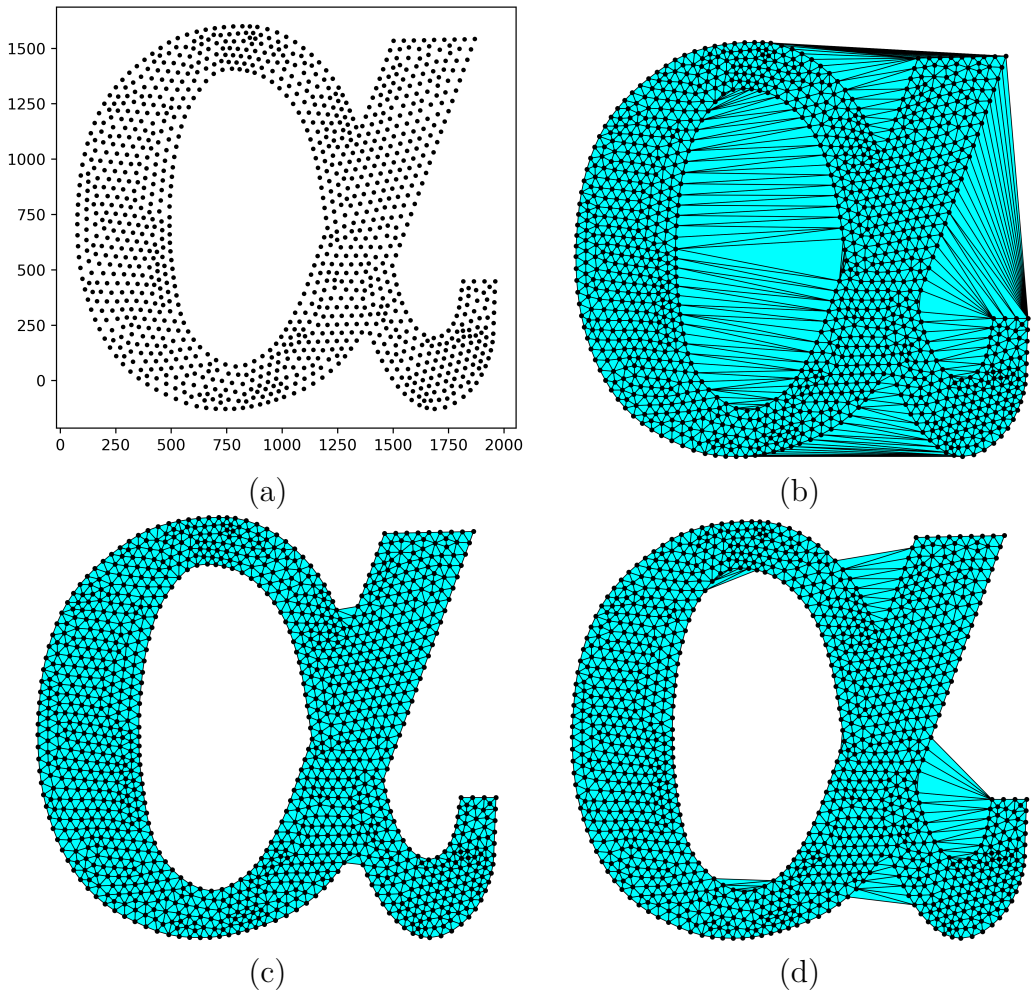


Figure 2.8: Set of points S (a), Delaunay triangulation DT(S) (b), α-shape for α=50 (c) et α-shape for α=200 (c).

2.3.3 Levenshtein Distance

The Levenshtein distance, also known as the edit distance, is a metric used to measure the difference between two sequences. This distance is computed as the minimum number of operations required to transform one sequence into another. The allowed operations are insertion, deletion, and substitution of characters. The Levenshtein distance is widely applied in fields such as computational biology, natural language processing, and data analysis, particularly when dealing with

strings or sequences where exact alignment is not guaranteed.

Mathematically, the Levenshtein distance $d(A, B)$ is computed using a dynamic programming approach, which creates a matrix in which each element represents the minimum number of operations needed to transform the substring of A up to position i into the substring of B up to position j. The recurrence relation for the Levenshtein distance is as follows:

$$d(i,j) = \min \begin{cases} d(i-1,j) + 1 & \text{(deletion)} \\ d(i,j-1) + 1 & \text{(insertion)} \\ d(i-1,j-1) + \text{cost} & \text{(substitution, where cost is 0 if } a_i = b_j \text{, else 1)} \end{cases}$$

For example, if we take the two sequences "123456" and "132654", the Levenshtein distance will create the table shown in Figure 2.9. This table concludes that the minimum number of operations is 4. These operations consist of 1 insertion, 1 deletion, and 2 substitutions.

		1	**3**	**2**	**6**	**5**	**4**
	0	1	2	3	4	5	6
1	1	0	1	2	3	4	5
2	2	1	1	1	2	3	4
3	3	2	1	2	2	3	4
4	4	3	2	2	3	3	3
5	5	4	3	3	3	3	4
6	6	5	4	4	3	4	**4**

Figure 2.9: Example of Levenshtein distance

Chapter 3

Quality Metric Derived from Shortest Paths in two Dimensional Space

The core idea behind the proposed method is that an effective low-dimensional representation should preserve the structural integrity of the high-dimensional (HD) space. This structure can be visualized as a "wireframe" in HD space, representing the network of shortest paths connecting the data points. By comparing these paths in the reduced, low-dimensional space with their counterparts in HD space, we expect that high-quality dimensionality reduction techniques will yield similar path structures, whereas poorly performing methods will produce inconsistencies.

In this thesis, we focus on reduction to 2D spaces, a widely used target dimensionality in data visualization that leverages the interpretability and practical advantages of 2D space. The goal of the proposed method is to compute the shortest paths between points in the 2D representation and compare them to paths in HD space. This approach has several benefits. First, computing paths in 2D is computationally manageable, as the shortest path problem is well-understood and can be efficiently solved with various algorithms. Second, it reduces the computational complexity in HD space by focusing only on points that lie along the 2D path, avoiding unnecessary comparisons and enabling a targeted examination of structural coherence.

However, this method also faces some limitations. To fully compare paths, we would need to calculate the shortest path between every pair of points, which scales poorly with the dataset size, requiring ($O(n^2)$) operations. While this can be mitigated, such workarounds may reduce the precision of our measure, a trade-off that will be discussed in detail in later sections.

Our expectations align with recent metrics for dimensionality reduction quality, anticipating results that reflect the known properties of specific techniques and the strengths of each in preserving different aspects of data structure.

3.1 Expected Behaviors

The methodology used to evaluate the effectiveness of the proposed approach is based on comparing the results against the expected behaviors exhibited by dimensionality reduction techniques. Specifically, this comparison hinges on the prioritization of local or global structure preservation inherent to each technique.

3.1.1 The 3D S-Curve Example

Consider the 3D S-curve depicted in Figure 3.1. This example provides an intuitive understanding of the trade-off between local and global structure preservation. Let us examine two specific distances:

1. The distance between the dark blue region at the beginning of the S-curve and the yellow region directly beneath it, following the curvature of the S.

2. The distance between the same dark blue region and the very light blue region on the opposite side of the S-curve.

Techniques that aim to preserve global structure will map the yellow region closer to the dark blue region compared to the light blue region. This is because global preservation prioritizes the true Euclidean distances in the original 3D space. However, in doing so, these methods may fail to maintain the continuity of the local structures (represented by the color transitions along the curve).

On the other hand, techniques designed to preserve local structure emphasize the manifold's continuity. Such techniques would embed the light blue region closer to the dark blue region than the yellow region, as the local path along the curve connects them more directly. In this process, the global geometry of the S-curve may be distorted, often reducing the S shape into a flattened band.

3.1.2 Trade-off in Dimensionality Reduction Techniques

This fundamental trade-off between local and global structure is a defining characteristic of dimensionality reduction methods. Each technique prioritizes one over the other or attempts to balance both based on its optimization objectives.

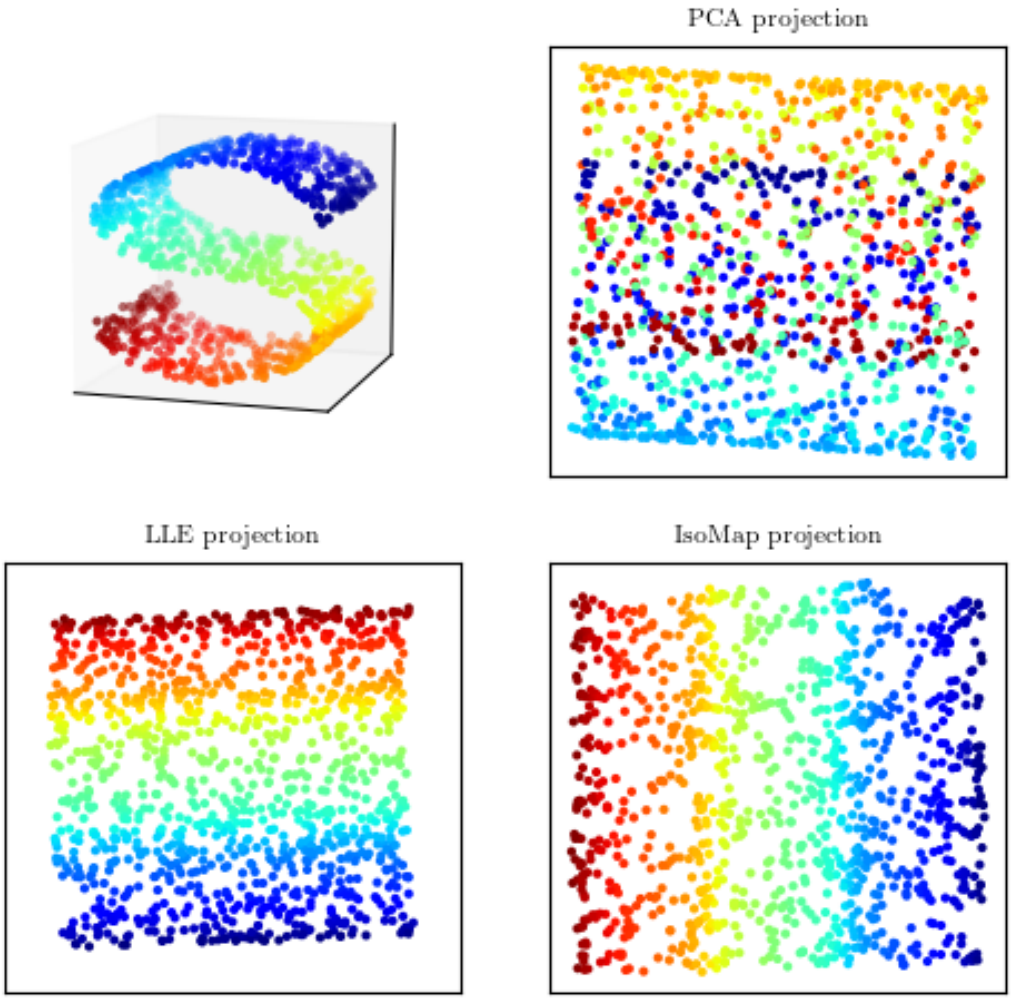


Figure 3.1: Illustration of the 3D S-curve dataset, highlighting regions of interest for analyzing local and global structure preservation.

Technique	**Tendency to Preserve Local Structure**	**Tendency to Preserve Global Structure**	**Description**
PCA	Low	High	Primarily focuses on maximizing variance along principal components, preserving global structure well but not ideal for local neighborhood relationships.
t-SNE	High	Low	Optimized to maintain local relationships and neighborhood structure, often at the cost of distorting global structure.
UMAP	High	Low	Focuses on preserving local neighborhoods similarly to *t*-SNE but with improved scalability and often better global structure than *t*-SNE.
Isomap	Mid	Mid	Aims to preserve both local and global structure by maintaining geodesic distances, effective for nonlinear manifolds.
MDS	Low (for metric MDS)	High	Maintains pairwise distances, capturing global structure; however, local relationships can be less accurately preserved.
SOMs	High	Low	Preserves local relationships by clustering similar data points on a grid, though global structure can be distorted.

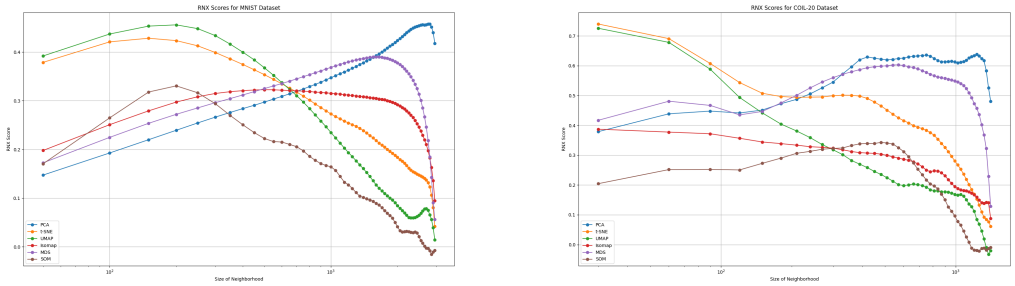
Table 3.1: Common Dimensionality Reduction Techniques and Their Structure Preservation Tendencies.

