Large Datasets Embedding and Visualization

Mateusz Smendowski, Michał Grela

June 2022

1 Introduction

Visualization makes it easier to understand and notice dependencies in the high dimensionality data that are not trivial to capture and perceive. It is an inseparable, far-reaching, and effectual concept of data analysis or its initial recognition, but also an autonomous tool and dextrous field of machine learning. Visualization allows checking whether there are groups of similar observations forming clusters and finally gain more priceless intuition and understanding about data. In the case of multi and high-dimensional ones, it is necessary to reduce their dimensions to at most three. The relationships in data are often non-linear, which rules out methods like PCA regarding separation quality (Figure 1). Therefore, it is required to use Manifold Learning techniques to discover the surface (manifold) on which the data is distracted and make reasonable projections into a space with the desired dimensionality.

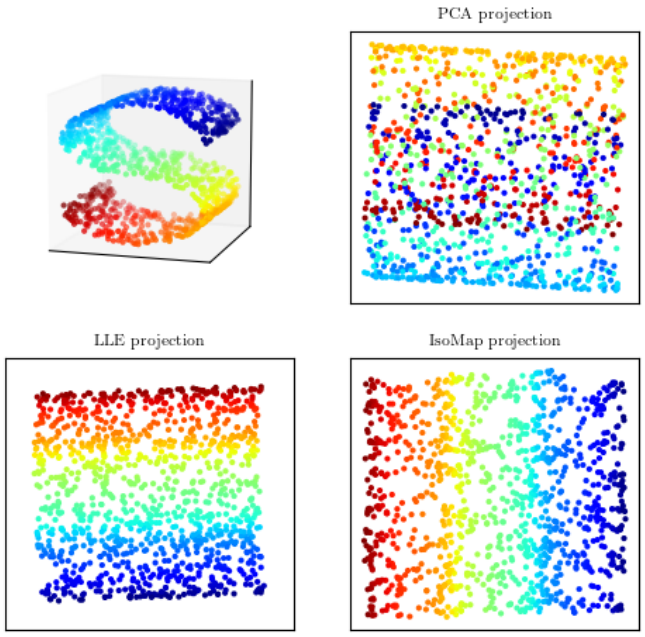

Figure 1: Comparison of PCA projection with manifold learning techniques..

This project aims to analyze and visualize the MINST, 20 News Groups, and RCV Reuters datasets using methods such as t-SNE, UMAP, ISOMAP, PaCMAP and IVHD. Therefore, the particular motivation is to show the concept of high-dimensional data visualization, assess multiple data embedding techniques, and highlight potential comparative criteria of data separation quality. This document is organized as follows: Section 2 introduces the used data sets. Subsequently, Section 3 includes theoretical descriptions of applied dimensionality reduction techniques and implemented metrics. The next part consists of visualizations and metric computations results. Finally, Section 5 states a space for a summary and conclusions.

2 Datasets description

In this section, analyzed datasets: MNIST, 20 News Groups and RCV Reuters are described.

2.1 MNIST

The MNIST (Modified National Institute of Standards and Technology database) dataset consists of the 70 000 size-normalized and centered images of hand-written digits (written by 500 different writers), where 60 000 samples make up the training set [1]. Typically, the remaining collection intends for testing. More specifically, the MNIST is a subset of a larger NIST dataset with a 28x28 size and black and white images of digits, which is a benchmark set not only for classification assessment but also for evaluating multiple methods of dimensionality reduction. Its great advantage is interpretability and intuitiveness. Because an average person can assess which of the digits are similar, it is possible to verify the data embedding techniques methods in terms of global separation and the reasonableness of the clusters created. Natural intuition suggests that representations of analogous digits such as 3 and 8 should be next to each other in the visualized space after dimensionality reduction. The size of the data set is 128 MB.

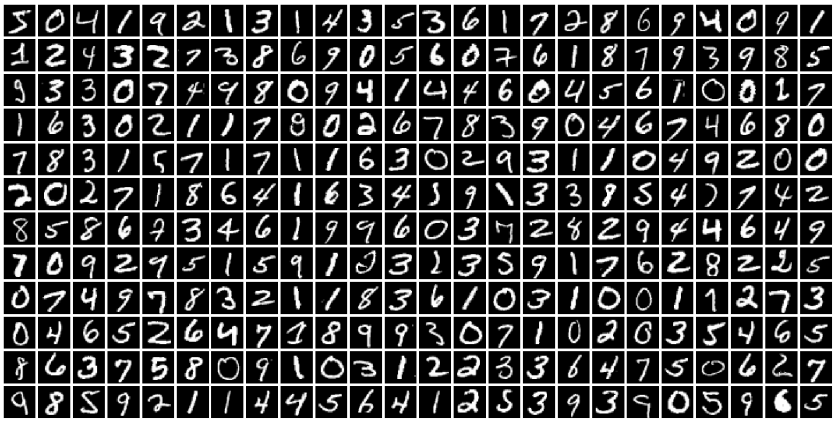

Figure 2: Visualization of sample digits from the MNIST data set.

2.2 20 News Groups

The 20NG (20 News Groups) dataset consists of 18 846 posts from 20 news groups [2]. Each post has a representation in the form of a TF-IDF 5000-word vector. The extracted TF-IDF vectors are very sparse, with an average score of less than .5% non-zero features. Furthermore, the collection includes 20 classes associated with thematic areas such as graphics, baseball, and electronics. Finally, the size of the entire data set is around 413 MB.

2.3 RCV Reuters

The RCV (Reuters Corpus Volume) Reuters dataset consists of 804 409 manually classified news [3]. The original dataset includes 103 classes and 47 236 dimensions. We have operated on its minimized version with attributes reduced to 30 via PCA (corpus of press articles preprocessed to 30D by PCA). Our investigation proved that exactly 59 documents have less or equal to 200 samples. In order not to introduce the phenomenon of noise that could significantly obscure and deteriorate the quality of the visualization, small classes have been removed, leaving the content of 24 unique documents. Consequently, the ultimate size of the data set occurred to be around 235 MB.

3 Used methods

In this section, used embedding methods and metrics to measure their quality are briefly described.

3.1 Embeddings

An embedding is a low-dimensional representation of high-dimensional data. Usually, it is impossible to fully map the structure of multidimensional data, but it is possible to map that is sufficient to solve a given problem. Many methods have been developed for data embedding. In this subsection, theoretical aspect of used embedding methods is presented. It is recommended to read the original papers and documentations to fully understand the issues presented in this subsection.

3.1.1 t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) was first proposed by Laurens van der Maaten and Geoffrey Hinton in 2008[4]. t-SNE is a variant of the Stochastic Neighbor Embedding (SNE), so in order to understand how t-SNE works it is necassary to introduce SNE method[5]. The method starts by finding the similarities between neighboring objects in high-dimensional space, which is determined by the conditional probability $p_{i|j}$ given by

$$p_{i|j} = \frac{\exp(-||x_i - xj||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{1}$$

where σ_i is either found by binary search for the value of it that gives probability distribution with a perplexity specified by the user or is set by hand. Whereas, in the low-dimensional space Gaussian neighborhoods are also used, but with a specified variance value set to $\frac{1}{2}$. So the conditional probability $q_{i|j}$ that object i picks j as its neighbor is given by the expression

$$q_{i|j} = \frac{\exp(-||y_i - yj||^2)}{\sum_{k \neq i} \exp(-||y_i - y_k||^2)} \tag{2}$$

In the SNE method, the cost function is defined as sum of Kullback-Leibler divergences between distributions given by

$$C = \sum_i \sum_j p_{i|j} \log \frac{p_{i|j}}{q_{i|j}} \tag{3}$$

However classic SNE has two main problems. First, it is based on a cost function that is difficult to optimize. Second, it encounters a problem called "crowding problem", which is caused by Gaussian "short tail". In order to solve this problem, t-SNE uses the Student's t-distribution with a single degree of freedom, which is heavy-tailed distribution. So in t-SNE probability q_{ij} is given by

$$q_{i|j} = \frac{(1 + ||y_i - y_j||^2)^-1}{\sum_{k \neq l}(1 + ||y_k - yl||^2)^-1} \tag{4}$$

Despite the fact that t-SNE gives quite good results, this method also has its disadvantages. First, it is sensitive to the curse of the intrinsic dimensionality of the data, due to its local nature. The second, which is the major problem of this method, is the necessity to select several optimization parameters as the cost function is not convex.

3.1.2 UMAP

Uniform Manifold Approximation and Projection (UMAP) is a modern technique for dimension reduction developed by Leland McInnes, John Healy, James Melville in 2018[6]. It is characterized by faster run time and better preservation of the global structure of data than t-SNE. The UMAP is based on three assumptions about the data:

1. The data is uniformly distributed on a Riemannian manifold
2. The Riemannian metric is locally constant
3. The manifold is locally connected

The first step of UMAP algorithm is constructing a fuzzy topological representation. In order to construct the initial high-dimensional graph, UMAP creates a fuzzy simplicial complex, which can be described as a graph with edge weights representing the likelihood that two points are connected. From each point a radius is determined which, when superimposed on a different radius, makes connection between those points. In UMAP the radius is chosen locally, based on the distance to each point's n nearest neighbor. Each point must be connected to at least one closest neighbor. The prepared fuzzy topological representation of the data is used to convert it into a low dimensional representation. It is achieved by optimizing the layout of a low dimensional analogue to be as similar as possible[7].

UMAP allows to easily control the balance between the local and global structure, thanks to the used parameters. The parameter that has the greatest influence on it is `n_neighbors`, which defines the size of the local neighborhood UMAP will look at when attempting to learn the manifold structure of the data. Low values of this parameter lead to concentrating on local structures, while large values contribute to more global structures. The second most important parameter is `min_dist`, which controls tightness of data in low dimensional space. It determines the minimum distance apart that points are allowed to be in the low dimensional representation[8].

Despite giving very good results, UMAP also has some disadvantages. This method is characterized by a lack of interpretability, dimensions of the embedding space have no specific meaning unlike, for example, Principal Component Analysis (PCA). UMAP can tend to find manifold structure within the noise of a dataset. With a small sample of noisy data, structure obtained from noise is more likely, as more data is sampled the amount of structure evident from noise tends to decrease[6].

3.1.3 ISOMAP

ISOMAP is one of the earliest approaches to manifold learning. It can be described as extension of Kernel PCA or Multi-dimensional Scaling (MDS). It extends the classical multidimensional scaling (MDS) method by exploiting the use of geodesic distances of the underlying nonlinear manifold. The working principle of the algorithm can be briefly described in the following steps:

1. Determine k neighboring points and create neighborhood graph
2. Find the shortest path in the graph, usually in this step Dijkstra's algorithm or Floyd-Warshall algorithm is used
3. Construct a d-dimensional embedding by a partial eigenvalue decomposition. The embedding is encoded in the eigenvectors corresponding to the d largest eigenvalues of the $N \times N$ ISOMAP kernel[9]

If the manifold contains holes and is not well sampled, Isomap perform poorly. Slightly wrong parametrization can produce bad results, due to the neighborhood graph creation[10].

3.1.4 PaCMAP

Pairwise Controlled Manifold Approximation Projection (PaCMAP) is another graph construction-based method. After the graph is created, the solution is initialized and iterative optimization with gradient descent algorithm. In this method, in graph construction, it is crucial to distinguish edges into three types: neighbor pairs, mid-near pairs, and further pairs. The first group consists of the n_{NB} nearest neighbors from each observation in the high-dimensional space. PaCMAP uses a scaled distance metric given by the expression

$$d_{ij}^{2,select} = \frac{||x_i - x_j||^2}{\sigma_i \sigma_j} \tag{5}$$

Where σ_i is the average distance between i and its Euclidean nearest fourth to sixth neighbors. The calculated distances $d_{ij}^{2,select}$ are only used for neighbors selection. The second type includes $N \cdot n_{MN}$ pairs selected by taking 6 additional, random observations from each observation and using the second smallest of them as a mid-near pair. Finally, the random selected n_{FP} further points from each observation are chosen. The number of mid-near and further points is determined by the specified parameters.

