



AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY

ENGINEERING AND TECHNOLOGY
INFORMATION AND COMMUNICATION TECHNOLOGY

DOCTORAL THESIS

IN SEARCH OF THE MOST EFFICIENT AND MEMORY-SAVING
VISUALIZATION OF HIGH DIMENSIONAL DATA

Autor:

Bartosz Minch

Promotor:

Witold Dzwinel, Ph.D, Professor

Promotor pomocniczy:

Dariusz Jamróz, Ph.D

Praca wykonana:

AGH University of Science and Technology,
Institute of Computer Science

Cracow, 2023



AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

NAUKI INŻYNIERYJNO-TECHNICZNE
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA

ROZPRAWA DOKTORSKA

**W POSZUKIWANIU NAJBARDZIEJ WYDAJNEJ I OSZCZĘDZAJĄCEJ
PAMIĘĆ WIZUALIZACJI DANYCH WIELOWYMIAROWYCH**

Autor:

mgr inż. Bartosz Minch

Promotor:

prof. dr hab. inż. Witold Dzwinel

Promotor pomocniczy:

dr Dariusz Jamróz

Praca wykonana:

Akademia Górniczo-Hutnicza,
Instytut Informatyki

Kraków, 2023

ABSTRACT

Interactive visual exploration of large, high-dimensional datasets plays a very important role in various fields of science, which requires aggregated information about mutual relationships between numerous objects. It enables not only to recognize their important structural features and forms, such as clusters of vertices and their connectivity patterns, but also to assess their mutual relationships in terms of position, distance, shape, and connection density. The structural properties of these large datasets can be scrutinized throughout their interactive visualization. We argue that the visualization of very high-dimensional data is well approximated by the two-dimensional ($2D$) problem of embedding undirected kNN -graphs. In the advent of the big data era, the size of complex networks (datasets) $G(V, E)$ ($|V|=M \sim 10^{6+}$) represents a great challenge for today's computer systems and still requires more efficient $ND \rightarrow 2D$ dimensionality reduction (DR) algorithms. The existing DR methods, which involve more computational and memory complexities than $O(M)$, are too slow for interactive manipulation on large networks that involve millions of vertices. We show that high-quality embeddings can be produced with minimal time&memory complexity. Very efficient IVHD (interactive visualization of high-dimensional data) and IVHD-CUDA algorithms are presented and then compared to the state-of-the-art DR methods (both CPU and GPU versions): t-SNE, UMAP, TriMAP, PaCMAP, BH-SNE-CUDA, and AtSNE-CUDA. We show that the memory and time requirements for IVHD are radically lower than those for the baseline codes. For example, IVHD-CUDA is almost 30 times faster in embedding (without the kNN graph generation procedure, which is the same for all methods) of one of the largest datasets used, YAHOO ($M=1.4 \cdot 10^6$), than AtSNE-CUDA. We conclude that at the expense of a minor deterioration of embedding quality, compared to baseline algorithms, IVHD well preserves the main structural properties of ND data in $2D$ for a significantly lower computational budget. We also present a meta-algorithm that enables using any unsupervised DR method in supervised fashion and as a result allows for flexible control of global and local properties of the embedding. Thus, our methods can be a good candidate for an interactive visualization of truly big data ($M=10^{8+}$) and can be further used to inspect and interpret relationships between alternative representations of observations learned by artificial neural networks (ANN). Additionally, we have provided a framework for testing the trade-off between preservation of global structure and local structure of DR.

STRESZCZENIE

Interaktywna, wizualna eksploracja dużych, wielowymiarowych zbiorów danych odgrywa bardzo ważną rolę w różnych dziedzinach nauki, która wymaga zagregowanej informacji o wzajemnych relacjach między wieloma obiektami. Umożliwia ona nie tylko rozpoznanie ich istotnych cech i form strukturalnych, takich jak skupiska wierzchołków i ich wzorce połączeń, ale także ocenę ich wzajemnych relacji w zakresie położenia, odległości, kształtu i gęstości połączeń. Twierdzimy, że wizualizacja wielowymiarowych danych (ND) jest dobrze przybliżana przez problem dwuwymiarowego ($2D$) osadzania nieukierunkowanych grafów najbliższych sąsiadów (ang. k NN graphs). W dzisiejszych czasach, rozmiar złożonych sieci (zbiorów danych) $G(V, E)$ ($|V|=M \sim 10^{6+}$) stanowi duże wyzwanie dla dzisiejszych systemów komputerowych i wciąż wymaga bardziej wydajnych algorytmów osadzania danych wielowymiarowych. Istniejące metody redukcji wymiarowości danych, które wymagają większej złożoności obliczeniowej i pamięciowej niż $O(M)$, są zbyt wolne do interaktywnej manipulacji na dużych sieciach obejmujących miliony wierzchołków. Pokazujemy, że osadzenia wysokiej jakości mogą być produkowane przy minimalnej złożoności czasowej i pamięciowej. Przedstawiamy bardzo wydajne algorytmy IVHD oraz IVHD-CUDA, a następnie porównujemy je z najnowszymi i najpopularniejszymi metodami redukcji wymiarowości (zarówno w wersji dla CPU, jak i GPU): t-SNE, UMAP, TriMAP, PaCMAP, BH-SNE-CUDA oraz AtSNE-CUDA. Pokazujemy, że wymagania pamięciowe i czasowe dla IVHD są radykalnie niższe niż dla kodów bazowych. Na przykład, IVHD-CUDA jest prawie 30 razy szybsza w osadzaniu (bez procedury generowania grafu najbliższych sąsiadów, która jest taka sama dla wszystkich metod) jednego z największych użytych zbiorów danych, YAHOO ($M=1.4 \cdot 10^6$), niż AtSNE-CUDA. Stwierdzamy, że kosztem niewielkiego pogorszenia jakości osadzania, w porównaniu do algorytmów bazowych, IVHD dobrze zachowuje główne własności strukturalne danych ND w $2D$ przy znacznie niższym budżecie czasowym. Przedstawiamy również meta-algorytm, który umożliwia wykorzystanie dowolnej nienadzorowanej metody osadzania danych w sposób nadzorowany i w rezultacie pozwala na elastyczną kontrolę globalnych i lokalnych własności osadzenia. Dzięki temu, nasze metody mogą być dobrym kandydatem do interaktywnej wizualizacji naprawdę dużych zbiorów danych ($M=10^{8+}$) i mogą być dalej wykorzystywane do inspekcji i interpretacji zależności pomiędzy alternatywnymi reprezentacjami obserwacji wyuczonymi przez sztuczne sieci neuronowe (ANN).