3.1.3 Implications for Evaluation Metrics

Structural preferences can be visualized using metrics like $R_{\text{NX}}(K)$. For example:

- Techniques prioritizing local structure (e.g., t-SNE, UMAP) achieve higher $R_{\text{NX}}(K)$ scores at smaller neighborhood sizes, reflecting accurate local preservation.

- Techniques emphasizing global structure (e.g., PCA, MDS) perform better at larger neighborhood sizes, where overall geometry is preserved.



(a) $R_{\text{NX}}(K)$ on MNIST dataset.

(b) $R_{\text{NX}}(K)$ on COIL-20 dataset.

Figure 3.2: $R_{\text{NX}}(K)$ plots for six different embeddings on two different datasets: (a) MNIST and (b) COIL-20.

3.2 General implementation

3.2.1 Graph Construction in LD Space

To capture the geometric structure of the reduced data, a graph is constructed in the LD space using Delaunay triangulation. The edges of this graph are weighted by the pairwise distances in the LD space. The adjacency matrix of the graph, denoted as $G \in \mathbb{R}^{n \times n}$, is initialized such that:

$$G_{ij} = \begin{cases} d_{LD}(x_i, x_j), & \text{if } (x_i, x_j) \text{ is an edge in the graph,} \\ 0, & \text{if } i = j, \\ \infty, & \text{otherwise,} \end{cases}$$

where $d_{LD}(x_i, x_j)$ represents the Euclidean distance between points x_i and x_j in the LD space. Self-loops are explicitly removed.

3.2.2 Path Computation

To quantify connectivity, we compute the shortest paths between all pairs of points in the LD graph using Dijkstra's algorithm. Let P_{ij}^{LD} represent the sequence of

points forming the shortest path between x_i and x_j in the LD space. Each path is subsequently mapped to its corresponding points in the HD space, yielding P_{ij}^{HD}. The algorithm ensures efficient computation even for large datasets.

3.2.3 Sorting and Comparison of Points

For each path P_{ij}^{LD}, the subset of points constituting the path is sorted based on their distances from the origin x_i in both LD and HD spaces:

$$S_{ij}^{LD} = \text{Sort}(P_{ij}^{LD}, \text{by } d_{LD}(x_i, \cdot)),$$

$$S_{ij}^{HD} = \text{Sort}(P_{ij}^{HD}, \text{by } d_{HD}(x_i, \cdot)).$$

3.2.4 datasets used as example

To illustrate the effectiveness of the method for evaluating dimensionality reduction, we use two well-known datasets: MNIST [8] and COIL-20 [28], to which we have applied five different dimensionality reduction methods: PCA, t-SNE, UMAP, Isomap, and MDS. These datasets serve as representative examples and provide a broad view of how the edit distance-based metric can capture the preservation of structure across different types of data.

MNIST

The MNIST dataset, composed of images of handwritten digits, serves as a benchmark for evaluating dimensionality reduction methods. Below, we visualize the embeddings generated by each method and their corresponding graphs constructed using Delaunay triangulation.

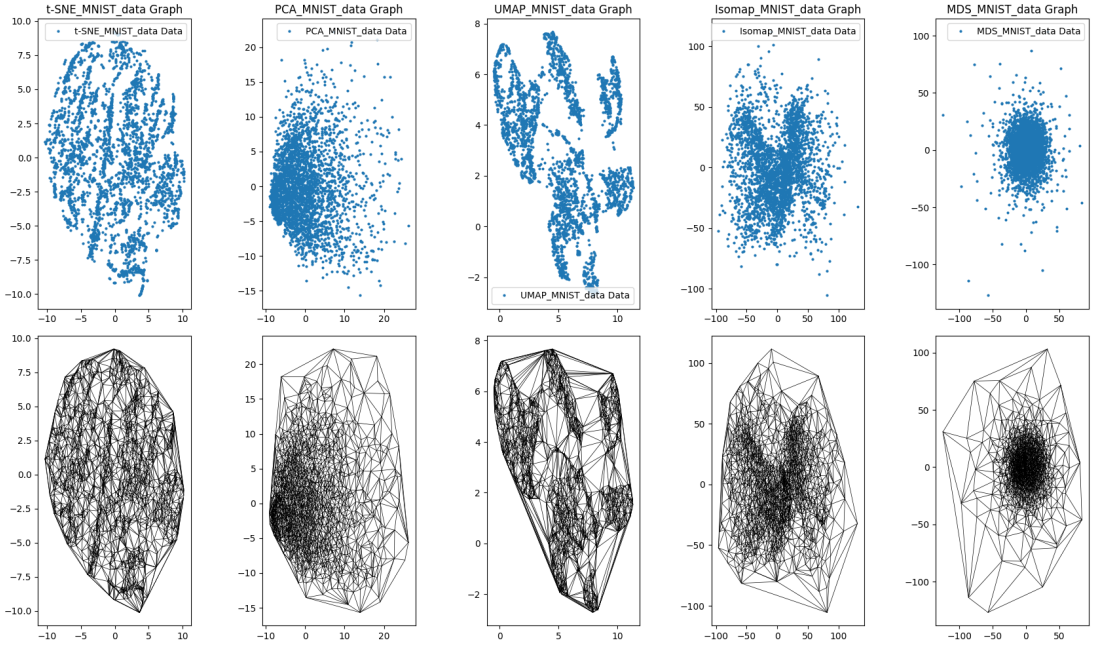


Figure 3.3: Embeddings of the MNIST dataset and their associated Delaunay triangulations for PCA, *t*-SNE, UMAP, Isomap, and MDS.

COIL-20

The COIL-20 dataset, consisting of 3D images of 20 different objects, presents a more complex challenge for dimensionality reduction due to its varied structures and higher levels of intra-class variation. For this dataset, we observe that the expected behaviors do not hold as clearly as they do with MNIST.

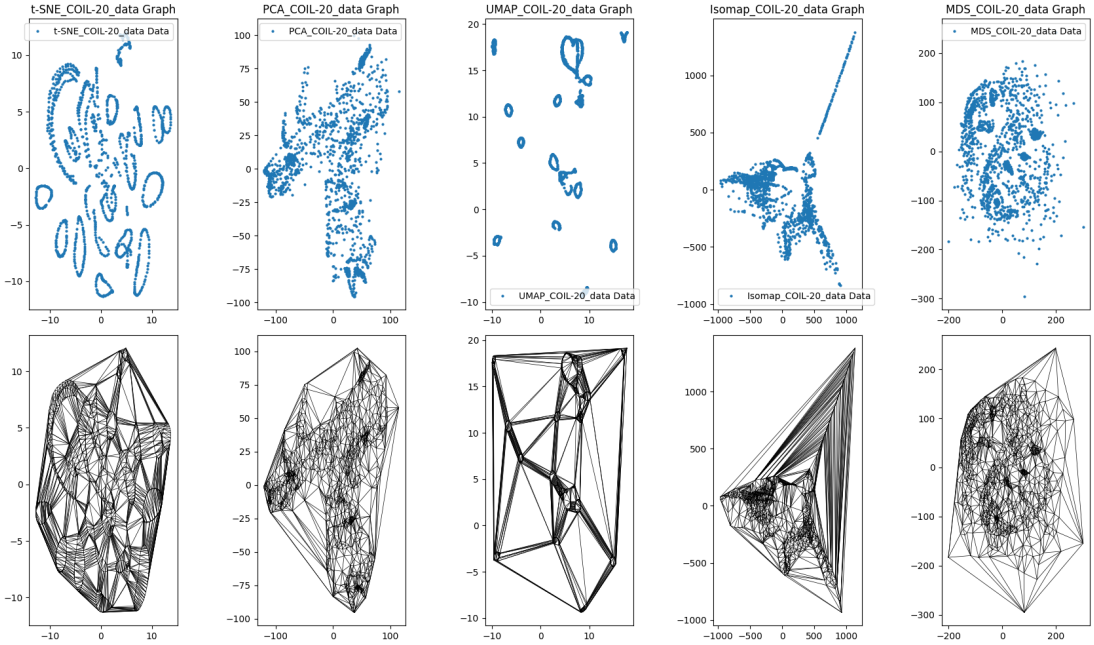


Figure 3.4: Embeddings of the COIL-20 dataset and their associated Delaunay triangulations for PCA, t-SNE, UMAP, Isomap, and MDS.

3.3 $R_{NX}(K)$ on Paths

This implementation introduces directionality into the computation of $R_{NX}(K)$ by applying its logic to multiple subsets of the data, which are determined through path computation. The aim of this approach is to assess the viability and effectiveness of a path-based evaluation framework for dimensionality reduction quality. By incorporating directionality, we focus on evaluating the preservation of structure along specific paths in the low-dimensional space, offering a more nuanced understanding of how the data is represented. This path-based method is expected to provide a more accurate and stable measure of neighborhood preservation, especially in datasets that may exhibit distortions or non-linearities in their low-dimensional embeddings.

3.3.1 Metric Aggregation

The overlap between LD and HD subsets is quantified for each n. The ratio of the intersection size to n is calculated:

$$r_{ij}^{n} = \frac{|S_{ij}^{LD}(1:n) \cap S_{ij}^{HD}(1:n)|}{n} \quad .$$

To aggregate the metric, the results are averaged over all paths:

$$\bar{r}^n = \frac{1}{|P|} \sum_{(i,j)\in P} r_{ij}^n \ ,$$

where P represents the set of all computed paths. To account for random effects, the expected result from a random embedding is subtracted:

$$\text{Adjusted } \bar{r}^n = \bar{r}^n - \mathbb{E}[r_{ij}^n \text{ (random)}] \ .$$

3.3.2 Implementation Details

The implementation was performed in Python using libraries such as `numpy`, `scipy`, and `networkx`. The following snippet demonstrates the core logic for computing and comparing paths:

```
for i in range(len(LD_data)):
    LD_paths, HD_paths = LDHDPathAll(graph, distance_matrix, i,
    LD_data, HD_data)
    for j in range(i + 1, len(LD_paths)):
        LD_path = LD_paths[j]
        HD_path = HD_paths[j]
        # Remove self-loops and sort by distance
        LD_sorted = sorted(LD_path[1:], key=lambda p:
    LD_distance_matrix[i][p])
        HD_sorted = sorted(HD_path[1:], key=lambda p:
    distance_matrix[i][p])
        # Compare top n elements
        for n in range(1, len(LD_sorted)):
            union_count = len(set(LD_sorted[:n]).intersection(set(
    HD_sorted[:n])))
            if n not in results:
                results[n] = [union_count / n]
            else:
                results[n].append(union_count / n)
```

Listing 3.1: Python implementation for path analysis

3.3.3 Comparison and Results

In this section, we evaluate the new quality metric using dimensionality reduction techniques on the MNIST and COIL-20 datasets. The methods considered include PCA, t-SNE, UMAP, Isomap, and MDS.

MNIST Dataset

The new quality metric successfully displays the expected behaviors of different dimensionality reduction techniques. Specifically:

- - Methods that prioritize local structure preservation, such as t-SNE and UMAP, show higher scores when evaluating shorter subsets of paths. This reflects their ability to maintain neighborhood relationships and local manifold continuity.
- - Techniques like PCA and MDS, which focus on global structure preservation, achieve better scores on longer path subsets. These methods are more adept at capturing overall dataset geometry and large-scale relationships.

This behavior aligns with the inherent optimization objectives of each technique and demonstrates that the new metric effectively distinguishes between local and global structure tendencies.

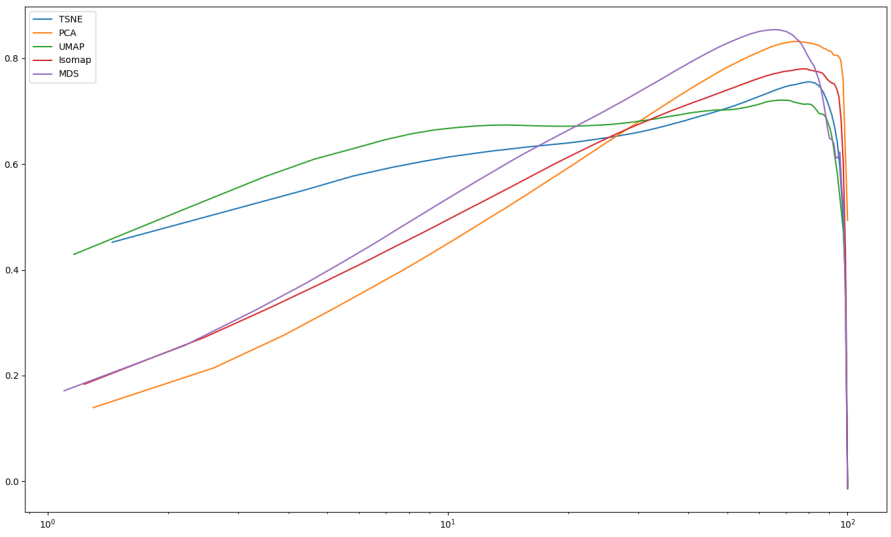


Figure 3.5: Path-based $R_{\mathrm{NX}}(K)$ scores for MNIST embeddings across different path lengths using PCA, t-SNE, UMAP, Isomap, and MDS.