In PaCMAP method for each type of pair, different loss function is used

$$Loss_{NB} = \frac{\overline{d}_{ij}}{10 + \overline{d}_{ij}}, Loss_{MN} = \frac{\overline{d}_{ik}}{10 + \overline{d}_{ik}}, Loss_{FP} = \frac{1}{1 + \overline{d}_{il}} \tag{6}$$

Where $\overline{d}_{ik} = ||y_a - y_b||^2 + 1$. The pairs are further weighted by the coefficients w_{NB}, w_{MN} and w_{FP}. These weights are updated dynamically during the algorithm as part of the optimization process.

The optimization process is divided into three phases, designed to avoid local optimas. In the first phase, the goal is to improve placement of embedded points in such a way as to preserve both the global and local structures, but mainly the global structure. In the second phase, the goal is to improve the local structure while maintaining the global structure. In the last, third phase, local structure is improved. This stage, as the authors pointed out, seems to have a larger effect on datasets with primarily local structure compared to datasets with global structure[11].

3.1.5 IVHD

Interactive Visualization of High-Dimensional Data (IVHD) is a modern and very fast method. This technique is the simplified force-directed implementation of the MDS. To obtain linear-time and memory complexity of data embedding, only limited number of distances from high-dimensional space and corresponding distances in low-dimensional space are used. The first step in IVHD is to create a graph of nearest neighbors (*nn-graph*). The main assumption, on which IVHD is based, is that the small number of nearest neighbors (nn) is needed to get a good approximation of data manifold. However, small number of nn leads to largely unconnected *nn-graph*. So apart from nn edges, randomly selected neighbors (rn) are connected. To obtain 2D embedding, the stress function is minimized, this function is given by an equation

$$E(||D - d||) = \sum_i \sum_{j \in (O_{nn}(i) \bigcup O_{rn}(i))} w_{ij}(\delta_{ij} - d_{ij})^2 \tag{7}$$

which is similar to the stress function used in classical MDS. $O_{nn}(i)$ and $O_{rn}(i)$ are the sets of indices respectively the nearest neighbors and random neighbors of a feature vector i. Also, to reduce memory load and computational complexity binary distances between data vectors are used, for example in the

following way

$$\forall y_i \in Y \begin{cases} \delta_{ij} = 0, & j \in O_{nn}(i) \\ \delta_{ij} = 1, & j \in O_{rn}(i) \end{cases} \tag{8}$$

The results given by IVHD are good, but obtained local neighborhood by this method is not as good as results given by other SNE clones. However, this method is worth attention in view of its computational and memory complexity, and the small number of required parameters[12].

3.2 Metrics

Metrics and measures are required to evaluate dimensionality reduction (DR) techniques and asses both local-neighborhood preservation and global structure holding. Multiple of them give a clear picture of the data separation quality. At the assessment stage of the dimensionality reduction technique, it is vital not only to attempt a visual interpretation but also to verify the metrics values to compare many embedding methods with each other. Table 1 represents selected methods available for evaluating the quality of DR algorithms. The already available metrics mustn't constitute a limitation. They can even become an inspiration for their extension and the implementation of custom metrics for assessing the quality of dimensionality reduction.

Table 1: Summary of methods for evaluating the quality of DR algorithms [13].

Name of the metric	**Criterion**
Sheppard Diagram	Global
Kruskal Stress Measure	Global
Sammon Stress	Global
Spearman's Rho	Local
Topological Product	Local
Topological Function	Local
Residual Variance	Global
Konig's Measure	Local
Trustworthiness	Local
Local Continuity Meta-Criterion	Local
Agreement Rate/Corrected Agreement Rate	Local
Mean Relative Rank Errors	Local
Procrustes Measure/Modified Procrustes Measure	Local
Co-ranking Matrix	Local
Global Measure	Local and Global
The Relative Error	Global
Normalization independent embedding quality assessment	Local/Global

3.2.1 Distance matrix-based

The first custom metric implemented is a measure that uses distance matrices. For each class included in the embedded data, we calculate the ratio between the mean Euclidean distance for samples with the same labels and points targeted differently. Finally, the metric's value is the average of the obtained partial results for each class. Distance matrix-based metric allows to recognize the strength of the relationship between points with the same labels and therefore know how well and compactly the clusters are formed - if samples from other classes appear among the point group, they weaken the purity of cluster. The lower the value, the better the quality of local separation - more tight clusters.

3.2.2 Distance matrix-based with KMeans optimization

Distance matrix-based metric with KMeans optimization approximate inter-class distances by distances between centroids (KMeans clusters centers). The obtained results are very similar to the form in which the Distance Matrix is calculated for all points. The advantage of this version is a much lower time complexity, with the risk of slight inaccuracies from the approximations taken.

3.2.3 Co-ranking matrix

Assessing the quality of dimensionality reduction can be performed by co-ranking matrix-based methods. Based on the co-ranking matrix itself, some of the most significant statistics that could be indicated are:

- QNX - Average Normalized Agreement Between K-ary Neighborhoods [14] - measure the quality of the data embedding technique in terms of how well it preserves the local neighborhood around observations. For a given value of k, the k closest points for each sample are retrieved. QXN is the number of shared neighbors between the original dimensionality and the reduced one, additionally normalized by k. QNX yields values from 0 to 1. 1 inform about supreme neighborhood preservation. On the other hand, the QNX value is close to 0 when there is no neighborhood preservation.
- RNX - Rescaled Agreement Between K-ary Neighborhoods [15] - is the scaled version of QNX. The RNX measures the quality of data embedding technique in terms of the shared number of k nearest neighbors. RNX yields values from 0 to 1. 1 means that the neighborhoods are supremely preserved. On the contrary, 0 means that the embedding resemble the random one.
- AUC - Area Under RNX Curve [16] - as with the aforementioned two statistics, the value of one indicates the neighborhood preservation is ideal after dimensionality reduction.

3.2.4 DR quality and KNN gain

DR quality and KNN gain are metrics derived from the co-ranking matrix and nearest neighbors approach. They are kept in a very similar convention to the previously presented metric. Their values depend on the size of the neighborhood, thanks to which it is possible to assess which of the dimensional reduction techniques works either best locally or globally. DR quality indicates the fidelity of the neighborhood representation based on the already mentioned RNX measure. Furthermore, KNN Gain means the fraction depicting the amount of additional samples of the same class contained in a neighborhood of a given size in a reduced dimensionality compared to the original one.

3.2.5 Shepard diagram

The Shepard diagram is a scatterplot of the distances between points in the original high dimensionality and the reduced one, deftly visualizing the goodness-of-fit [17]. Typically, the x-axis represents primary distances in the space, while the y-axis indicates those after dimensionality reduction. The accurate dimensionality reduction technique produces a straight line. However, as dimensionality reduction is naturally associated with the lossy compression of information, an ideal straight will be rare to observe. Finally, the greater the dispersion around the straight line, the weaker preservation of the distances between the samples in the original space and the reduced one.

3.2.6 Trustworthiness

Trustworthiness is a numeric metric with values in the range [0, 1] and expresses how well the DR technique preserves the local structure and to what extent honors pairwise distances between samples [18]. The default metric to compute distances between point pairs is the euclidean one. However, there are many other possibilities to calculate pairwise distances - cosine distances.

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i^k} \max(0, (r(i, j) - k)) \tag{9}$$

Trustworthiness requires specifying the number of considered neighbors. Each invalid neighbor (unexpected nearest neighbor present in the lower dimensionality that was not in the original space) reduces the value of the metric and, therefore, lowers the reliability of the DR technique in terms of quality of local separation. The neighborhood ($\mathcal{N}_i^k$) of each sample `i` in the reduced dimensionality is compared with r(i, j)-th nearest neighbors for each sample `j` in the original space.

3.2.7 Spearman correlation based

The last of the custom metrics use the Spearman rank correlation that indicates monotonic relationships (also non-linear ones) between variables and varies between -1 and +1, with 0 implying no correlation [19]. In particular, it can highlight dependencies between low-dimensional and original data (measure the relationship between two datasets). For each class in the dataset, computations include the squares of euclidean distances between the samples for both dimensionalities. Then, the correlation coefficient investigates relations between the vector representations of the obtained matrices. The final value of the metric is the average of the values obtained for each class separately. Finally, the greater it is, the more accurate the data embedding technique keeps similarities of the distances within the same target class.

4 Visualization and analysis

In this section, visualizations of obtained embeddings in two dimensions of analyzed datasets are shown. Moreover, results and interpretation of used metrics are presented. Initially we wanted to use ISOMAP for every dataset, but ultimately because of its high computational cost, we changed it to a PaCMAP method for MNIST and RCV Reuters datasets. The plots have been squeezed to provide an easy comparison. A complete and more robust visualization of each embedding can be found on the GitHub repository[20].

4.1 MNIST

In this subsection, analysis of MNIST dataset is presented. A code which was used to get these results can be found in our repository [21].

4.1.1 Visualization of embeddings

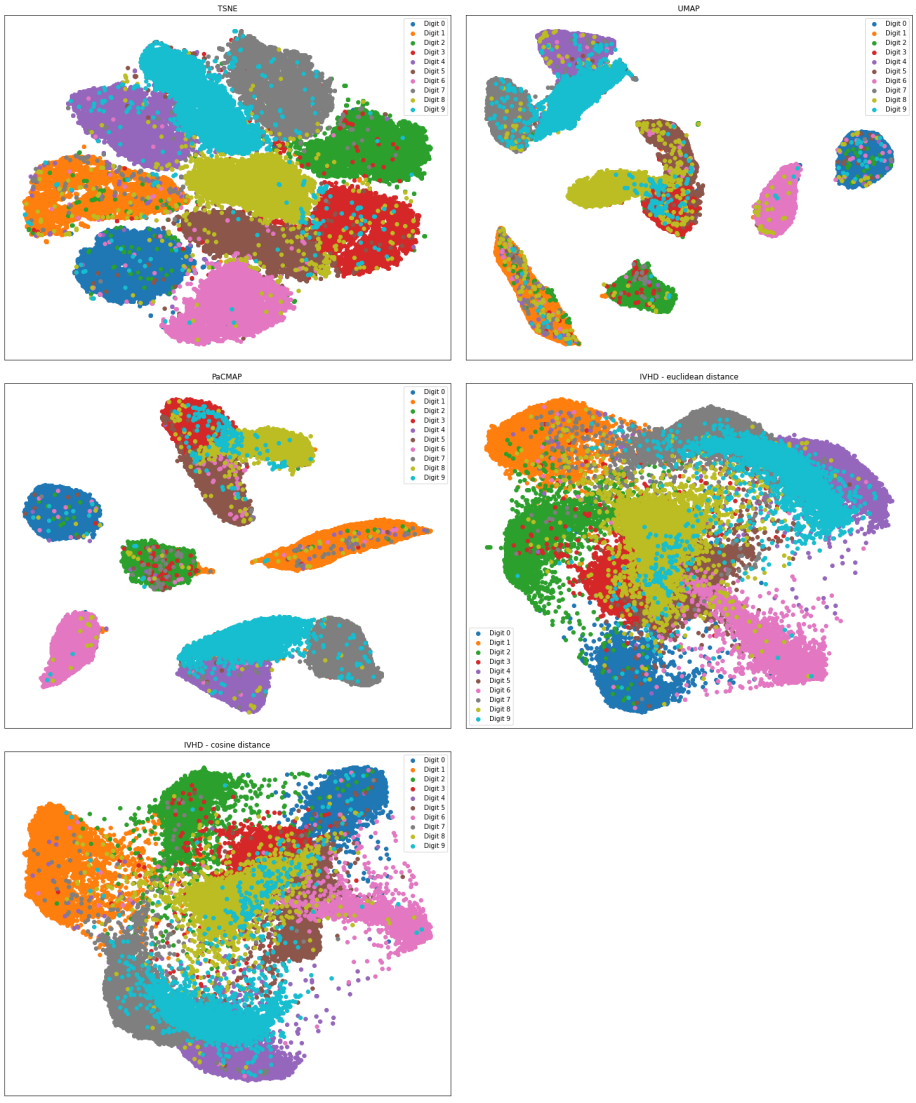

Figure 3: MNIST visualizations using t-SNE, UMAP, PaCMAP and IVHD.

MNIST dataset contains handwritten digits, so we have good intuition how this separation should look like. The default parameterization was used for t-SNE, UMAP and PaCMAP methods. In the case of IVHD, the parameters were set as follows: number of nearest neighbors - 3, number of random neighbors - 2, number of iterations - 5000, rest were set by default. As it can be seen in Figure 3, good local separation has been achieved for each method. UMAP and PaCMAP methods reflect global separation best. We can tell that there are two three-elements clusters: first includes digits 3, 5 and 8, second includes digits 4, 7, 9. And the rest of the numbers are far from the rest. It coincides with what our intuition would tell us. However, in addition to visual analysis, it is necessary to use metrics to efficiently assess the quality of each embedding.

4.1.2 Distance matrix-based metric

Table 2: Distance matrix-based metric.
Results for MNIST embeddings using t-SNE, UMAP, ISOMAP and IVHD.

t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
0.346	0.168	0.230	0.250	0.206

4.1.3 Distance matrix-based metric with KMeans optimization

Table 3: Distance matrix-based metric with KMeans optimization.
Results for 20 News Groups embeddings using t-SNE, UMAP, ISOMAP and IVHD.

t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
0.331	0.203	0.212	0.241	0.221

Distance matrix-based metric and its optimized version in terms of the execution time return similar results. However, the non-approximation one may be considered more reliable. It is worth recalling that this metric is the quotient of the average distances of points belonging to the same class and inter-class ones. Generally, if the mean distances between points with the same labels decrease, the resulting clusters are tighter and more compact. Furthermore, the value of the metric decreases as the average distances between classes increase, and thus the quality of global separation gets better. Therefore, UMAP yielded the most distinguished clusters. PaCMAP and IVHD turned out to be not much worse. Distance matrix-based metric indicated t-SNE as the worst in terms of the considered criterion.

4.1.4 DR quality and KNN gain

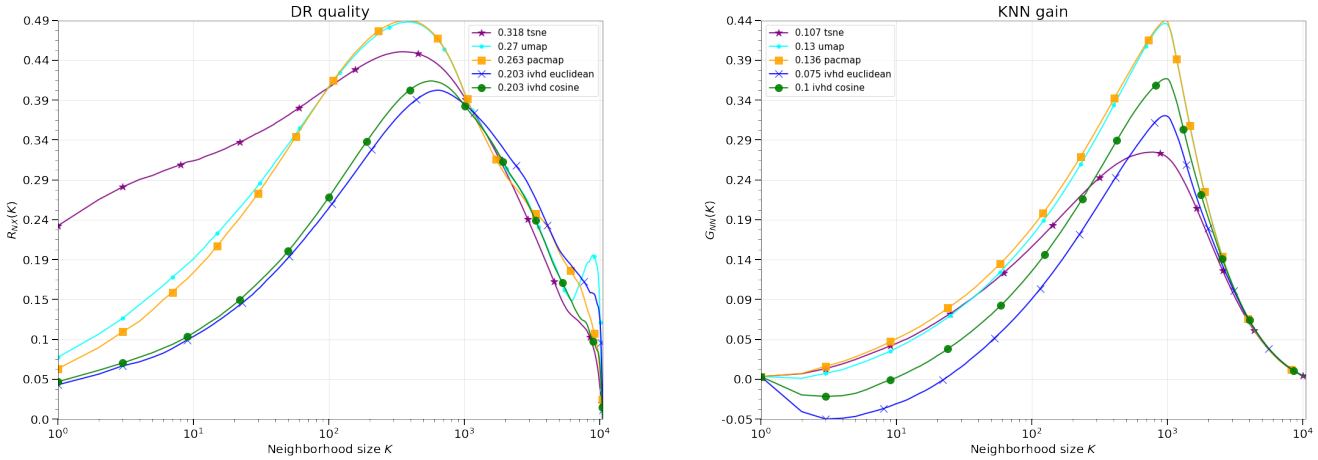

Figure 4: DR quality and KNN gain for MNIST

Referring to the DR quality plot, t-SNE provides the highest quality of local separation (up to neighborhood size equal to 10^2). In the next part of the chart, the dominance of the PaCMAP and UMAP occurs. Furthermore, for the bigger neighborhood size, t-SNE returns worse results than the other dimensionality reduction techniques. Finally, strengthening the analysis with the KNN gain chart, UMAP and PaCMAP performed best.

4.1.5 Co-ranking matrix

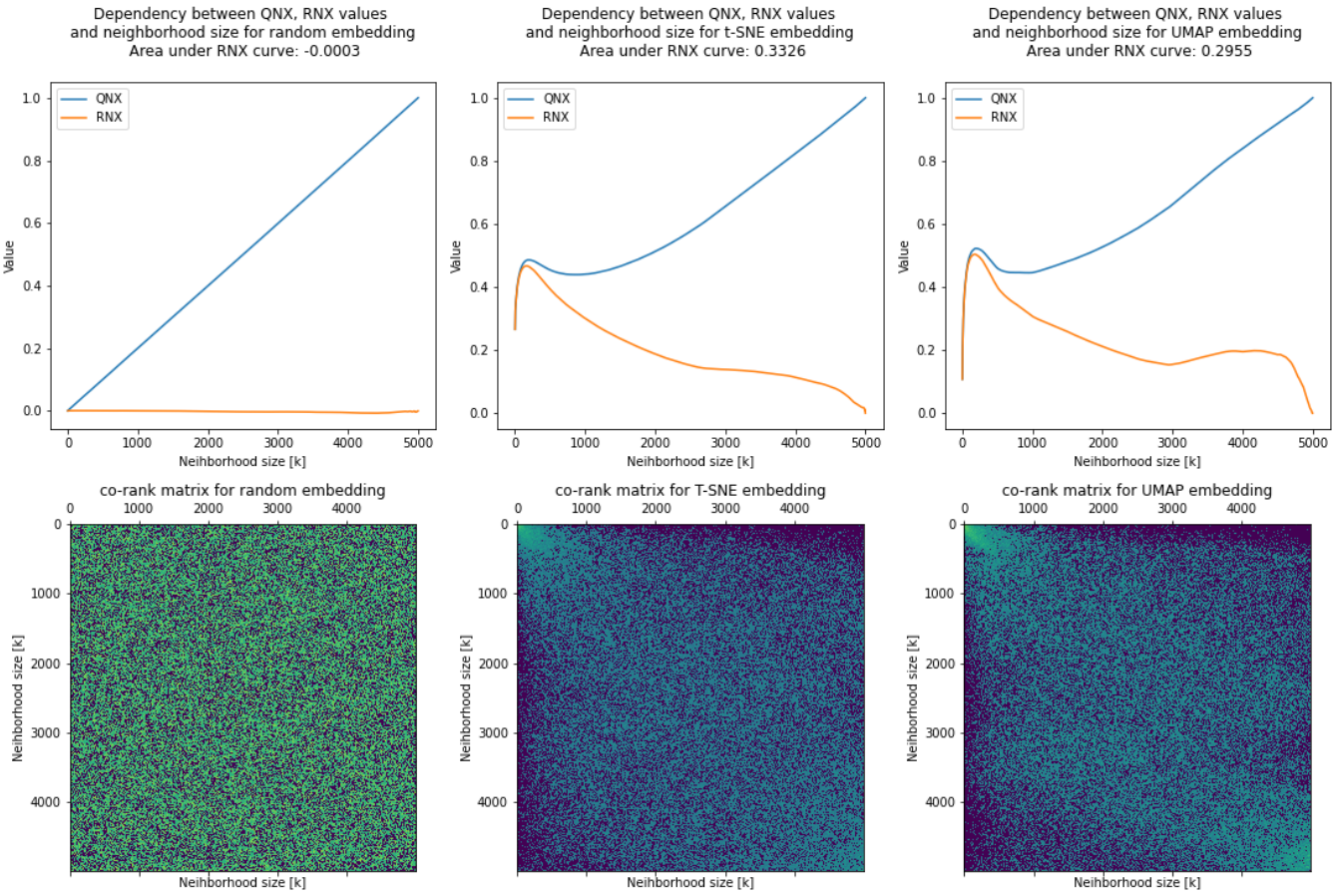

Figure 5: Co-ranking matrix-based metrics.
Results for MNIST embeddings using random embedding, t-SNE and UMAP.

Random embedding pretends the most inaccurate of possible mappings, where the co-ranking matrix visualization looks like complete noise. For real data embedding techniques, it is possible to observe a green cone created, indicating that the mapping to the low-dimensional space is not accidental. Since RNX is a scaled version of QNX, and the area under the RNX curve returns the exact numerical value, it can determine the process of analyzing the accuracy of dimension reduction techniques using this metric. Additionally, for each of the representations, the QNX curve converges to the value 1. It is natural because assuming that P indicates the number of the nearest neighbors and there are P samples in the original dimensionality and reduced one, computations qualify each sample as a neighbor.

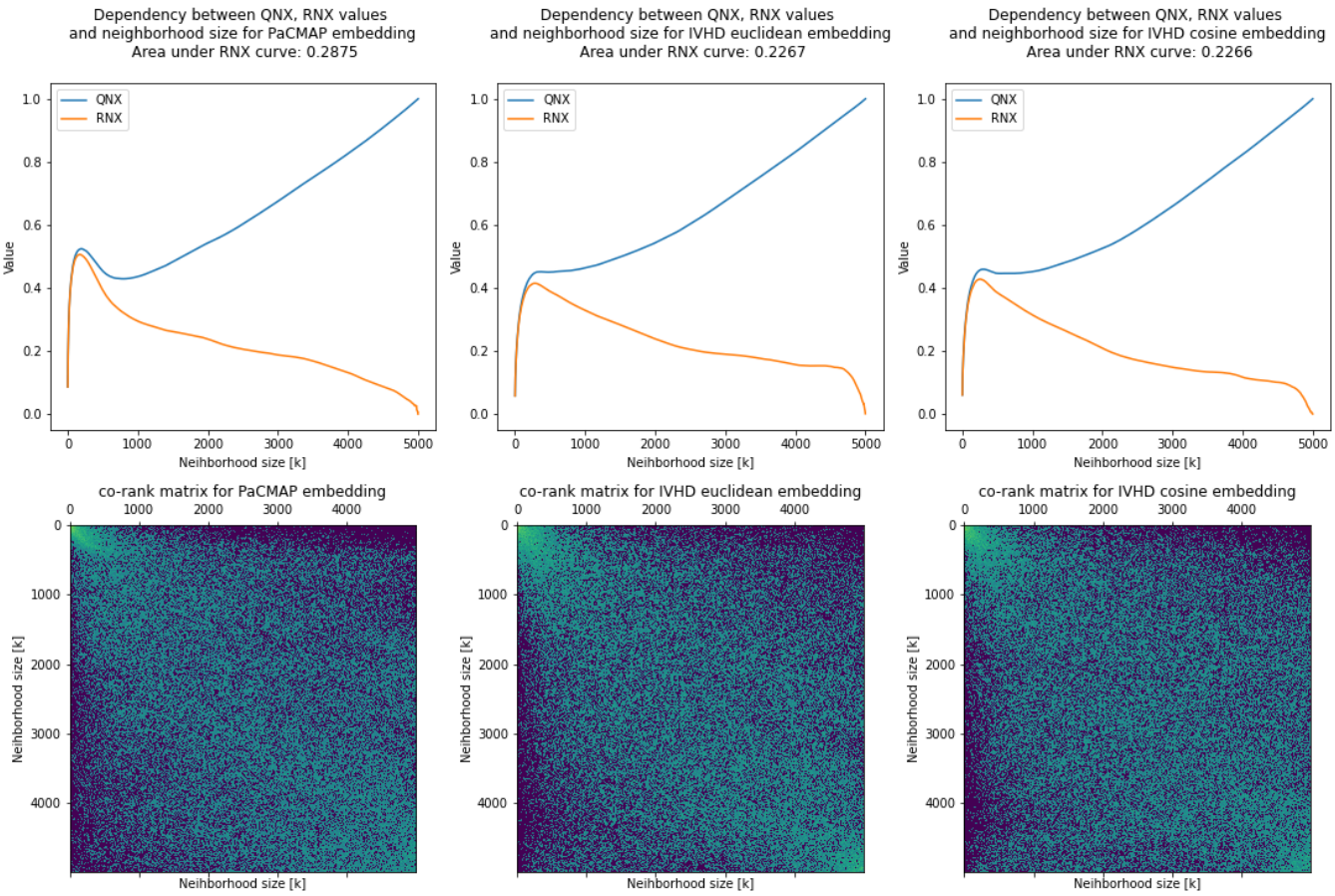

Figure 6: Co-ranking matrix-based metrics.
Results for MNIST embeddings using PaCMAP and IVHD.