The results show that the new metric accurately reflects the expected prioritization of local and global structures by each technique. For example, t-SNE and UMAP yield better performance for shorter paths, while PCA and MDS excel at longer paths. This demonstrates that the new metric captures the trade-offs inherent

in dimensionality reduction techniques and validates its utility in evaluating the behavior of embeddings across path-based metrics.

COIL-20 Dataset

While t-SNE and UMAP typically show better preservation of local structure in lower-dimensional embeddings for datasets like MNIST, they do not perform as expected with COIL-20. Instead of tightly preserving local neighborhoods, the embeddings produced by t-SNE and UMAP in this case exhibit large, spaced-out loops that are poorly connected. This results in a graph with many long, weak links, which distorts the path structure. These embeddings fail to represent the local relationships accurately, and the path-based quality metric reflects this discrepancy, with results showing poorer performance for local structures.

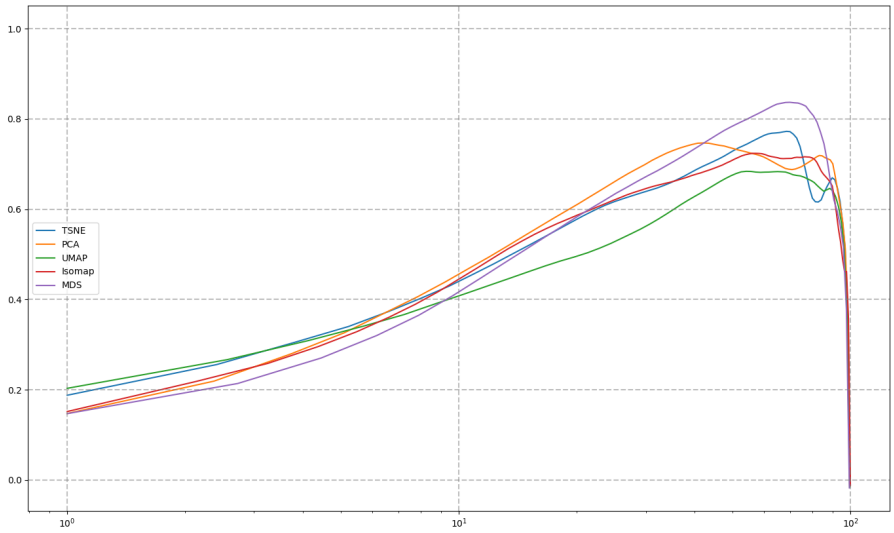


Figure 3.6: Path-based $R_{\mathrm{NX}}(K)$ scores for COIL-20 embeddings, showing poor preservation of local structure and reduced quality of the paths compared to MNIST.

This issue is particularly evident in the path-based evaluation, where techniques such as t-SNE and UMAP do not yield significantly better results for shorter paths, as would be expected in cases where local structure is preserved. The long, poorly connected links in the graph lead to path distances that do not accurately reflect the underlying structure of the data, resulting in suboptimal performance according to the new quality metric.

This discrepancy between the expected and observed results suggests that while the new metric is effective for simpler, more well-defined datasets like MNIST, additional adjustments or considerations may be needed when applying it to more complex datasets like COIL-20. The presence of large, spaced-out loops and the distortion of local relationships in the embeddings highlights potential challenges when using path-based evaluations on datasets with more intricate structures and varying intra-class relationships.

3.3.4 conclusion

The path-based $R_{NX}(K)$ metric demonstrates its effectiveness in evaluating the preservation of both local and global structures in dimensionality reduction techniques. It provides a nuanced understanding of how data is represented in low-dimensional spaces, reflecting the behavior of various methods like t-SNE, UMAP, PCA, and MDS. The metric successfully distinguishes between local and global preservation, as seen in datasets like MNIST, where techniques that prioritize local structure show better performance for shorter paths, while those focused on global structure excel with longer paths. However, the evaluation on more complex datasets, such as COIL-20, highlights challenges when local neighborhood preservation is distorted, suggesting the need for further refinements to adapt the metric to more intricate data structures. Overall, the proposed path-based $R_{NX}(K)$ metric offers a promising framework for assessing the quality of dimensionality reduction embeddings, but its application may require additional considerations for datasets with more complex relationships.

3.4 Edit Distance Method

Building on the insights gained from the path-based $R_{NX}(K)$ metric, the Edit Distance-based method introduces a new approach for evaluating dimensionality reduction techniques by focusing on the sequence of nodes along paths in high-dimensional (HD) and low-dimensional (LD) spaces. Edit distance measures the minimum number of operations—insertions, deletions, or substitutions—required to transform one sequence into another. In this context, the sequences represent the order of nodes encountered along specific paths in the HD and LD spaces.

By comparing these sequences, the edit distance provides a quantitative assessment of how well the relative order of nodes is preserved between the original high-dimensional space and its low-dimensional counterpart. A lower edit distance indicates that the path structure in the low-dimensional space closely matches that in the high-dimensional space, suggesting better preservation of the data's inherent structure during dimensionality reduction.

This method offers an alternative to traditional distance-based metrics by focusing on sequence alignment, which captures more subtle differences in path structure, particularly when the embeddings involve non-linear distortions or complex relationships between data points. It is expected that this approach will complement the path-based $R_{NX}(K)$ method by providing a more direct measure of sequence consistency and allowing for a deeper analysis of the dimensionality reduction process.

3.4.1 Implementation

The implementation methodology follows a similar approach to the previous chapter. The computation of graphs and paths remains unchanged, but here the comparison is performed using edit distance. A smaller edit distance indicates greater similarity in the order of points between the HD and LD paths, reflecting better preservation of structure.

To quantify the overall preservation of structure, the edit distance is computed for each path and aggregated across all pairs of points.

Edit distance has a worst-case value equal to the length of the sequence. However, experiments show that for a random permutation of two sequences containing the same alphabet, the expected edit distance is approximately length - 1.7 . These two lines are shown in the figures 3.7 and 3.8.

The following snippet demonstrates the core logic for computing and comparing paths:

```
for i in range(len(LD_data)):
    LD_paths, HD_paths = LDHDPathAll(graph, distance_matrix, i,
   LD_data, HD_data)
    for j in range(i + 1, len(LD_paths)):
        LD_path , HD_path = LD_paths[j], HD_paths[j]
        distance = levenshteinDistanceDP(LD_path, HD_path)
        path_len = len(LD_path)
        results[index][path_len].append(distance)

    results[index] = {k: np.mean(v) for k, v in sorted(results[
   index].items())}
```

Listing 3.2: Python implementation for path analysis

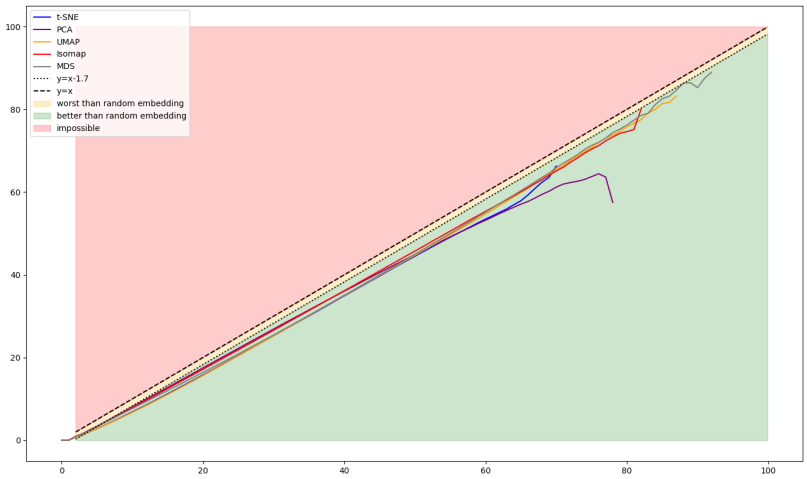


Figure 3.7: Edit distance metric for MNIST

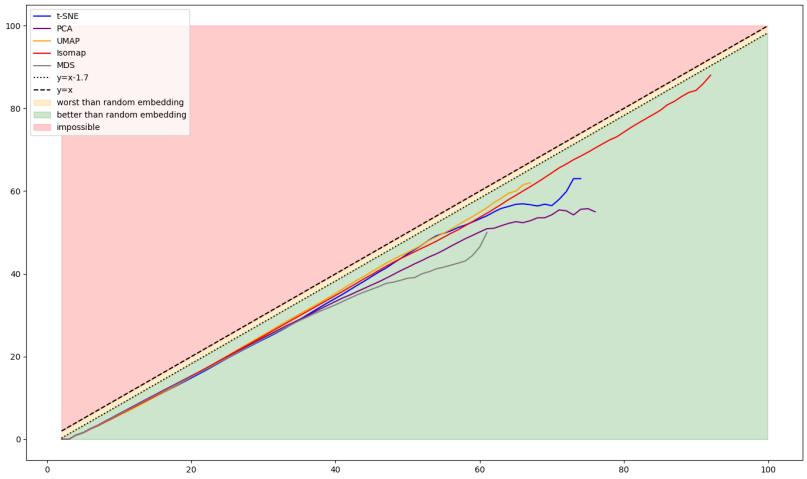


Figure 3.8: Edit distance metric for COIL-20

Here, we can see in red the area where the edit distance is impossible for a specific length. This corresponds to the worst-case scenario for the edit distance of two sequences, where the maximum possible distance is the length of the sequence itself. In addition, the orange band represents the area of the graph where the quality would be worse than random, meaning that the sequences are less aligned than

expected by chance. Below this orange band, we find the region where the results are better than random. The further we go down, the better the dimensionality reduction at this length.

For improved interpretability, the metric is rescaled to a range from 0 to 1. The rescaling is performed relative to a random embedding, as in the case of the $R_{NX}(K)$ metric. The transformation applied to the edit distance is:

$$d(S_1, S_2) = \frac{(\text{len}(S_1) - 1.7) - \text{EditDistance}(S_1, S_2)}{\text{len}(S_1)}$$

This transformation normalizes the edit distance to account for the expected distance in a random embedding, providing a clearer comparison of how well the structure is preserved. While this rescaling may slightly distort the best-fit lines, dotted lines are added to the graph to indicate specific quality values from the original metric, now scaled to the 0–1 range. This approach ensures that the visualizations remain intuitive and interpretable while offering a direct comparison to the random baseline.

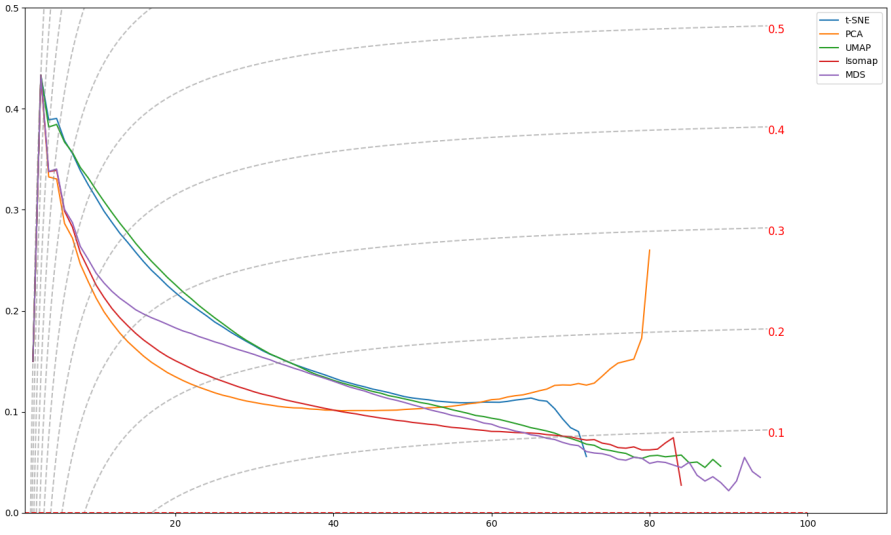


Figure 3.9: Scaled edit distance metric for MNIST

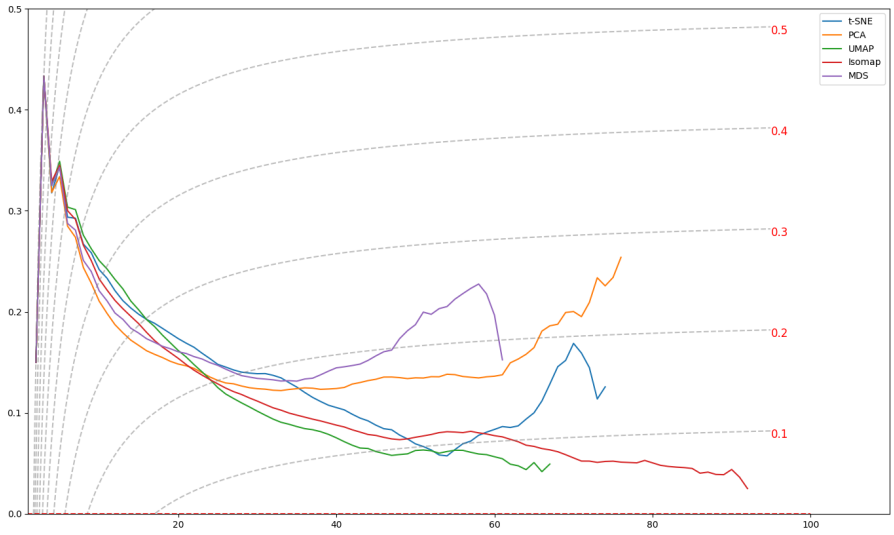


Figure 3.10: Scaled edit distance metric for COIL-20

In the visualization, the x-axis represents the path length, while the y-axis shows the corresponding average quality. The graphs confirm the expected behavior: methods emphasizing local structure exhibit better quality for shorter paths, while techniques prioritizing global structure perform better for longer paths. However, this trend is less pronounced for COIL-20, as previously discussed in the preceding chapter.

This representation has certain limitations. First, different dimensionality reduction techniques often produce varying maximum path lengths. For example, in the MDS (Multidimensional Scaling) case, the maximum path length extends to 130, which creates disparities in the range of values displayed. Such inconsistencies can hinder readability and complicate direct comparisons between methods. Simply rescaling the data to a 0–1 range exacerbates the issue, as it distorts the alignment of lines representing specific quality values differently for each method.

Additionally, visualization suffers from an increase in jitter at longer path lengths. This effect arises because fewer paths exist at these lengths, resulting in less reliable averaging and reducing the representativeness of the quality metric as path length increases.

3.4.2 Color map analysis

The metric implementation involves computing the average from all possible starting points. Each starting point generates its own individual metric, reflecting the localized quality of the embedding. While the overall metric is derived as the average of these individual metrics, analyzing them separately allows for a more granular understanding of the embedding's quality. By integrating the specific quality associated with each point and comparing it to the overall average, we can create a color map that visualizes the local quality across the embedding. This provides an intuitive and detailed view of how well different regions of the embedding perform.

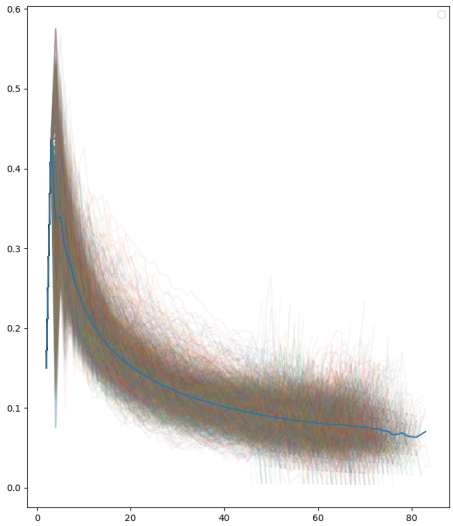


Figure 3.11: Individual quality for every starting point

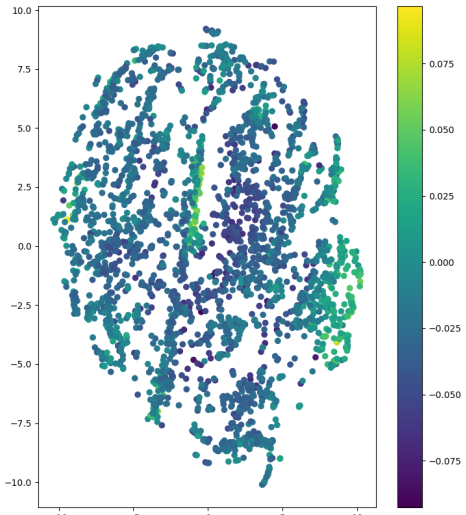


Figure 3.12: Local quality of the embedding

We can visualize regions of the embedding that displayes better quality. In this example we see a area in the center and an area on the botom far right of the graph.

3.4.3 Single Path Analysis

In this section, we examine the overall quality of paths generated from individual starting points in the embedding. This analysis provides insights into both local and global characteristics of the embedding, enabling a deeper understanding of the data structure and the performance of the embedding technique.

to compute the quality of ind

High Global Scores and Path Quality

For points with high global scores, we observe an abundance of paths with above-average quality. These results indicate that regions of the embedding associated with high global scores tend to preserve local structure effectively, as demonstrated in Figure 3.13.

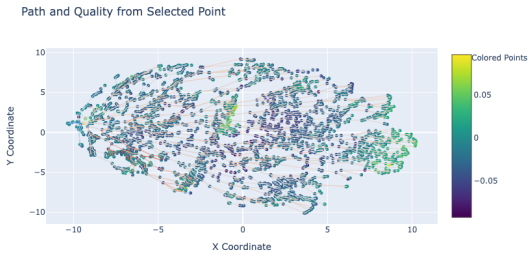


Figure 3.13: Path quality for high global scores, showing an abundance of above-average quality paths.

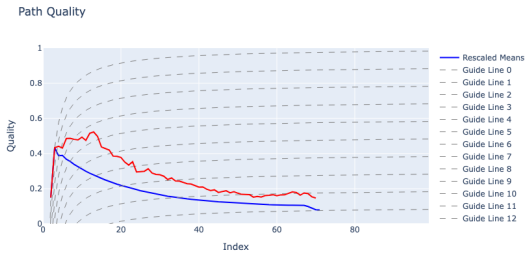


Figure 3.14: Localized quality distribution for high global scores, highlighting regions of well-preserved structure.

Low Global Scores and Path Quality

Conversely, points with low global scores exhibit a predominance of paths with below-average quality. These regions of the embedding suffer from poor structural preservation, as illustrated in Figure 3.15. This pattern suggests that these areas may be sparse, distorted, or poorly aligned with the overall data structure.

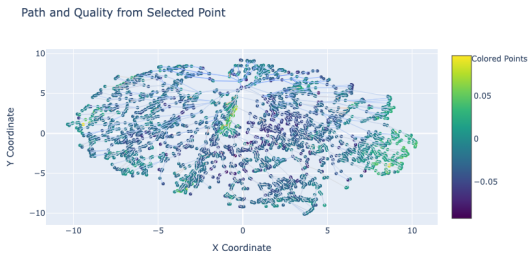


Figure 3.15: Path quality for low global scores, showing a predominance of below-average quality paths.

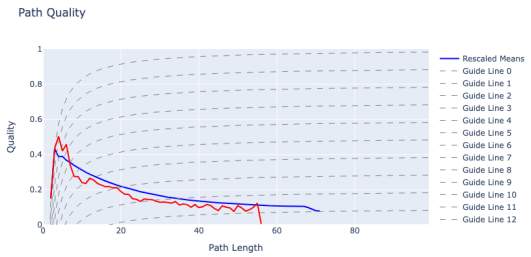


Figure 3.16: Localized quality distribution for low global scores, indicating regions of poor structural preservation.

Insight from moderately performing points

For points with moderate quality we observe increase in quality while crossing areas of the graph with better overall quality. In this example the quality increase for points that have a long length, these paths are mostly situated on the far right of the graph, a zone were the metric seems to suggest a better quality of the reduction. In the embedding we see that the paths in this zone tend to white , suggesting an increase from the blue paths it originates from.

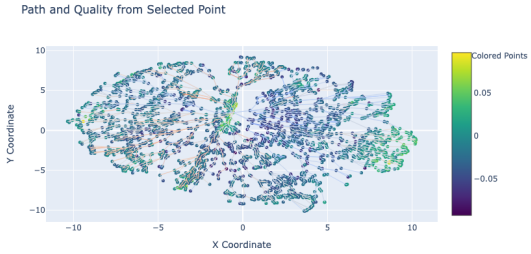


Figure 3.17: Path quality for moderately high global scores, highlighting the improvement driven by the longest paths.

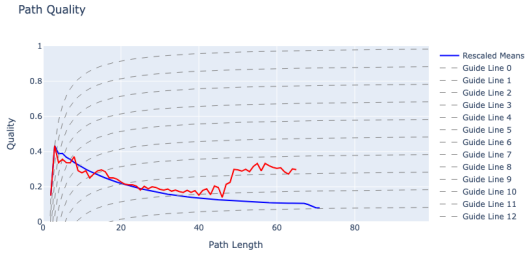


Figure 3.18: Localized quality distribution for moderately high scores, with longest paths leading to higher quality.

We can see a similar behavior in the folowing example, the overall quality is better for shorter path. This behavior is expected as the points is chosen in a region with good quality.

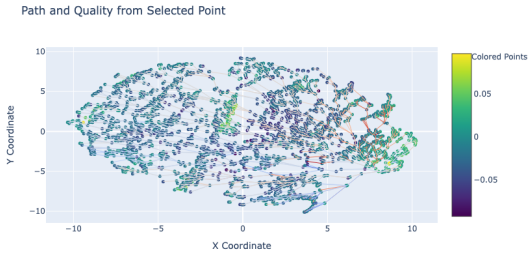


Figure 3.19: Path quality for moderately high global scores, highlighting the improvement driven by the longest paths.

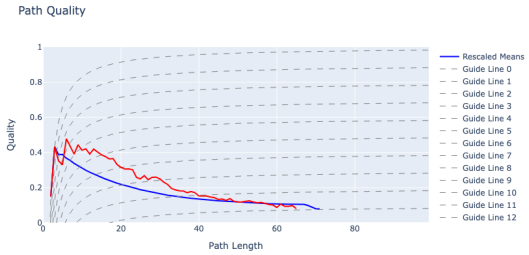


Figure 3.20: Localized quality distribution for moderately high scores, with longest paths leading to higher quality.

Granular Insights from Single Path Analysis

This granular analysis can be applied to every starting point, providing unique insights into the local structure of the embedding. For example, while high global scores indicate overall well-preserved regions, analyzing individual paths reveals whether these scores are uniformly distributed or skewed by specific patterns, such as long paths connecting better-quality regions.

By integrating these observations, we can identify areas of the embedding that require further optimization, enhancing both local and global structure preservation.

3.5 Deformation impact

A key limitation of $R_{NX}(K)$ lies in its sensitivity to distortions in the low-dimensional representation of data. This limitation stems from $R_{NX}(K)$'s isotropic approach to evaluating neighborhood preservation. While the overall structure of the data may remain intact under distortion, the relative order of neighbors around each point is often disrupted. This disruption can lead to significant variations in $R_{NX}(K)$ results, even if the embedding retains a globally coherent structure.

Distortions in low-dimensional representations frequently occur in dimensionality reduction methods, particularly those employing iterative processes like t-SNE and UMAP. These methods are highly sensitive to initialization and hyperparameter settings, which can result in slight variations in the output embedding. While the general geometric structure of the data is often preserved, iterative algorithms can introduce distortions that are non-linear, localized, or scale-dependent. These distortions disproportionately affect isotropic measures like $R_{NX}(K)$ by altering local neighborhood arrangements without fundamentally changing the global structure.

The new method overcomes this limitation by adopting a directional approach to

neighborhood evaluation. Rather than considering all directions simultaneously, it evaluates neighborhood preservation along specific paths, effectively isolating the analysis from distortions that affect secondary or less significant directions. By focusing on the shortest paths between points, this approach minimizes the impact of distortions and ensures greater consistency in the evaluation. For example, even when iterative methods produce embeddings with slight shifts or stretches, the shortest paths are less likely to change, allowing the new method to provide a more stable assessment of embedding quality.

This directional methodology is particularly advantageous for datasets prone to high variability in embedding outcomes. For instance, embeddings of complex datasets such as high-dimensional manifolds or non-linear structures often exhibit localized distortions. These distortions can lead to inconsistencies in isotropic evaluations like $R_{NX}(K)$ but are mitigated under the new framework. Additionally, by reducing the sensitivity to distortion, the method ensures a fairer comparison between embeddings generated under different initialization conditions or algorithmic variations.

To illustrate the robustness of the new method, we applied a distortion factor of 1.5 along the X-axis to a multidimensional scaling (MDS) embedding on a Diabetes dataset [15]. The results demonstrate that the new method maintains high robustness to deformations for shorter path lengths. However, for longer paths, a noticeable difference emerges. This discrepancy arises due to the behavior of MDS, which tends to place distant points at the periphery of the embedding. These peripheral points are more strongly impacted by the distortion, as changes to the Delaunay triangulation structure significantly affect the shortest path costs to reach them.

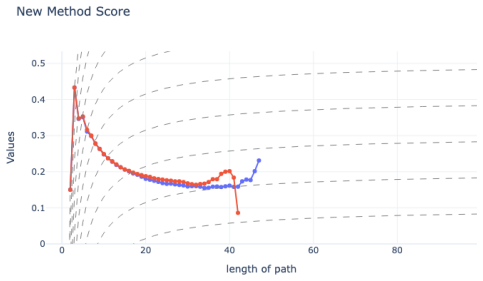


Figure 3.21: Performance of the new method: Blue represents the distorted embedding, while red represents the normal embedding. The method remains robust for shorter paths.

Figure 3.22: $R_{NX}(K)$ performance comparison: Blue represents the distorted embedding, while red represents the normal embedding. $R_{NX}(K)$ is more affected by distortions.

These results highlight the strengths of the path-based approach, particularly in addressing the limitations of isotropic measures like $R_{NX}(K)$ under deformations. The new method's reliance on stable geometric constructs, such as Delaunay triangulation and directional analysis, ensures robust evaluations that reflect the true quality of the embedding, even in the presence of significant distortions.