The largest area under the RNX curve is for t-SNE, which means that t-SNE achieves the best results for the local neighborhood (small K size).

4.1.6 Shepard diagrams

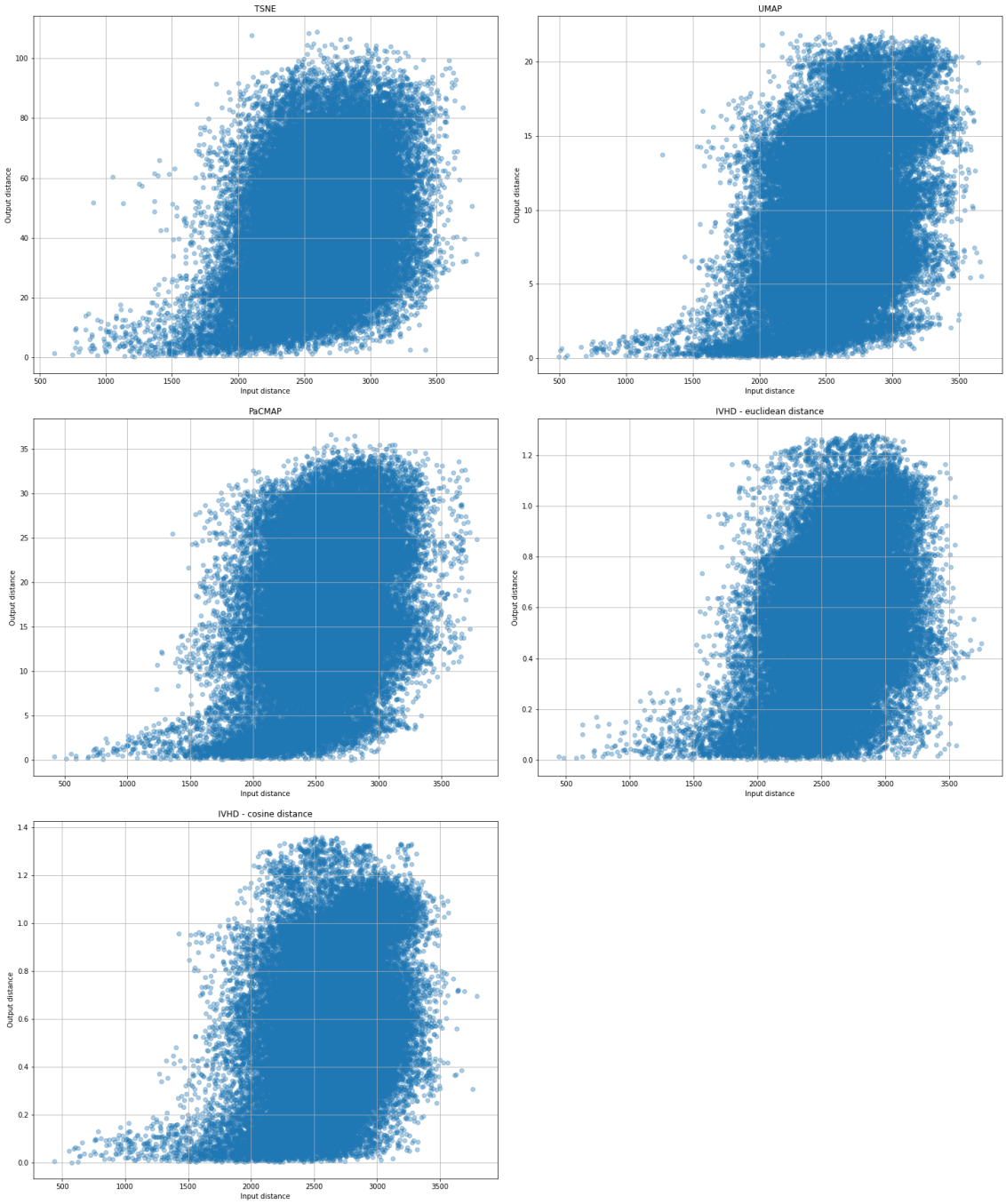

Figure 7: Shepard diagrams of MNIST embeddings using t-SNE, UMAP, PaCMAP and IVHD.

Ideally, the points in the Sheppard diagram should be on a straight line which would mean that the distances in the reduced dimensionality are perfectly rescaled values from the original space. We can see that the points are rather heavily scattered around than spread along the straight line. For PaCMAP and UMAP we can notice that the group of points is the most compact and there are the fewest points far away from the straight line.

4.1.7 Trustworthiness

Table 4: Trustworthiness with euclidean metric for pairwise distances.
Results for MNIST embeddings using t-SNE, UMAP, ISOMAP and IVHD.

K	t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
5	0.980	0.959	0.957	0.887	0.896
15	0.970	0.955	0.956	0.884	0.893
50	0.948	0.947	0.946	0.877	0.889
100	0.927	0.936	0.936	0.870	0.883
150	0.908	0.924	0.924	0.863	0.875

The high values of the trustworthiness metric obtained for all embedding techniques indicate that each method coped relatively well with local separation. This conclusion coincides with the results of visualizations, where each class creates a distinctive cluster. Consistency in good results should be taken into account, which indicates that UMAP and PaCMAP methods were the best at maintaining local relationships between points.

Table 5: Trustworthiness with cosine metric for pairwise distances.
Results for MNIST embeddings using t-SNE, UMAP, ISOMAP and IVHD.

K	t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
5	0.984	0.969	0.967	0.892	0.910
15	0.975	0.967	0.965	0.890	0.909
50	0.959	0.961	0.960	0.885	0.905
100	0.941	0.953	0.953	0.879	0.900
150	0.924	0.943	0.944	0.873	0.892

4.1.8 Spearman correlation-based metric

Table 6: Spearman correlation-based metric.
Results for MNIST embeddings using t-SNE, UMAP, PaCMAP and IVHD.

t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
0.666	0.683	0.700	0.545	0.575

The metric based on Spearman's correlation places t-SNE, UMAP, and PaCMAP on the podium, but the last of them turned out to be the most faithful in maintaining the correlation between the distances in both dimensions.

4.2 20 News Groups

In this subsection, analysis of 20 News Groups dataset is presented. A code which was used to get these results can be found in our repository[22].

4.2.1 Visualizations

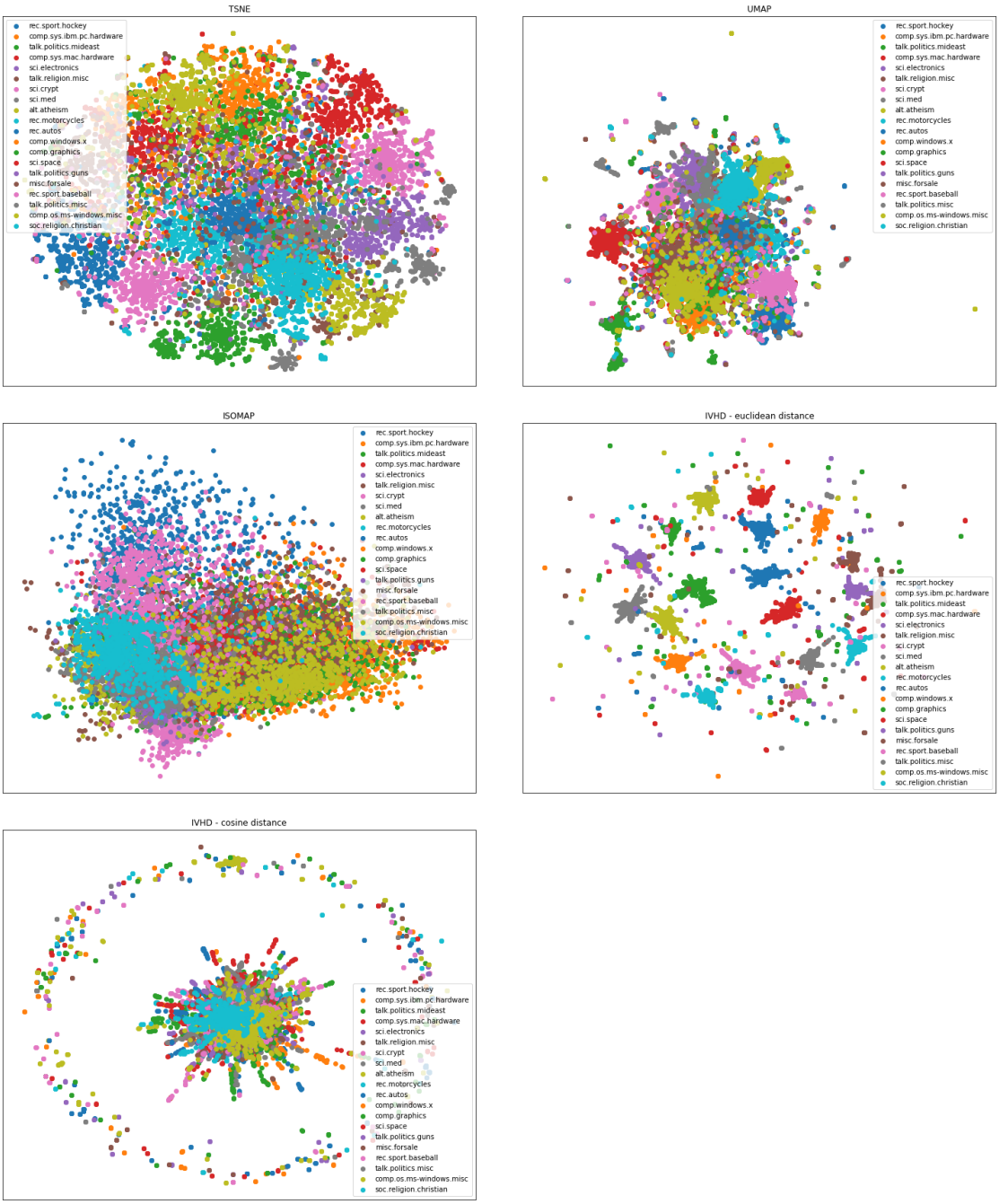

Figure 8: 20 News Groups visualizations using t-SNE, UMAP, ISOMAP and IVHD.

In case of this dataset in t-SNE method perplexity parameter was set to 15, in other method default parametrization was used. IVHD with cosine distance metric gave worst results, but with Euclidean distance metric obtained results were very good separated globally and locally. For different datasets, different distance metrics give better results. As we can see in Figure 8 every method gives good local separation. Global separation is only easy to see in IVHD with Euclidean distance metric, in other methods this separation is a bit blurry. Also, IVHD creates some noise in embedding.

4.2.2 Distance matrix-based metric

Table 7: Distance matrix-based metric.
Results for 20 News Groups embeddings using t-SNE, UMAP, ISOMAP and IVHD.

t-SNE	UMAP	ISOMAP	IVHD euclidean	IVHD cosine
0.574	0.639	0.637	0.174	0.314

4.2.3 Distance matrix-based metric with KMeans optimization

Table 8: Distance matrix-based metric with KMeans optimization.
Results for 20 News Groups embeddings using t-SNE, UMAP, ISOMAP and IVHD.

t-SNE	UMAP	ISOMAP	IVHD euclidean	IVHD cosine
0.547	0.556	0.582	0.182	0.291

Distance matrix-based metric indicated IVHD euclidean as unrivaled in terms of the quality of the created clusters for samples from the same classes. According to this metric, it is precisely after the application of IVHD cosine that the most compact clusters can be obtained.

4.2.4 DR quality and KNN gain

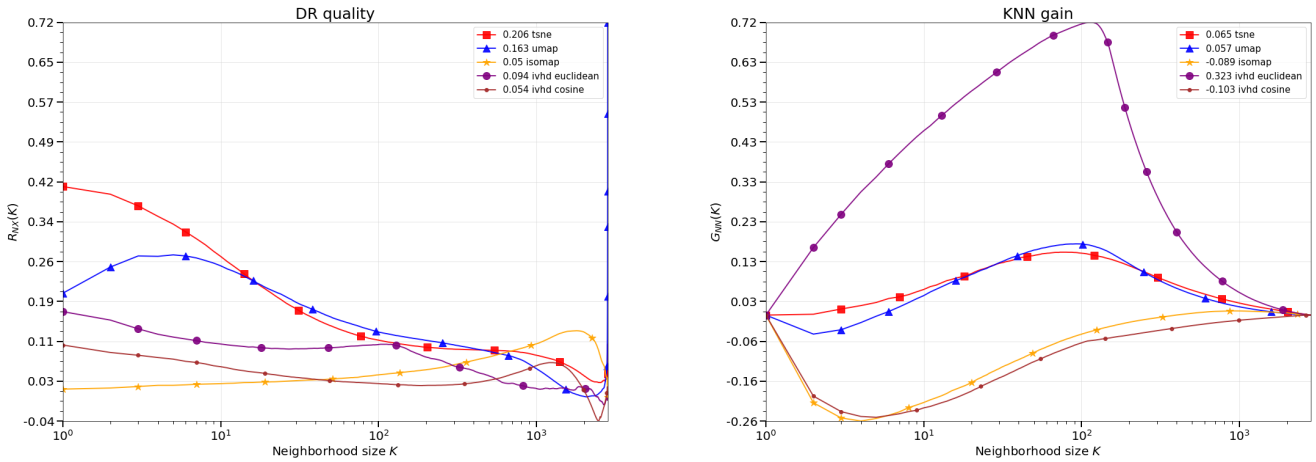

Figure 9: DR quality and KNN gain for 20 News Groups

The KNN gain graph clearly illustrates the contrast between IVHD euclidean and IVHD cosine. The first variant allows to achieve by far the best separation - the resulting clusters are separated from the others. On the other hand, the IVHD cosine failed at outlining dependencies from the original space. Again, t-SNE proved to be the best for local separation. However, when confronting the results jointly with visualizations and both metric charts, IVHD euclidean indisputably brings the best result of global separation.

4.2.5 Co-ranking matrix

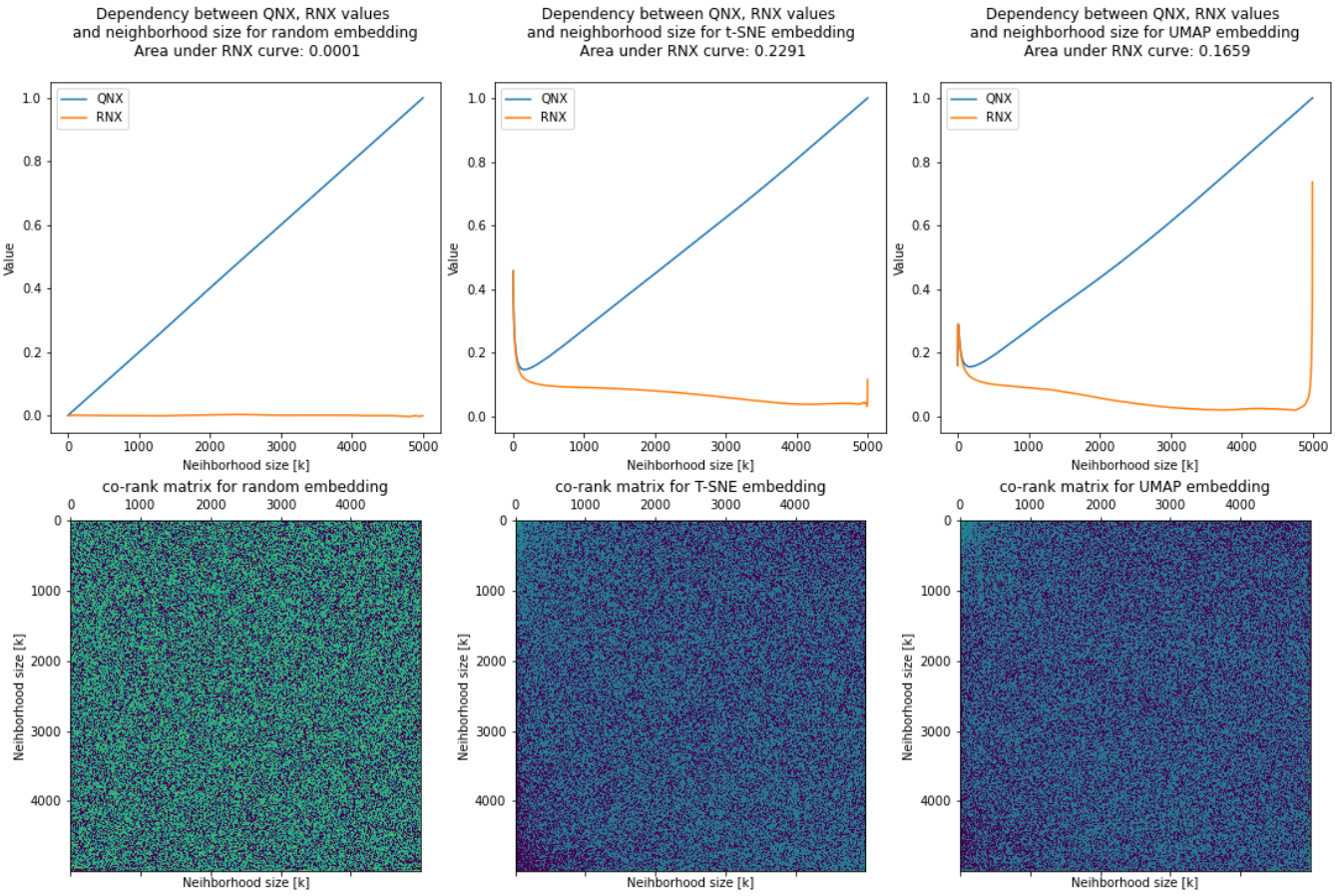

Figure 10: Co-ranking matrix-based metrics.
Results for 20 News Groups embeddings using random embedding, t-SNE and UMAP.

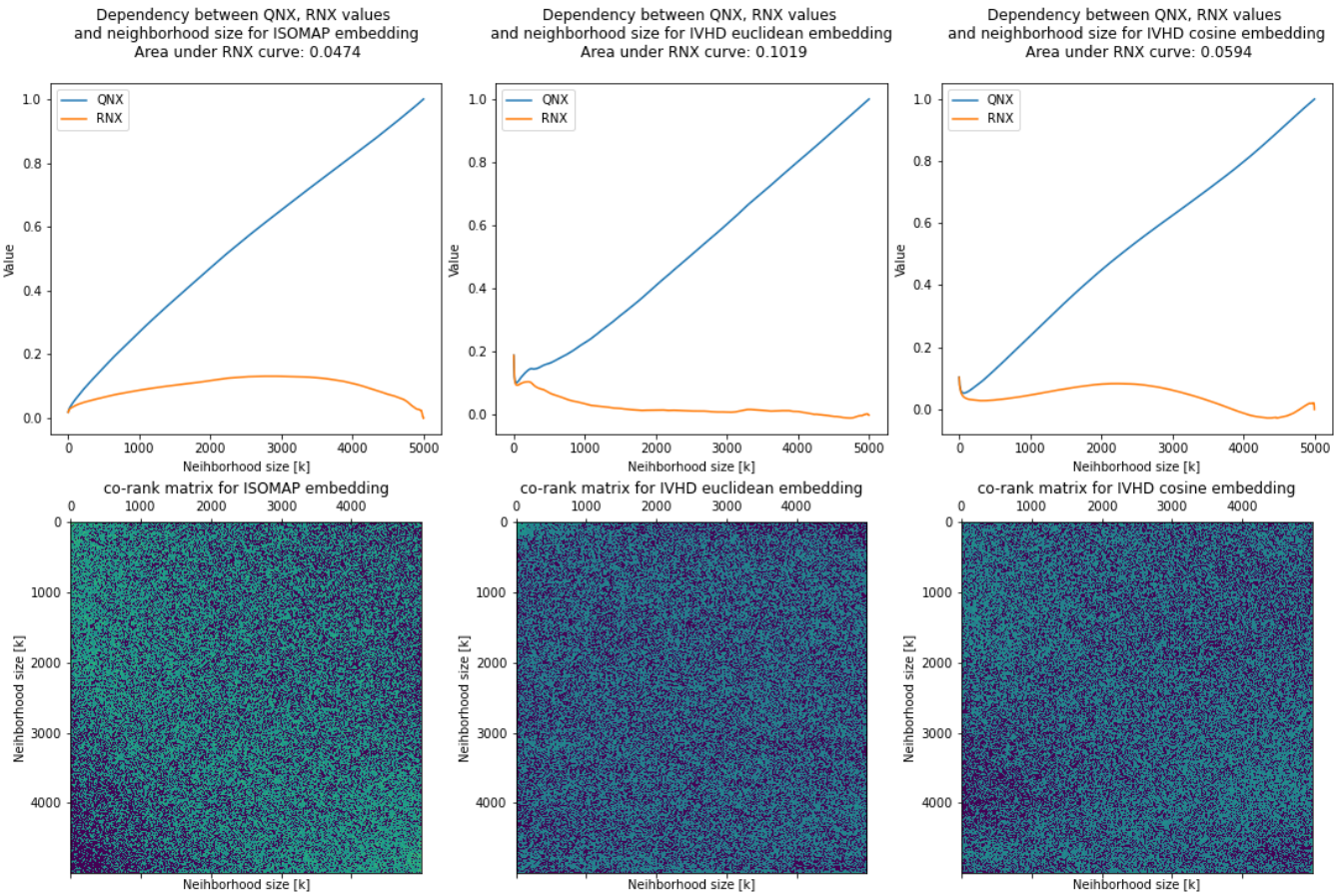

Figure 11: Co-ranking matrix-based metrics.
Results for 20 News Groups embeddings using ISOMAP and IVHD.

The results indicate a significant challenge for data embedding techniques to preserve the local neighborhood in the reduced dimensionality. However, the best results are for t-SNE (the initial most steep trend of the RNX curve) and simultaneously the highest value of the area under the RNX curve.