3.6 Acceleration

The computational cost of the algorithm is substantial, as it requires calculating paths between all pairs of points. This results in a time complexity of $\mathbb{O}(n^2 \log(n))$, since Dijkstra's algorithm must be run for each starting point. For large datasets, this approach quickly becomes infeasible.

To mitigate this issue, we propose using the convex hull of the low-dimensional embedding as the set of starting points for Dijkstra's algorithm. The intuition behind this acceleration is that paths generated using the convex hull points form a representative subset of all possible paths within the embedding. These paths provide sufficient coverage of the total space while significantly reducing computational overhead.

3.6.1 Single-Layer Convex Hull Acceleration

Using only the convex hull as the starting points accelerates computation, but it introduces a limitation: for most embeddings, the convex hull comprises only a small fraction of the total points. As a result, the quality of the approximation may suffer.

in the following we display the convex hull of two different dataset. We can see in those graph that the number of points considered in the convex hull is very dependent on the geometry of the embedding. This

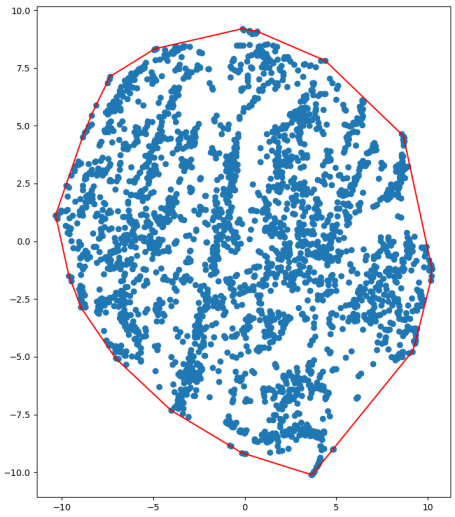


Figure 3.23: convex hull on t-SNE reduction of the mnist dataset, the convex hull contains 29 points out of 3000

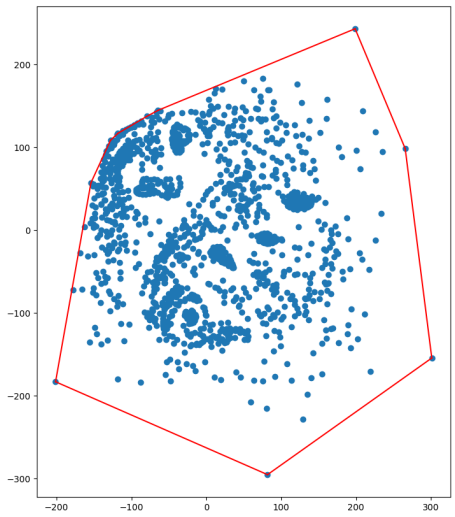


Figure 3.24: convex hull on MDS reduction of the coil 20 dataset, the convex hull contains 10 points out of 1440

In the following graphs, we display side-by-side the evaluation of the metric based on the convex hull against the true metric. In the second plot, we show the points considered for each number of layers. Points in color represent those added in each successive layer, corresponding to the new "peels" captured.

The results are promessing as they seems to tend to fit the actual curve. There is still a lot of improvement to be made as the number of points is often not sufficient for an accurate approximation.

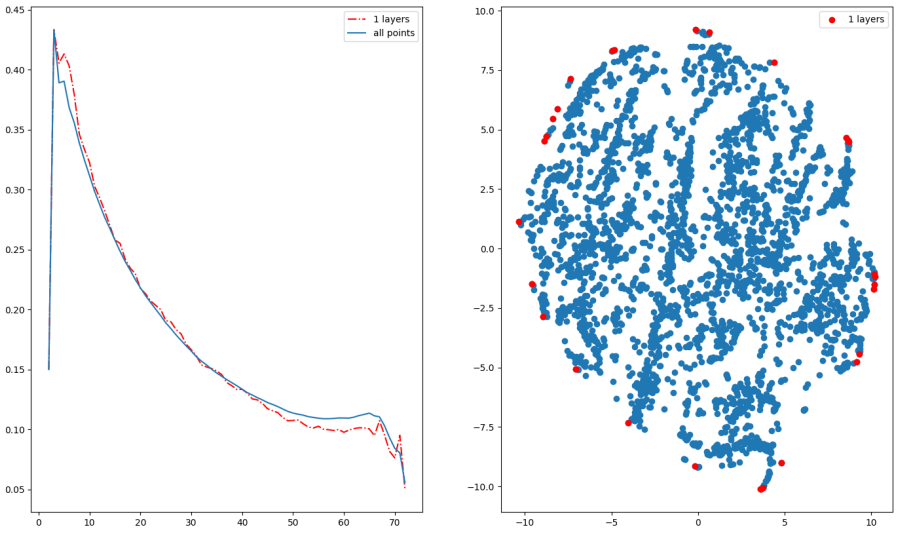


Figure 3.25: approximation based on alpha value for t-SNE reduction of the mnist dataset

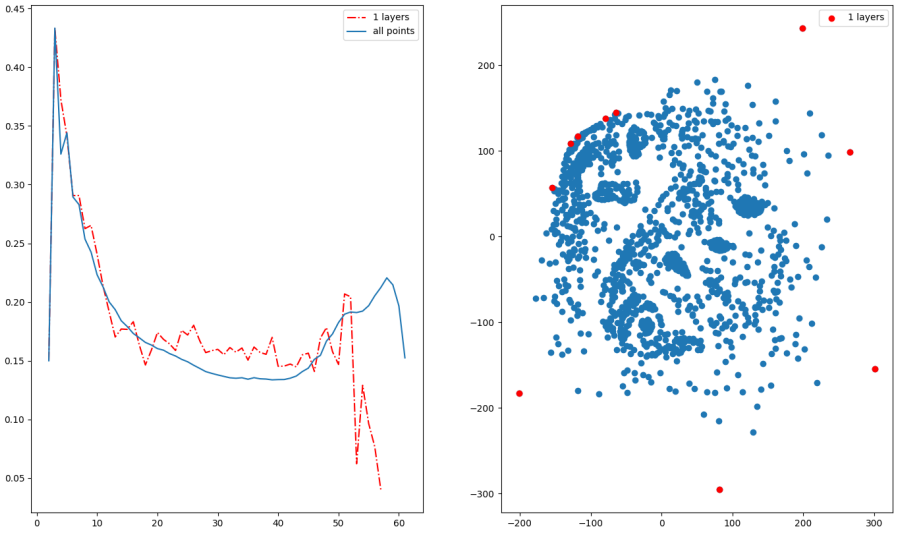


Figure 3.26: approximation based on alpha value for MDS reduction of the coil 20 dataset

3.6.2 Multi-Layer Convex Hulls

To address the limitations of a single convex hull, we introduce a multi-layer convex hull approach. In this method, we iteratively compute the convex hull of the remaining points after removing the current convex hull layer. This process effectively "peels" the embedding, capturing additional points in each subsequent layer.

In the following, we display the layers obtained from three consecutive convex hull computations. As we progress through the layers, we observe an increase in the number of points while retaining those from the same region, thereby increasing the width of the convex hull.

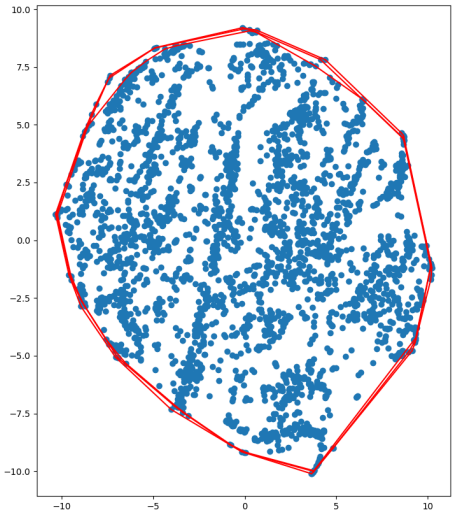


Figure 3.27: 3 consecutive convex hull on t-SNE reduction of the mnist dataset, the convex hulls contains 29, 30 and 32 points for a total of 91 points out of 3000

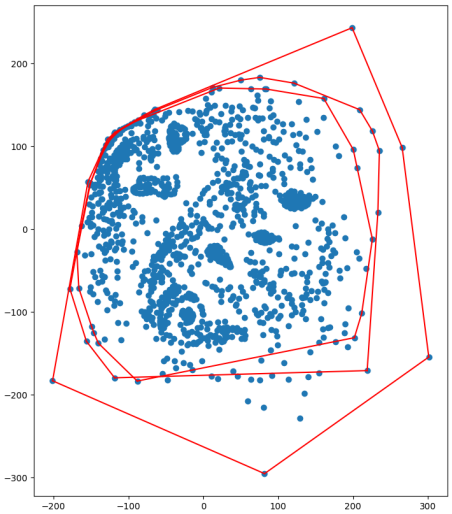


Figure 3.28: 3 consecutive convex hull on MDS reduction of the coil 20 dataset, the convex hull contains 10, 20 and 23 points for a total of 53 points out of 1440

In the following graphs, we display side-by-side the evaluation of the metric for different numbers of considered layers (represented by dotted lines) against the true metric. In the second plot, we show the points considered for each number of layers. Points in color represent those added in each successive layer, corresponding to the new "peels" captured. As the number of points increases, the approximation improves, indicating that our initial intuition is correct.

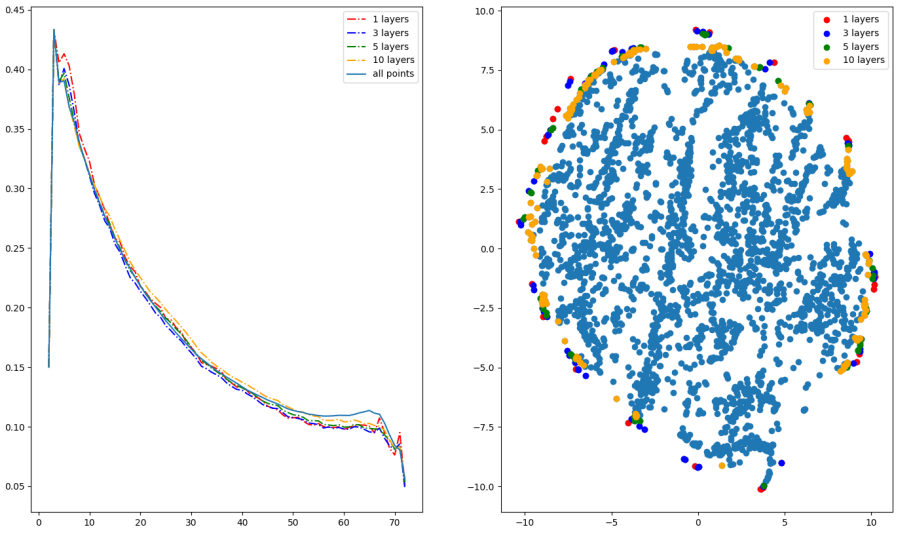


Figure 3.29: approximation based on alpha value for t-SNE reduction of the mnist dataset

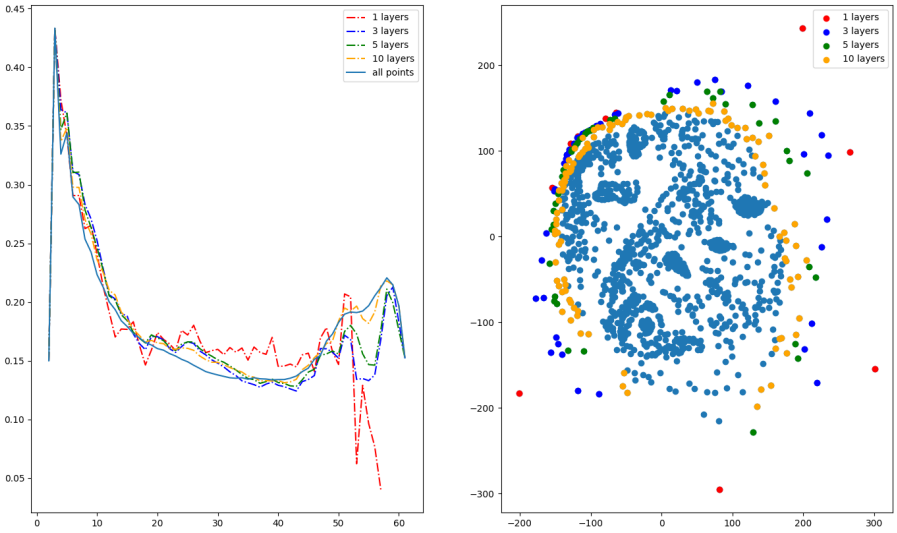


Figure 3.30: approximation based on alpha value for MDS reduction of the coil 20 dataset

To assess the effect of increasing the number of layers used to compute the metric, we evaluate the mean squared error (MSE) between the generated curve and the curve obtained by considering all the points in the dataset. This evaluation is conducted across a variety of datasets and dimensionality reduction techniques.