4.2.6 Shepard diagrams

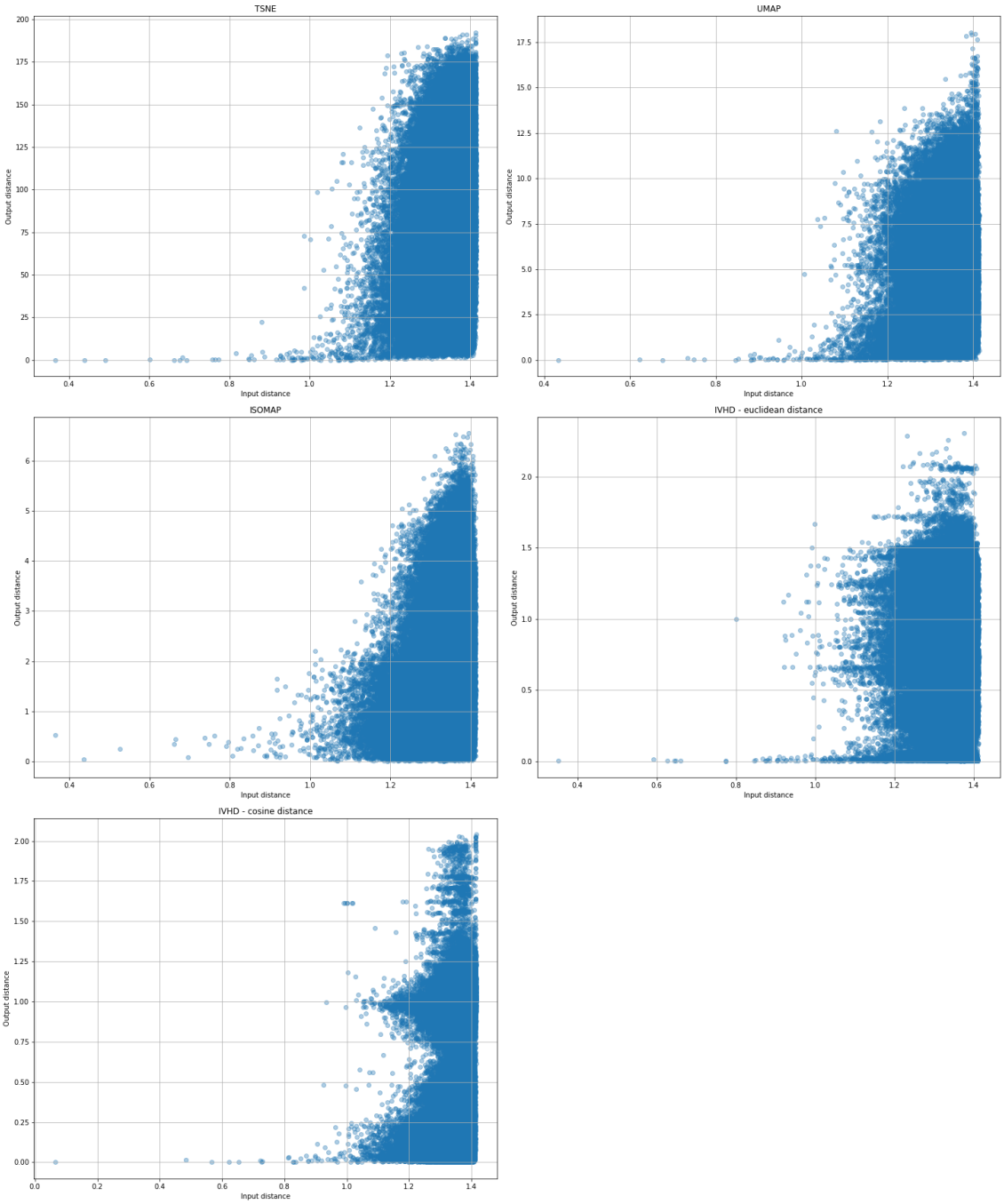

Figure 12: Shepard diagrams of 20 News Groups embeddings using t-SNE, UMAP, ISOMAP and IVHD.

The Shepard diagram for all techniques except IVHD has the most concise and least jagged form. In summarizing, all scatter plots are far from ideal. Finally, in the case of IVHD euclidean, it is possible to observe a lot of points detached from the group, which may indicate that, apart from the distinguished clusters, many points identifies as noise.

4.2.7 Trustworthiness

Table 9: Trustworthiness with euclidean metric for pairwise distances.
Results for 20 News Groups embeddings using t-SNE, UMAP, ISOMAP and IVHD.

K	t-SNE	UMAP	ISOMAP	IVHD euclidean	IVHD cosine
5	0.880	0.838	0.574	0.704	0.583
15	0.784	0.775	0.570	0.670	0.559
50	0.678	0.686	0.571	0.648	0.541
100	0.638	0.650	0.573	0.643	0.537
150	0.662	0.634	0.575	0.636	0.536

Both for euclidean and cosine distances, trustworthiness yielded around identical values. For a small neighborhood, the t-SNE method obtained the best results, but the farther away, the gains resulting from IVHD euclidean can be observed.

Table 10: Trustworthiness with cosine metric for pairwise distances.
Results for 20 News Groups embeddings using t-SNE, UMAP, ISOMAP and IVHD.

K	t-SNE	UMAP	ISOMAP	IVHD euclidean	IVHD cosine
5	0.880	0.838	0.574	0.704	0.583
15	0.784	0.775	0.570	0.670	0.559
50	0.678	0.686	0.571	0.648	0.541
100	0.638	0.650	0.573	0.643	0.537
150	0.622	0.634	0.575	0.636	0.536

4.2.8 Spearman correlation-based metric

Table 11: Spearman correlation-based metric.
Results for 20 News Groups embeddings using t-SNE, UMAP, ISOMAP and IVHD.

t-SNE	UMAP	ISOMAP	IVHD euclidean	IVHD cosine
0.259	0.286	0.181	0.132	0.112

Relatively low metric values may indicate that the 20 News Groups turned out to be challenging to separate. However, the UMAP and t-SNE retained the strongest correlation between distances in the original space and reduced dimensionality.

4.3 RCV Reuters

In this subsection, analysis of RCV Reuters dataset is presented. A code which was used to get these results can be found in our repository[23].

4.3.1 Visualizations

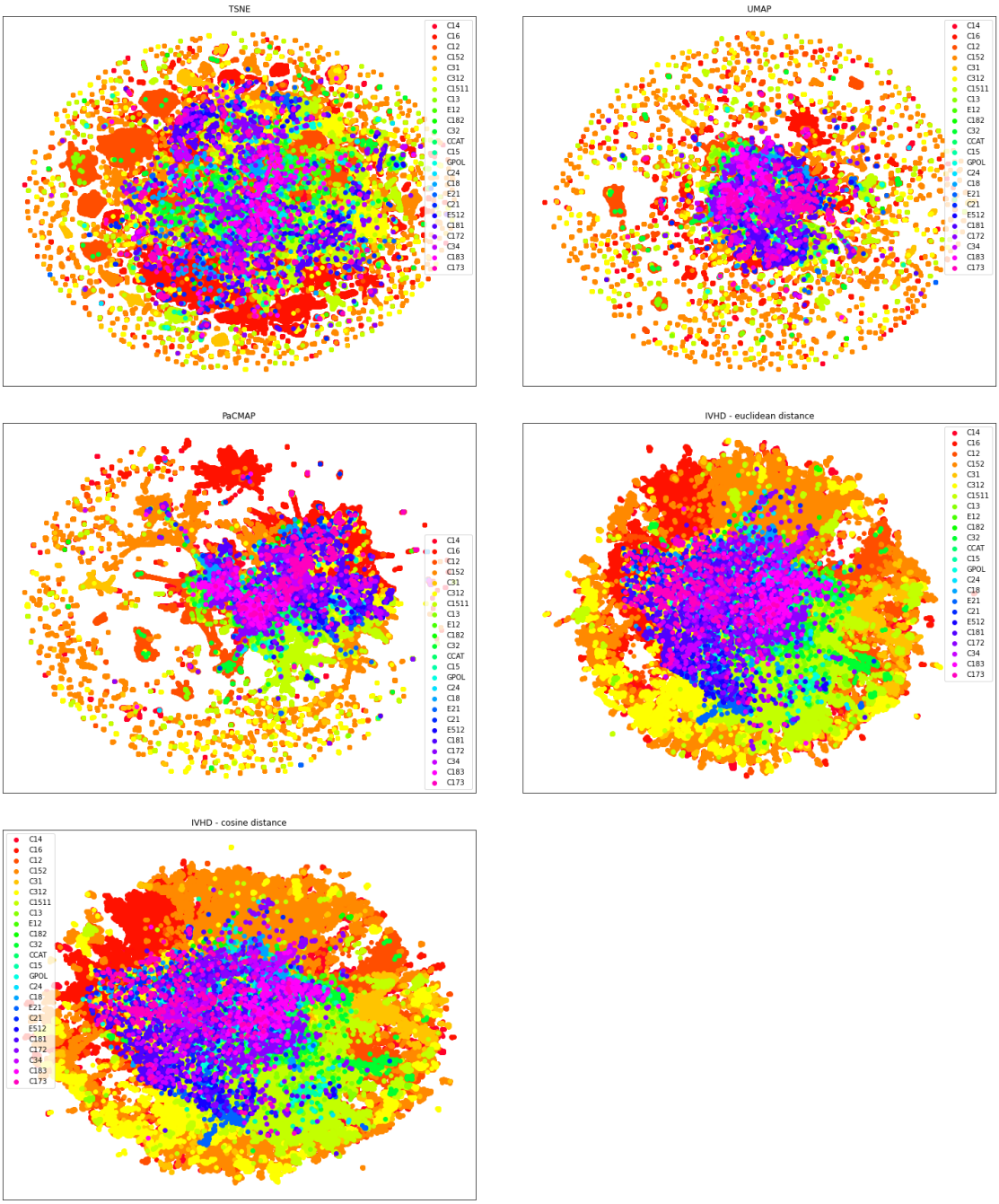

Figure 13: RCV Reuters visualizations using t-SNE, UMAP, PaCMAP and IVHD.

In case of this dataset the following parametrization was used: for t-SNE method perplexity was set to 40, for IVHD number of nearest neighbors was set to 5, number of iterations to 5000 and c to 0.1, other parameters were left default. RCV Reuters is a very complex dataset so as it can be seen in Figure 13 presented methods did not reflect the data structure very well. t-SNE, UMAP and PaCMAP methods leave a lot of points alone. Visually, the best local separation is given by IVHD. Neither of the methods reproduced global separation very well.

4.3.2 Distance matrix-based metric

Table 12: Distance matrix-based metric.
Results for RCV Reuters embeddings using t-SNE, UMAP, PaCMAP and IVHD.

t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
0.804	0.650	0.613	0.551	0.548

4.3.3 Distance matrix-based metric with KMeans optimization

Table 13: Distance matrix-based metric with KMeans optimization.
Results for RCV Reuters embeddings using t-SNE, UMAP, PaCMAP and IVHD.

t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
0.728	0.501	0.452	0.503	0.508

The metric based on the distance matrix indicates t-SNE as the one that yields the highest ratio between the mean distances of points in the same class and samples labeled otherly. In fact, in t-SNE, it is hard to distinguish regions where the concentration of samples occurs - it is an even distribution. By citing the value of distance matrix-based metrics without KMeans optimization, IVHD gets the best score.

4.3.4 DR quality and KNN gain

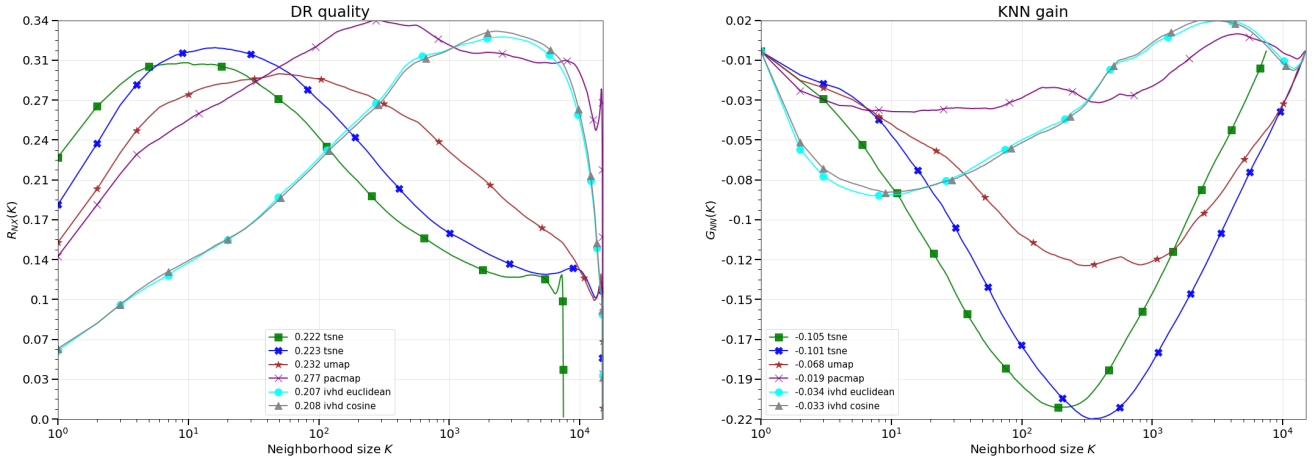

Figure 14: DR quality and KNN gain for RCV Reuters

Again, the metric indicates superior preservation of the local neighborhood by the t-SNE. Generally, IVHD behaves better and better when neighborhood size arises. KNN gain additionally enforces such a conclusion. Furthermore, PaCMAP results are consistent, and its embedding maintains a high level of mapping throughout the verified neighborhood size range. However, the key turns out to be global separation, where IVHD dominates.

4.3.5 Co-ranking matrix

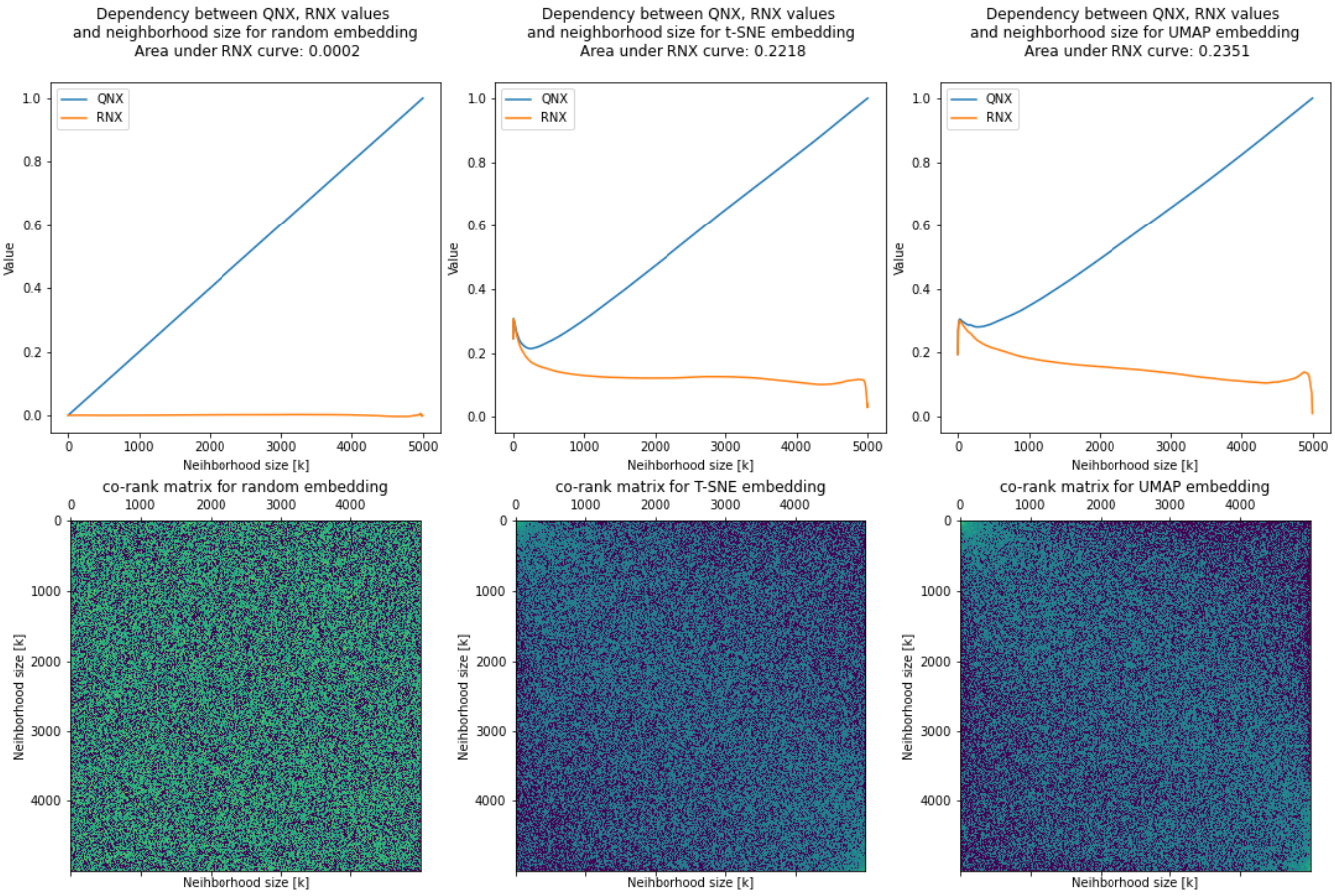

Figure 15: Co-ranking matrix-based metrics.
Results for RCV Reuters embeddings using random embedding, t-SNE and UMAP.

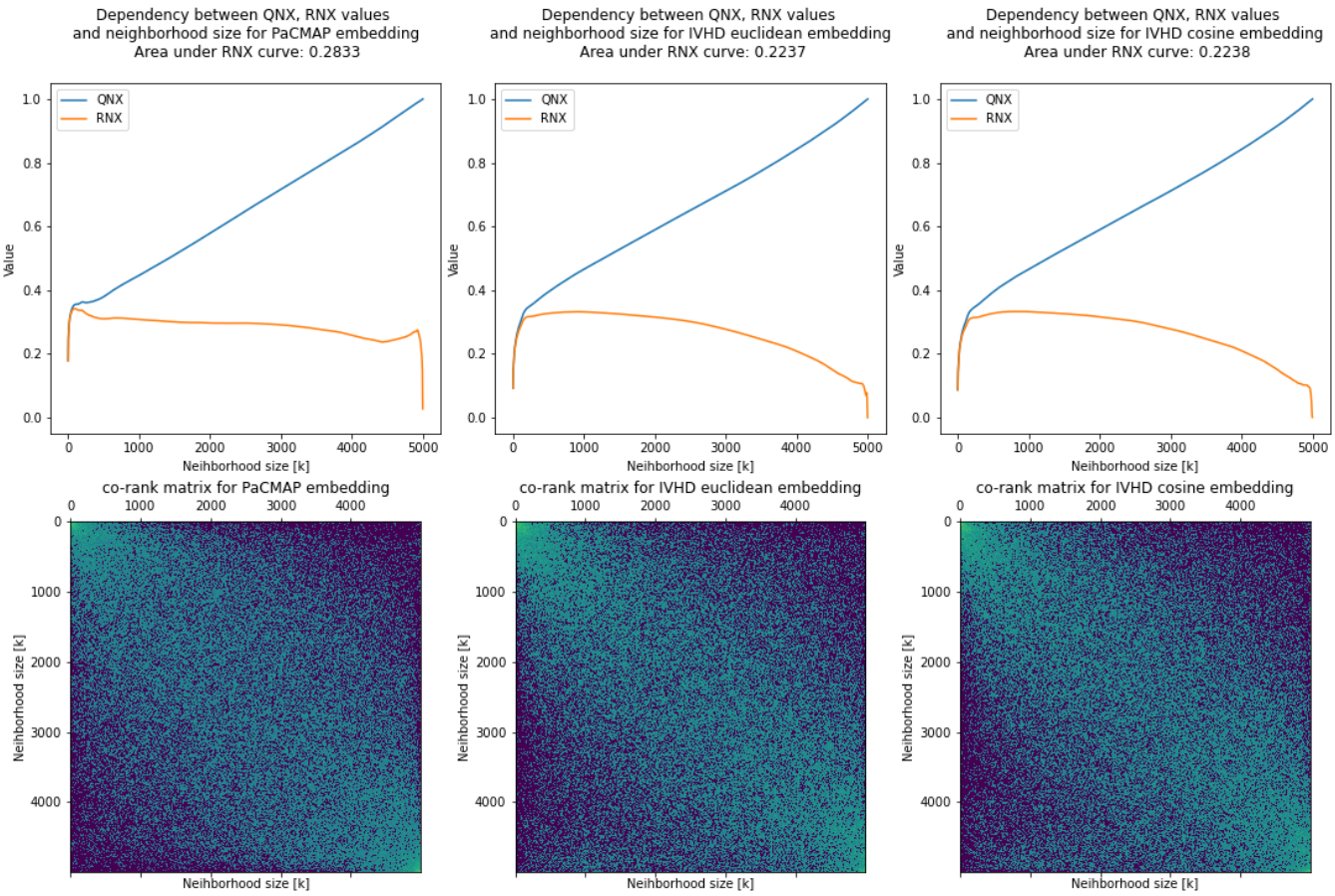

Figure 16: Co-ranking matrix-based metrics.
Results for RCV Reuters embeddings using PaCMAP and IVHD.

The metric characterizes PaCMAP as a method that has practically retained an almost constant level of quality and mapping for the sets of the analyzed neighbors. Its profit is visible for the increasing number of neighbors compared to the results of other methods.