Table 3.2: MSE from Layers (1, 3, 5, 10) with Number of Points

Dataset	**1 layer**	**3 layers**	**5 layers**	**10 layers**
Isomap MNIST	$1.95 * 10^{-4}$ (13)	$2.75 * 10^{-4}$ (45)	$2.40 * 10^{-4}$ (74)	$1.56 * 10^{-4}$ (174)
Isomap COIL-20	$4.14 * 10^{-4}$ (6)	$1.49 * 10^{-4}$ (18)	$1.42 * 10^{-4}$ (27)	$7.67 * 10^{-5}$ (60)
MDS MNIST	$7.30 * 10^{-4}$ (9)	$6.60 * 10^{-4}$ (32)	$8.47 * 10^{-4}$ (58)	$7.17 * 10^{-4}$ (127)
MDS COIL-20	$1.68 * 10^{-4}$ (10)	$4.78 * 10^{-4}$ (53)	$2.99 * 10^{-4}$ (95)	$5.80 * 10^{-5}$ (207)
PCA MNIST	$2.86 * 10^{-3}$ (15)	$1.91 * 10^{-3}$ (58)	$1.45 * 10^{-3}$ (108)	$8.34 * 10^{-4}$ (246)
PCA COIL-20	$3.28 * 10^{-4}$ (17)	$1.88 * 10^{-4}$ (59)	$1.19 * 10^{-4}$ (94)	$5.92 * 10^{-5}$ (193)
t-SNE MNIST	$7.86 * 10^{-5}$ (29)	$5.19 * 10^{-5}$ (91)	$2.19 * 10^{-5}$ (150)	$2.07 * 10^{-5}$ (309)
t-SNE COIL-20	$3.91 * 10^{-4}$ (53)	$2.14 * 10^{-4}$ (103)	$1.31 * 10^{-4}$ (151)	$8.42 * 10^{-5}$ (255)
UMAP MNIST	$1.07 * 10^{-4}$ (26)	$1.07 * 10^{-4}$ (98)	$8.94 * 10^{-5}$ (166)	$9.81 * 10^{-5}$ (315)
UMAP COIL-20	$1.57 * 10^{-4}$ (18)	$1.86 * 10^{-4}$ (53)	$2.17 * 10^{-4}$ (77)	$1.83 * 10^{-4}$ (143)

This multi-layer approach provides increasingly accurate approximations of the quality curve as more layers are added. However, while effective, it can still miss finer structures within the embedding, especially for datasets with intricate geometries or multiple subgroups.

3.6.3 Alpha Shapes for Improved Coverage

To further enhance the approximation, we use alpha shapes instead of traditional convex hulls derived from Delaunay triangulation. Alpha shapes allow for a more detailed representation of the dataset's geometry, as their "convex hull" tightly follows the shape of the data. This has several advantages: Increased Coverage: Alpha shapes include more points compared to a single convex hull layer, capturing finer details of the embedding. Multi-Group Embeddings: Alpha shapes can penetrate deeper into the graph, even when dealing with embeddings that contain multiple subgroups.

However, alpha shapes also introduce some challenges:

1. Disconnected Graphs: Alpha shapes can result in disconnected components, which is why we continue to use Delaunay triangulation for the computation of quality metrics.
2. Choice of Alpha Parameter: The selection of the α parameter significantly impacts the number of points included, influencing both computation time and approximation accuracy. Choosing an appropriate α depends on the geometry of the embedding, adding complexity for the end user.

In the following, we display the convex hulls based on the alpha shape for two different datasets, each with two distinct alpha values. These graphs illustrate that the number of points included in the convex hull strongly depends on the geometry of the embedding. Consequently, the choice of the alpha value cannot be arbitrary across all embeddings. Instead, it should be determined using exploratory methods, such as grid search or binary search, or left to the discretion of the user.

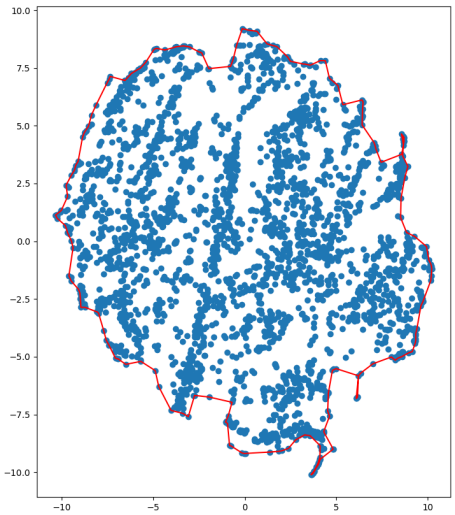


Figure 3.31: convex hull of the alpha shape with alpha = 0.3 on t-SNE reduction of the mnist dataset, the convex hulls contains points 178 out of 3000

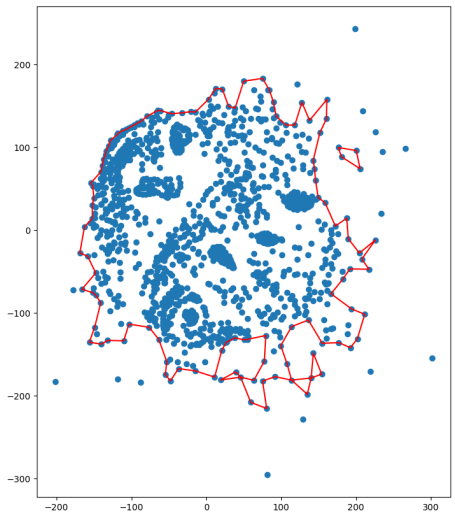


Figure 3.32: convex hull of the alpha shape with alpha = 0.3 on MDS reduction of the coil 20 dataset, the convex hulls contains points 125 out of 1440

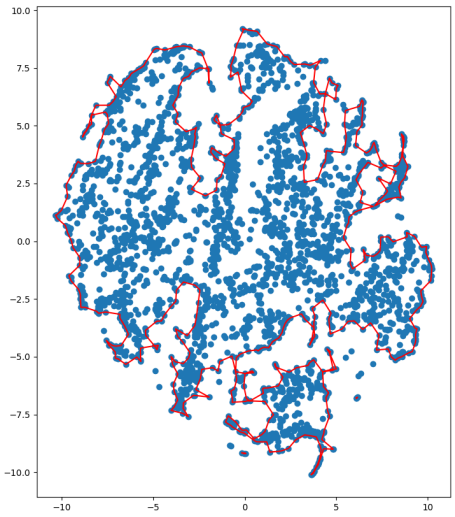


Figure 3.33: convex hull of the alpha shape with alpha = 0.5 on t-SNE reduction of the mnist dataset, the convex hulls contains points out of 3000

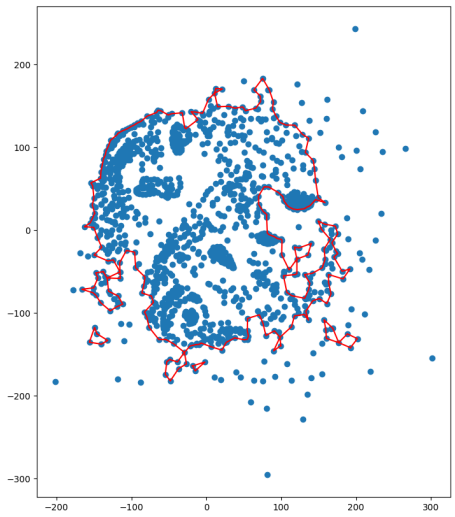


Figure 3.34: convex hull of the alpha shape with alpha = 0.5 on MDS reduction of the coil 20 dataset, the convex hulls contains points 236 out of 1440

In the following graphs, we display side-by-side the evaluation of the metric for different alpha values (represented by dotted lines) against the true metric. In the second plot, we show the points considered for each number of layers. Points in color represent those added in each successive increase of alpha. As the number of points increases, the approximation improves, indicating that our initial intuition is correct

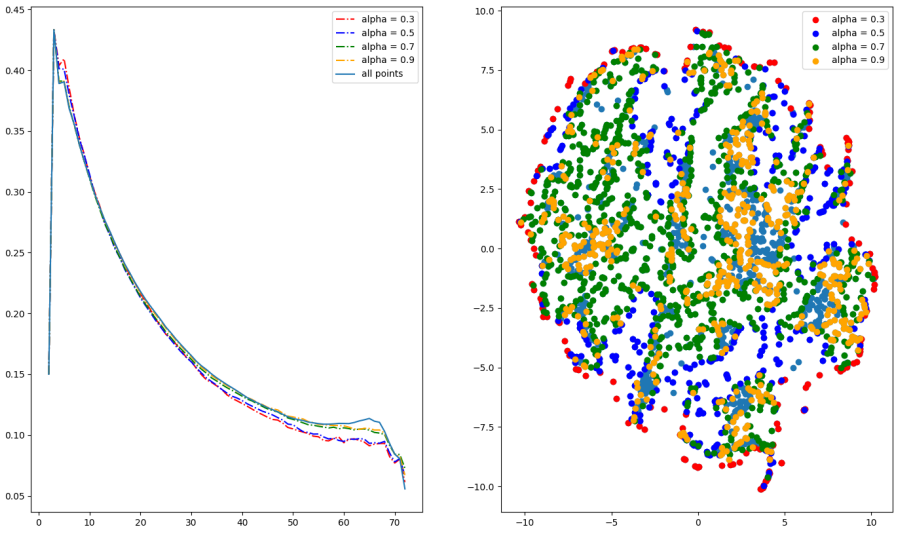


Figure 3.35: approximation based on alpha value for t-SNE reduction of the mnist dataset

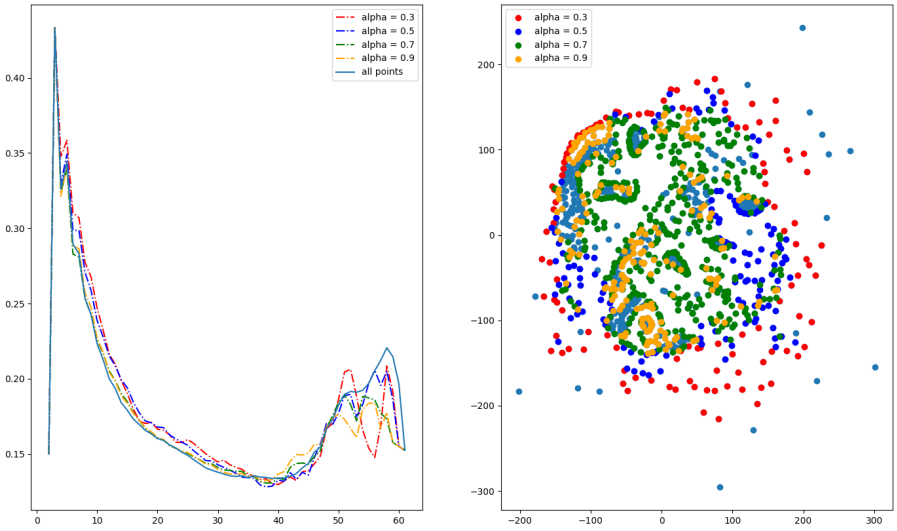


Figure 3.36: approximation based on alpha value for MDS reduction of the coil 20 dataset

To assess the effect of increasing the alpha value used to compute the metric, we evaluate the mean squared error (MSE) between the generated curve and the curve obtained by considering all the points in the dataset. This evaluation is conducted across a variety of datasets and dimensionality reduction techniques.

Table 3.3: MSE from Alphas (0.3, 0.5, 0.7, 0.9) with Number of Points

Dataset	**Alpha 0.3**	**Alpha 0.5**	**Alpha 0.7**	**Alpha 0.9**
Isomap MNIST	$2.42 * 10^{-4}$ (127)	$1.68 * 10^{-4}$ (254)	$7.73 * 10^{-5}$ (497)	$4.24 * 10^{-5}$ (716)
Isomap COIL-20	$5.75 * 10^{-5}$ (181)	$2.10 * 10^{-5}$ (324)	$9.28 * 10^{-6}$ (410)	$1.48 * 10^{-5}$ (515)
MDS MNIST	$7.97 * 10^{-4}$ (73)	$5.66 * 10^{-4}$ (116)	$4.33 * 10^{-4}$ (139)	$3.36 * 10^{-4}$ (222)
MDS COIL-20	$2.80 * 10^{-4}$ (125)	$1.15 * 10^{-4}$ (236)	$1.68 * 10^{-4}$ (631)	$2.08 * 10^{-4}$ (770)
PCA MNIST	$8.12 * 10^{-4}$ (128)	$2.74 * 10^{-4}$ (316)	$1.20 * 10^{-4}$ (423)	$2.03 * 10^{-5}$ (602)
PCA COIL-20	$1.43 * 10^{-4}$ (182)	$1.03 * 10^{-4}$ (392)	$3.71 * 10^{-5}$ (728)	$7.74 * 10^{-5}$ (791)
t-SNE MNIST	$8.73 * 10^{-5}$ (178)	$6.65 * 10^{-5}$ (541)	$1.33 * 10^{-5}$ (1424)	$3.59 * 10^{-6}$ (1902)
t-SNE COIL-20	$1.05 * 10^{-5}$ (719)	$8.69 * 10^{-6}$ (1006)	$6.12 * 10^{-6}$ (1005)	$1.47 * 10^{-5}$ (1023)
UMAP MNIST	$3.46 * 10^{-5}$ (353)	$3.06 * 10^{-5}$ (514)	$1.44 * 10^{-5}$ (794)	$7.71 * 10^{-6}$ (1255)
UMAP COIL-20	$2.05 * 10^{-5}$ (446)	$1.04 * 10^{-5}$ (561)	$4.80 * 10^{-6}$ (758)	$3.59 * 10^{-6}$ (852)

The alpha-shape approach provides increasingly accurate approximations of the quality curve as more layers are added. However, it considers significantly more points than the multi-layer approach while offering only marginal improvements in results. In the context of acceleration, this trade-off becomes critical, as the computational cost of including more points can outweigh the slight gains in accuracy.

3.6.4 Random selection

To validate our initial hypothesis, we compare the results obtains by the two previous methods against a random selection of points of the dataset. We can see in the following graphs that the hull based methods seems to yeild as good or even sometimes worst results than a simple random selection of the points.

This unexpected result suggests that the additional geometric structure captured by the hull-based methods does not necessarily translate into better approximations of the metric.

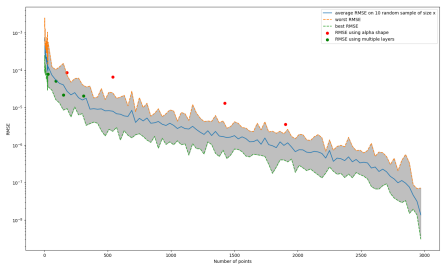


Figure 3.37: MSE of random selection of point on t-SNE applied to MNIST dataset

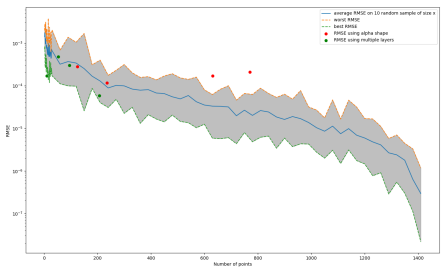


Figure 3.38: MSE of random selection of point on MDS applied to COIL-20 dataset

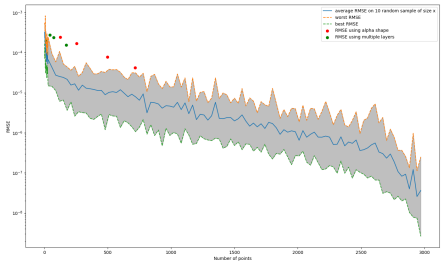


Figure 3.39: MSE of random selection of point on Isomap applied to MNIST dataset

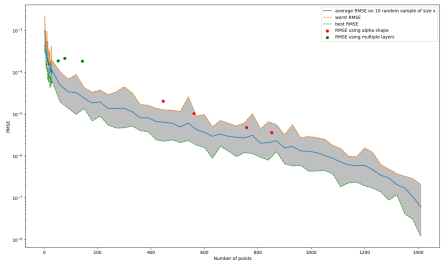


Figure 3.40: MSE of random selection of point on UMAP applied to COIL-20 dataset

Nevertheless, the results indicate that using only a fifth of the total number of points already yields a very good approximation. Since the total computation time for the metric scales linearly with the number of points considered, this approach reduces the computation time by a factor of five when employing a random selection.

The combination of multi-layer convex hulls and alpha shapes offers a structured approach to approximating quality metrics while reducing computational time. Alpha shapes provide greater flexibility and accuracy by capturing finer geometric details of the embedding, though careful selection of the α parameter remains critical to balancing computational cost and approximation quality.

However, our findings suggest that a simpler random selection of points can achieve comparable, or even better, results in significantly less time. With just a fifth of the total dataset points, random sampling yields a strong approximation of the metric, reducing computational time by a factor of five due to its linear scaling with the number of points.

Chapter 4

Conclusion and perspectives

In this thesis, I have investigated the assessment of quality in dimensionality reduction techniques. Historically, quality metrics in this field were designed to serve specific methods and were used primarily used as optimization targets. These early metrics were too method-specific, limiting their ability to facilitate direct comparisons between different techniques. The introduction of trustworthiness and continuity [5] represented the first attempt to create a more general quality metric, laying the foundation for further advancements. Building on the framework established in [24], my thesis work introduces a novel approach to quality assessment.

Existing metrics typically rely on neighborhoods to evaluate quality. These methods are effective, I believe the neighborhood-based approach has notable limitations, primarily because it fails to account for directionality within the data. Reducing the data structure to a one-dimensional list of nearest neighbors oversimplifies its complexity, inevitably leading to the loss of valuable information. This makes metrics like $R_{NX}(K)$ highly sensitive to distortions in low-dimensional (LD) space, even when the global structure is preserved.

To address these challenges, I propose a new paradigm based on shortest paths within the LD space. By considering paths instead of neighborhoods, this approach forms a network—or "web"—around a starting point, rather than a simple linear list. This web captures significantly more information and more faithfully reflects the original structure of the data. Moreover, this web-like representation aligns more closely with an observer intuitively perceives a cloud of points in two dimensions. Analyzing the order of points along these paths offers a richer more nuanced perspective on data quality. Unlike isotropic neighborhood-based metrics, which assume a uniform sphere of influence, the path-based approach adapts to the inherent structure of the LD space, providing a more intuitive and faithful representation of dimensionality reduction quality.

In this thesis, I propose two methods for evaluating quality using shortest paths:

1. A path-based adaptation of $R_{NX}(K)$: This adapts the metric from [24], applying it to shortest paths instead of neighborhoods.
2. An edit-distance-based comparison of paths: This provides an alternative perspective by analyzing path similarity.

Both methods yield results consistent with $R_{NX}(K)$ at the global level, reinforcing its validity. Specifically, higher values at shorter path lengths are associated with methods that preserve local structures, while higher values at longer path lengths indicate better preservation of global structures. These observations hold across different dimensionality reduction techniques and provide a solid foundation for exploring the finer details revealed by the path-based metrics.

With these new methods, it becomes possible to detect poorly positioned elements in the embedding, as well as to identify zones of higher or lower quality where multiple elements deviate from the global quality. Moreover, by leveraging the path-based quality information, it is possible to directly interpret the quality of connections between specific pairs of points. This dual perspective helps users understand the relationships between different regions of the graph and provides insights into the origins of quality diminutions.

In zones of poorer quality, examining paths with low-quality scores can reveal the sources of distortion. If all paths in a region exhibit poor quality, this may indicate a zone with a complex high-dimensional structure that the dimensionality reduction algorithm struggled to simplify. Conversely, if specific paths within a low-quality zone exhibit higher scores, this can highlight localized challenges in mapping particular portions of the high-dimensional space.

These insights enable users to better understand where the dimensionality reduction algorithm successfully captured the global structure of the data and where artifacts of the reduction method dominate. By distinguishing between these scenarios, users gain a more nuanced understanding of the embedding's fidelity, which can guide further refinement and interpretation of the results.

This approach presents promising new insights, it also introduces computational challenges. Calculating and comparing all paths does not scale well with the number of data points. However, this thesis demonstrates that randomly sampling paths allows the global quality curve to converge efficiently, reducing computation time. Full-path calculations can be reserved for areas of interest selected by the user.

Representing the web-like structure of the data also poses challenges. Initially, the aim was to create a visualization similar to a heat map, but I struggled to develop a representation that felt intuitive and satisfactory. Currently, the quality is displayed either at each node or along the paths being traversed. A well-designed,

accessible visualization could greatly enhance the interpretability and utility of this approach, making the insights more comprehensible and actionable for researchers and practitioners alike.

This path-based paradigm for quality assessment offers an exciting direction for future research. Further exploration could involve optimizing graph construction within the LD space or leveraging geodesic distances in the high-dimensional (HD) space to refine the method. These developments could improve the reliability of path-based metrics in capturing the true structure of the data.

As the volume and complexity of data continue to grow across diverse fields—from biology and finance to artificial intelligence—the tools we use to analyze and interpret this data must evolve to meet the increasing demands for clarity and insight. This thesis introduces a novel, path-based approach to dimensionality reduction quality assessment, bridging intuitive understanding with rigorous analysis. By offering a more nuanced and reliable framework for evaluating data structure, this research contributes not only to the advancement of dimensionality reduction techniques but also to the broader pursuit of data-driven discoveries.

To conclude, I believe these methods have the potential to empower researchers and practitioners across various fields, enabling them to uncover deeper patterns and relationships, and ultimately contributing to more transparent and impactful decision-making. While much work remains in refining and expanding these ideas, I hope my thesis serves as a starting point for future research. I look forward to seeing this work inspire further exploration and collaboration, advancing our understanding of data visualization and quality assessment, and fostering a future where the true structure of data can be more effectively understood and utilized.

Acknowledgment

I would like to express my sincere gratitude to my supervisor, Professor John Lee, for proposing this engaging and challenging thesis topic and for his continuous guidance and support throughout this journey.

I also extend my heartfelt thanks to Pierre Lambert for his invaluable assistance and insightful feedback, which have greatly contributed to the refinement of this work.

My appreciation goes to the jury members, including Professor Michel Verleysen, for their time and effort in reviewing this thesis.

Additionally, I would like to thank Geoffroy Panis and his formation supervisor, Antoine Gilliard, for providing me with the opportunity to present my thesis at the "BeCode AI Bootcamp, Pôle Image de Liège".

Thank you all for your contributions and support.

To be complete, I utilized DeepL and ChatGPT to improve the clarity of my text, and refine the English language. All outputs were reviewed and validated by my self.

Bibliography

[1] C. H. Achen. *Interpreting and Using Regression*. Sage, Beverly Hills, 1982.

[2] C. H. Achen. What does "explained variance" explain?: Reply. *Political Analysis*, 2(1):173–184, 1990.

[3] H. Bauer and K. Pawelzik. Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks*, 1992.

[4] H.-U. Bauer, K. Pawelzik, and T. Geisel. A topographic product for the optimization of self-organizing feature maps. *Advances in Neural Information Processing Systems*, 4, 1991.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

[6] L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104(485):209–219, 2009.

[7] B. Delaunay. Sur la sphère vide. a la mémoire de georges voronoï. . , (6):793–800, 1934.

[8] L. 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.

[9] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, 29(4):551–559, 1983.

[10] S. France and D. Carroll. Development of an agreement metric based upon the rand index for the evaluation of dimensionality reduction techniques. In *Lecture Notes in Computer Science*, volume 4571. Springer, 2007.

[11] Y. Goldberg and Y. Ritov. Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. *Machine learning*, 77:1–25, 2009.

[12] J. E. Goodman, J. O'Rourke, and C. D. Tóth. Handbook of discrete and computational geometry. In J. E. Goodman, J. O'Rourke, and C. D. Tóth, editors, *Handbook of Discrete and Computational Geometry*, chapter 27, page 711. CRC Press LLC, Boca Raton, FL, 3rd edition, 2017.

[13] A. Gracia, S. González, V. Robles, and E. Menasalvas. A methodology to compare dimensionality reduction algorithms in terms of loss of quality. *Information Sciences*, 270:1–27, 2014.

[14] H. Handa. On the effect of dimensionality reduction by manifold learning for evolutionary learning. *Evolving Systems*, 2(4):235–247, 2011.

[15] A. D. Khare. Diabetes dataset, 2020. Accessed: 2024-12-17.

[16] D. Kobak and P. Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):5416, 2019.

[17] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[18] A. Konig. Interactive visualization and analysis of hierarchical neural projections for data mining. *IEEE Transactions on Neural Networks Learning Systems*, 2000.

[19] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964.

[20] J. B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[21] J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, New York, London, 2007.

[22] J. Lee and M. Verleysen. Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods. In *Journal of Machine Learning Research (JMLR) - Proceedings Track*, 2008.

[23] J. Lee and M. Verleysen. Rank-based quality assessment of nonlinear dimensionality reduction. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2008.

[24] J. Lee and M. Verleysen. Quality assessment of dimensionality reduction: rank-based criteria. *Neurocomputing*, 2009.

[25] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

[26] D. M. Y. Leung and Z. Xu. A new quality assessment criterion for nonlinear dimensionality reduction. *Neurocomputing*, 2011.

[27] W. Lueks, B. Mokbel, M. Biehl, and B. Hammer. How to evaluate dimensionality reduction?-improving the co-ranking matrix. *arXiv preprint arXiv:1110.3917*, 2011.

[28] S. A. Nene, S. K. Nayar, H. Murase, et al. Columbia object image library (coil-20). *Citeseer*, 1996.

[29] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[30] J. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 1969.

[31] L. Saul and S. Roweis. Think globally, low dimensional manifolds. *Journal of Machine Learning Research*, 2003.

[32] R. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 1962.

[33] R. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function, ii. *Psychometrika*, 1962.

[34] S. Sidney. Nonparametric statistics for the behavioral sciences. *The Journal of Nervous and Mental Disease*, 125(3):497, 1957.

[35] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[36] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[37] L. van der Maaten. The matlab toolbox for dimensionality reduction, 2012. Available at http://lvdmaaten.github.io/drtoolbox/.

[38] J. Venna. *Dimensionality reduction for visual exploration of similarity structures*. Helsinki University of Technology, 2007.