4.3.6 Shepard diagrams

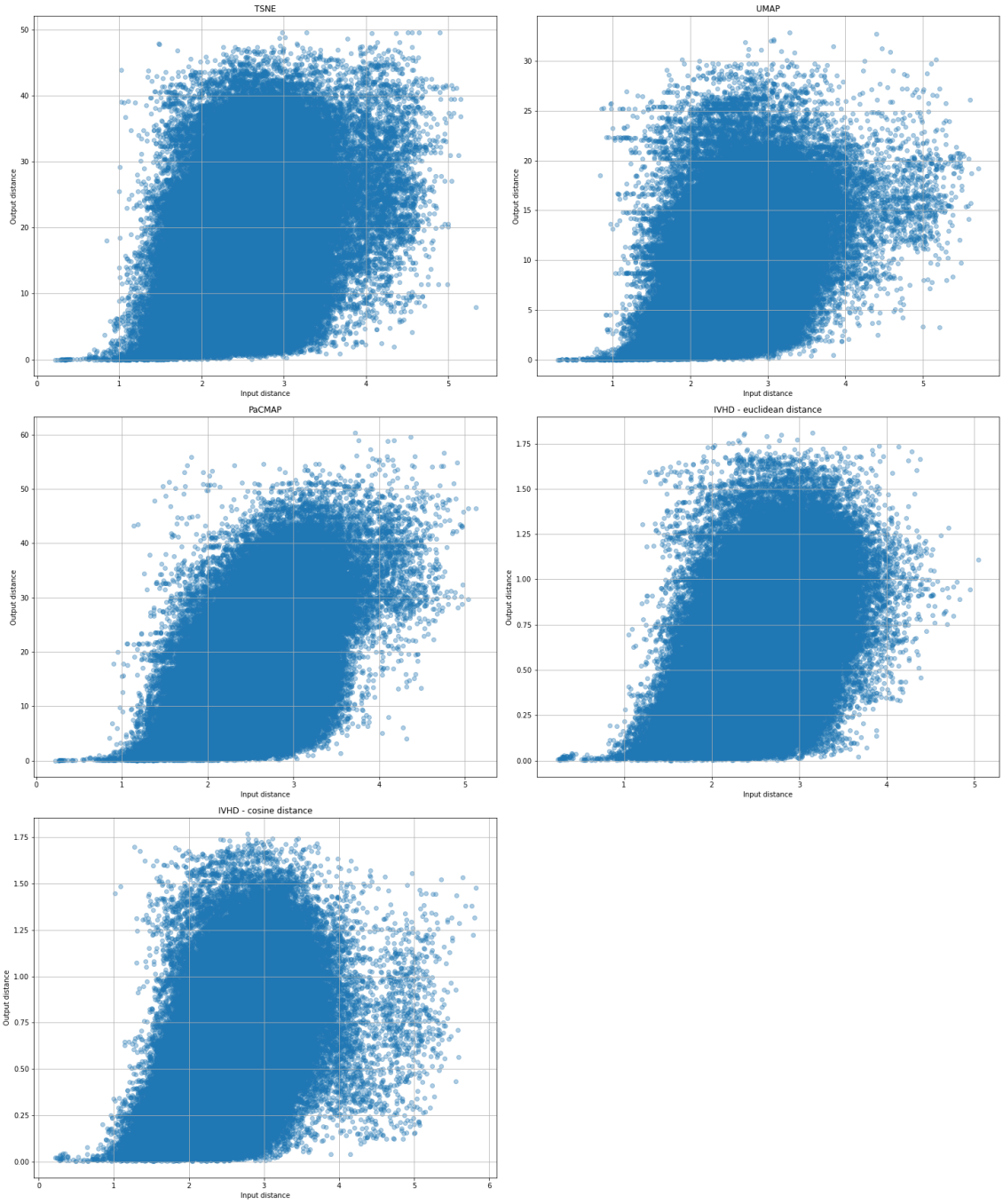

Figure 17: Shepard diagrams of RCV Reuters embeddings using t-SNE, UMAP, PaCMAP and IVHD.

The Shepard diagrams for PaCMAP and IVHD euclidean represent points centered around a straight line quite well. The greater the input distance, the more difficult it is to rescale it in a space with reduced dimensions. Therefore, for all methods (especially t-SNE, UMAP, and IVHD cosine) and high input distance values, a significant number of points are separated from the group.

4.3.7 Trustworthiness

Table 14: Trustworthiness with euclidean metric for pairwise distances.
Results for RCV Reuters embeddings using t-SNE, UMAP, PaCMAP and IVHD.

K	t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
5	0.907	0.915	0.935	0.862	0.864
15	0.847	0.883	0.919	0.854	0.855
50	0.763	0.828	0.897	0.845	0.845
100	0.717	0.793	0.876	0.834	0.834
150	0.692	0.772	0.860	0.830	0.830

Table 15: Trustworthiness with cosine metric for pairwise distances.
Results for RCV Reuters embeddings using t-SNE, UMAP, PaCMAP and IVHD.

K	t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
5	0.922	0.933	0.951	0.884	0.884
15	0.868	0.905	0.940	0.877	0.878
50	0.788	0.857	0.923	0.870	0.870
100	0.741	0.823	0.904	0.861	0.862
150	0.714	0.801	0.890	0.857	0.858

In the case of t-SNE and UMAP, we can observe the relatively large variability of the range of trustworthiness values for the examined number of neighbors compared to the trustworthiness of other methods - PaCMAP and IVHD. Therefore, PaCMAP and IVHD are definitely at the forefront in terms of local structure preservation. Additionally, the latter can be favored by the same quality of trustworthiness, regardless of the number of neighbors in a given range.

4.3.8 Spearman correlation-based metric

Table 16: Spearman correlation-based metric.
Results for RCV Reuters embeddings using t-SNE, UMAP, PaCMAP and IVHD.

t-SNE	UMAP	PaCMAP	IVHD euclidean	IVHD cosine
0.388	0.397	0.541	0.537	0.536

The metric based on the Spearman correlation confirms the conclusions drawn based on the trustworthiness metric. The PaCMAP and IVHD methods obtained the best results of similarity of the distances in both dimensionalities.

5 Summary and Conclusions

In this section, the results of the metrics obtained for the analyzed sets and the methods of data dimensionality reduction were summarized and compared. Additionally, this section was enriched with overall conclusions about high dimensional data visualization.

Table 17: Summary of metrics results for MNIST.

Metric	**t-SNE**	**UMAP**	**PaCMAP**	**IVHD euclidean**	**IVHD cosine**
Distance matrix-based metric		X			
Distance matrix-based metric with KMeans		X			
DR quality and KNN gain		X	X		
Co-ranking matrix	X				
Sheppard diagram		X	X		
Trustworthiness		X	X		
Spearman correlation-based metric		X	X		

Table 18: Summary of metrics results for 20 News Groups.

Metric	**t-SNE**	**UMAP**	**ISOMAP**	**IVHD euclidean**	**IVHD cosine**
Distance matrix-based metric				X	
Distance matrix-based metric with KMeans				X	
DR quality and KNN gain				X	
Co-ranking matrix	X				
Sheppard diagram				X	
Trustworthiness				X	
Spearman correlation-based metric	X	X			

Table 19: Summary of metrics results for RCV Reuters.

Metric	**t-SNE**	**UMAP**	**PaCMAP**	**IVHD euclidean**	**IVHD cosine**
Distance matrix-based metric				X	X
Distance matrix-based metric with KMeans				X	X
DR quality and KNN gain				X	X
Co-ranking matrix			X		
Sheppard diagram			X	X	X
Trustworthiness			X	X	X
Spearman correlation-based metric			X	X	X

According to metrics value, UMAP gave the best results for the MNIST dataset in the context of both local and global separation, PaCMAP also gives reasonable results. In this case, IVHD gave the worst effects, none of the examined metrics showed its superiority over other ones. In the case of 20 News Group, IVHD with Euclidean distance metric gives considerably better results than other methods, but in contrast, IVHD with cosine distance metric gives the worst embedding. ISOMAP also acts poorly in this case. For the RCV Reuters, the IVHD method gives the best results, and PaCMAP also gives decent results. None of the considered metrics showed domination of t-SNE or UMAP in the case of this dataset. Multiple metrics are available to assess the quality of dimensionality reduction techniques. The vast majority of them relate to local separation. Unfortunately, not all data sets allow for an intuitive assessment of the meaningfulness of the resulting clusters. MNIST is a perfect example of the one with very high interpretability. Therefore it is worth using diversified metrics that allow for more in-depth analysis and provide a strong foundation for formulating justified assessments. The use of various methods, the inspection of visualizations supported by the conclusions drawn from metrics interpretation, allows noticing the challenges facing this field. A visualization is undoubtedly a powerful tool where each DR technique can show its advantages depending on the highlighted goal to achieve. Finally, one of the most vital issues is the appropriate parameterization of methods. It undoubtedly affects the quality of dimensionality reduction and can radically change the obtained visualizations. Unfortunately, there is no one unique method that works best for each data set. Evaluation and verification should be made to match the one appropriate to the issue under consideration.

The source code and complete visualizations are available at the following location:
`https://github.com/Smendowski/data-embedding-and-visualization`

References

[1] Christopher J.C. Burges Yann LeCun Corinna Cortes. *The MNIST database of handwritten digits*. URL: http://yann.lecun.com/exdb/mnist/ (visited on 06/18/2022).

[2] scikit-learn developers. *20 News Groups*. URL: https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html (visited on 06/14/2022).

[3] UCI Machine Learning Repository maintainers. *RCV Reuters*. URL: https://archive.ics.uci.edu/ml/datasets/reuters+rcv1+rcv2+multilingual,+multiview+text+categorization+test+collection (visited on 06/14/2022).

[4] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[5] Geoffrey Hinton and Sam Roweis. "Stochastic Neighbor Embedding". In: *Advances in neural information processing systems* 15 (2003), pp. 833–840. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.7959&rep=rep1&type=pdf.

[6] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: (2018). URL: https://arxiv.org/pdf/1802.03426.pdf.

[7] Andy Coenen and Adam Pearce. *Understanding UMAP*. 2019. URL: https://pair-code.github.io/understanding-umap/ (visited on 06/12/2022).

[8] Leland McInnes. *UMAP documentation*. 2018. URL: https://umap-learn.readthedocs.io/en/latest/parameters.html (visited on 06/12/2022).

[9] scikit-learn developers. *Manifold learning*. URL: https://scikit-learn.org/stable/modules/manifold.html#isomap (visited on 06/14/2022).

[10] Kumar Pal Ashwini. *Dimension Reduction - IsoMap*. 2018. URL: https://blog.paperspace.com/dimension-reduction-with-isomap/#:~:text=Isomap%20is%20a%20non%2Dlinear,between%20all%20pairs%20of%20points. (visited on 06/14/2022).

[11] Yingfan Wang et al. "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization". In: *Journal of Machine Learning Research* 22.201 (2021), pp. 1–73. URL: http://jmlr.org/papers/v22/20-1061.html.

[12] Witold Dzwinel, Rafał Wcisło, and Stan Matwin. "2-D Embedding of Large and High-dimensional Data with Minimal Memory and Computational Time Requirements". In: *ArXiv* abs/1902.01108 (2019). URL: https://doi.org/10.48550/arXiv.1902.01108.

[13] Antonio Gracia et al. "A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality". In: *Information Sciences* 270 (2014), pp. 1–27. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2014.02.068.

[14] John A. Lee and Michel Verleysen. "Quality assessment of dimensionality reduction: Rank-based criteria". In: *Neurocomputing* 72.7 (2009). Advances in Machine Learning and Computational Intelligence, pp. 1431–1443. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2008.12.017.

[15] John A. Lee et al. "Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation". In: *Neurocomputing* 112 (2013). Advances in artificial neural networks, machine learning, and computational intelligence, pp. 92–108. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2012.12.036.

[16] John A. Lee, Diego H. Peluffo-Ordóñez, and Michel Verleysen. "Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure". In: *Neurocomputing* 169 (2015). Learning for Visual Semantic Understanding in Big Data ESANN 2014 Industrial Data Processing and Analysis, pp. 246–261. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2014.12.095.

[17] Jake Hoare. *Goodness of Fit in MDS and t-SNE with Shepard Diagrams.* URL: https://www.displayr.com/goodness-of-fit-in-mds-and-t-sne-with-shepard-diagrams/.

[18] *Trustworthiness metric in sckit-learn.* URL: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.trustworthiness.html.

[19] *Spearman Rank.* URL: https://www.sciencedirect.com/topics/engineering/spearman-rank (visited on 06/18/2022).

[20] *The project repository – Data embeddings and visualization.* URL: https://github.com/Smendowski/data-embedding-and-visualization/blob/main/%5B3%5D%20Data%20embeddings%20and%20visualizations.ipynb.

[21] *The project repository – MNIST analysis.* URL: https://github.com/Smendowski/data-embedding-and-visualization/blob/main/%5B4%5D%20MNIST%20embedding%20analysis.ipynb.

[22] *The project repository – 20NG analysis.* URL: https://github.com/Smendowski/data-embedding-and-visualization/blob/main/%5B5%5D%2020NG%20embedding%20analysis.ipynb.

[23] *The project repository – RCV Reuters analysis.* URL: https://github.com/Smendowski/data-embedding-and-visualization/blob/main/%5B6%5D%20RCV%20Reuters%20embedding%20analysis.ipynb.