[39] J. Venna and S. Kaski. Local multidimensional scaling with controlled tradeoff between trustworthiness and continuity. In *Proceedings of 5th Workshop on Self-Organizing Maps*, pages 695–702, 2005.

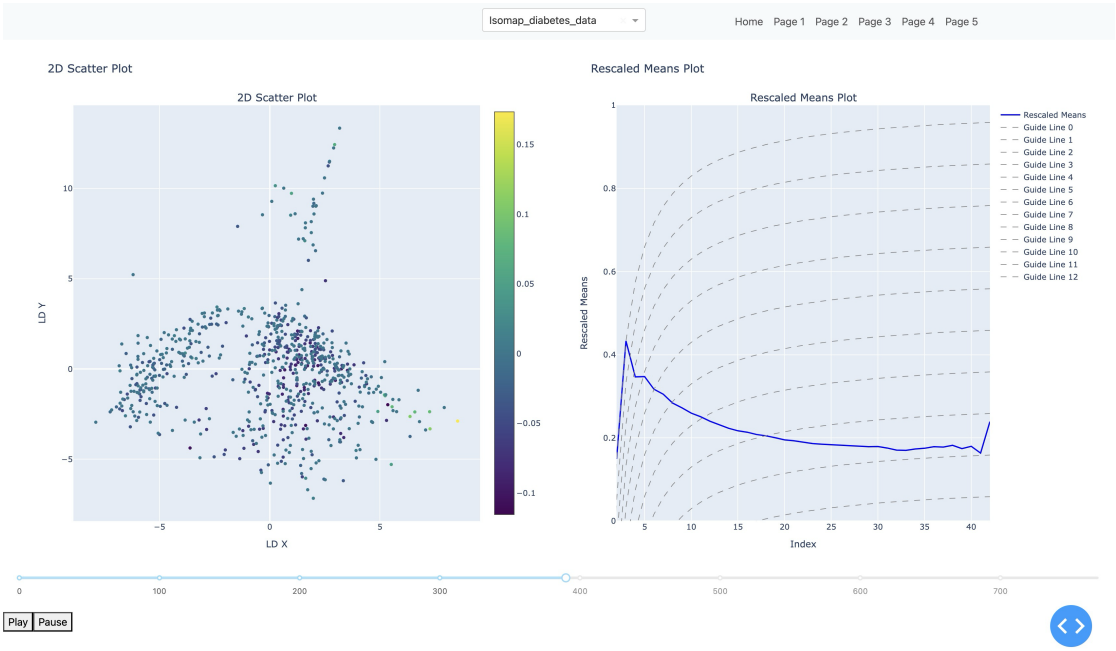
[40] T. Villmann, R. Der, and T. Martinetz. A new quantitative measure of topology preservation in kohonen's feature maps. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 2, pages 645–648. IEEE, 1994.

[41] K. Q. Weinberger, F. Sha, and L. K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the twenty-first international conference on Machine learning*, page 106, 2004.

Dashboard

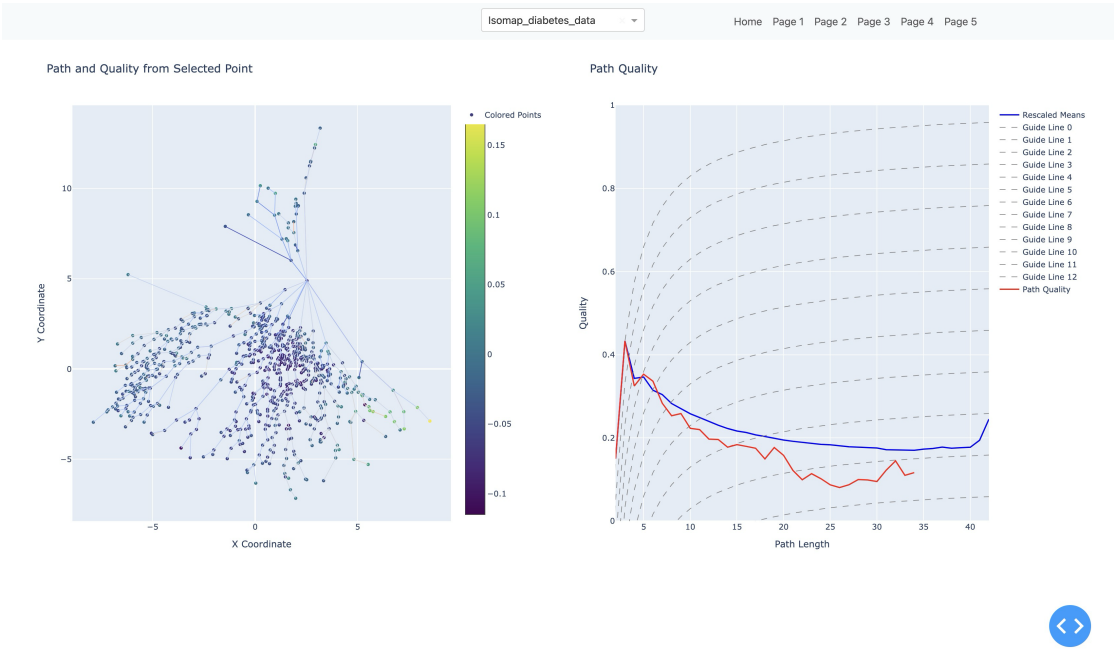
This appendix provides an in-depth overview of the dashboard developed for assessing path-based quality metrics. The dashboard consists of five main tabs, each dedicated to a specific aspect of the analysis.

Tab 1: Convergence Analysis



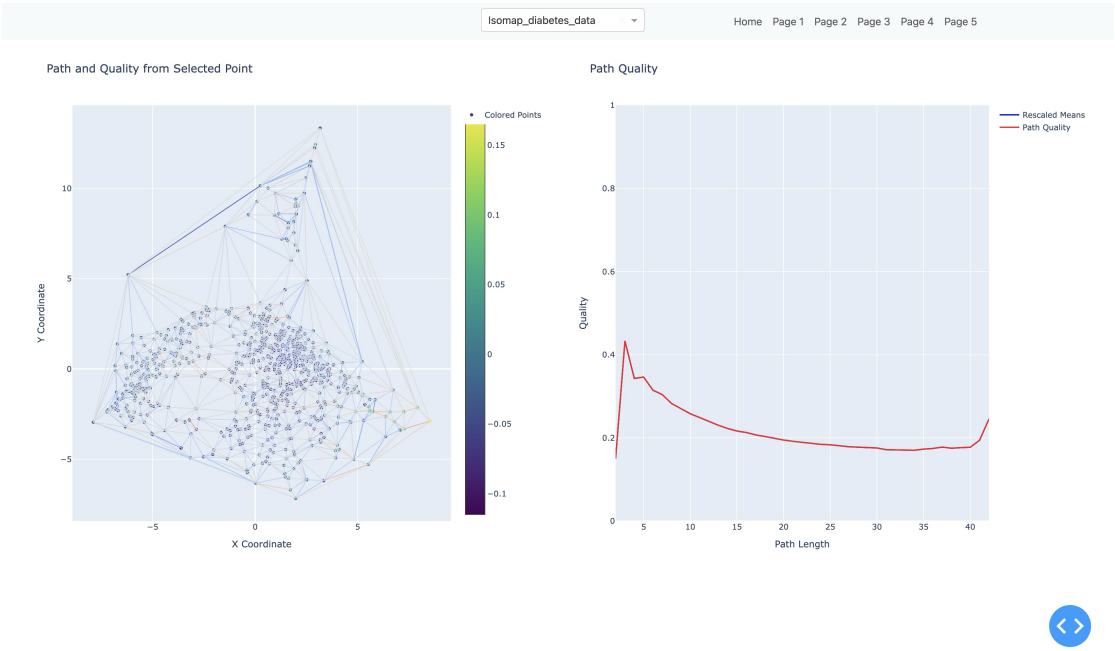
This tab examines the convergence of the path-based quality analysis algorithm, offering insights into its stability and reliability. A slider at the bottom allows users to adjust the number of points considered in the approximation. As points are introduced, they are colored according to their respective quality values.

Tab 2: Path-by-Path Analysis



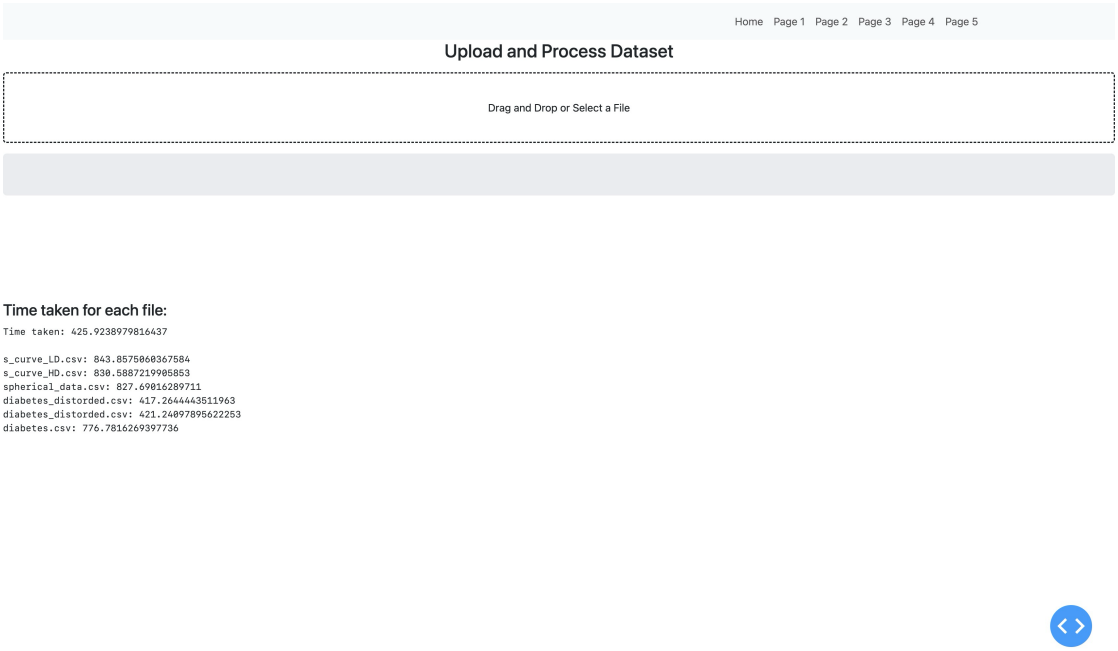
This tab enables detailed inspection of individual paths. By selecting a point in the LD space, all paths generated from that point are displayed along with their respective quality scores. Additionally, the quality graph is updated to reflect the contribution of the selected point.

Tab 3: Graph Representations



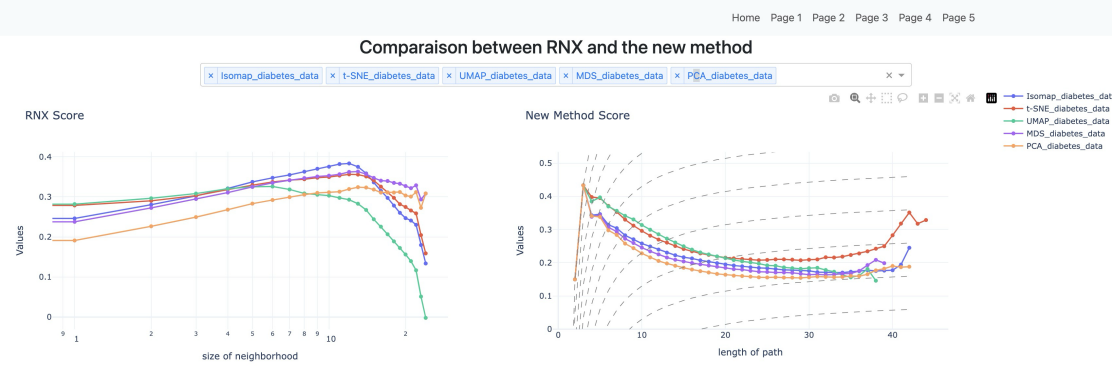
This tab performs a similar computation as the previous one, but for all points, presenting a total representation of the paths. However, due to the computational intensity of this process, it may take a significant amount of time to complete.

Tab 4: Dataset Upload and Processing



Users can upload their own HD dataset in this tab. The dashboard will process the data using Isomap, t-SNE, UMAP, MDS, and PCA, and compute the results for each representation. These results are stored locally and can be used in all other sections of the dashboard.

Tab 5: Comparison with $R_{NX}(K)$ Metrics



Dataset	RNX	New RNX	Rank RNX	Rank New RNX
Isomap_diabetes_data	0.3006495233335131	0.21675293794538028	4	2
t-SNE_diabetes_data	0.30528521014424226	0.2605081256121146	2	3
UMAP_diabetes_data	0.23971510956903072	0.2391848202361036	1	1
MDS_diabetes_data	0.3226213844714307	0.2111441884538407	5	4
PCA_diabetes_data	0.28860912687484985	0.19598595259027957	3	5

This tab offers a side-by-side comparison of the results generated by the thesis techniques and the $R_{NX}(K)$ metrics. It also displays the overall quality scores, computed as averages over the entire curve, and the order in which these quality values appear.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**