ACKNOWLEDGEMENTS

I express my deepest gratitude to Professor Witold Dzwiniel for guiding and encouraging me during my research efforts. I also thank Dr. Dariusz Jamróz for helpful discussions and invaluable advice that greatly improved this thesis.

I dedicate this work to my parents, my family and my friends.

The research for this thesis was supported by:

- Funds allocated to the AGH University of Science and Technology by the Polish Ministry of Science and Higher Education.
- Infrastructure and computing resources made available by ACK Cyfronet PL-Grid.

TABLE OF CONTENTS

Glossary of symbols	10
1. Introduction	11
1.1. Motivation	11
1.2. The thesis and goals	13
1.3. The structure of dissertation	14
2. Dimensionality reduction	15
2.1. Structured and unstructured data	16
2.2. Manifold learning	18
2.3. Spectral dimensionality reduction	19
2.3.1. Unsupervised methods	19
2.3.2. Supervised methods	25
2.4. Probabilistic dimensionality reduction	27
2.4.1. Uniform Manifold Approximation and Projection (UMAP) as a general approach to data embedding and visualization	27
2.4.2. Stochastic Neighborhood Embedding heuristics	32
2.4.3. State-of-the-art supervised and unsupervised improvements and simplifications. UMAP and t-SNE.	36
2.5. Neural network based dimensionality reduction	43
3. Towards optimal NN graph visualization	45
3.1. LargeVis	46
3.1.1. k NN graph construction	46
3.1.2. Probabilistic model for graph visualization	47
3.1.3. Differences between UMAP, SNE and LargeVis methods	48
3.2. t-SNE with Euclidean and binary distances	50
3.3. Interactive visualization of high-dimensional data	52
3.3.1. Improving classical MDS	54
3.3.2. Optimization methods	58
3.3.3. Force directed	59
3.3.4. Binary and Euclidean distances comparison	61
3.4. Improvements in IVHD algorithm	62

3.4.1. L1 norm in late stages of embedding.....	63
3.4.2. Reverse neighbors mechanism	64
4. Experimental setup	67
4.1. Quality assessment of dimensionality reduction.....	68
4.2. Datasets	72
4.3. Baseline methods.....	73
4.4. Experiments.....	77
4.4.1. Hardware	77
4.4.2. Synthetic datasets	77
4.4.3. Mid-scale baseline datasets	82
5. GPU-Embedding of kNN-Graph Representing Large and High-Dimensional Data	86
5.1. k NN graph generation	86
5.2. BH-SNE-CUDA, At-SNE, IVHD-CUDA	87
6. Meta-platform for supervised visualization.....	93
7. Application of HDD visualizations in ANNs inspection and interpretation .	102
7.1. Experimental protocol.....	102
7.1.1. Exploring the relationships between alternative representations of observations learned by ANN.....	104
7.1.2. Exploring the relationships between artificial neurons.....	108
8. Conclusions.....	110
A. Datasets.....	113
A.1. Datasets	114
B. Visualizations, timings and methods parametrization	119

GLOSSARY OF SYMBOLS

x	a scalar
\mathbf{x}	a vector
\mathbf{X}	a matrix (dataset)
\mathbb{R}	the set of real numbers
\mathbb{R}^N	high-dimensional space ($N \gg n$)
\mathbb{R}^n	low-dimensional space ($n \ll N$)
$ \mathbb{X} $	number of elements in \mathbb{X}
\mathbf{X}^T	transpose of matrix \mathbf{X}
x_i	i -th element of vector \mathbf{x}
$\mathbf{X}_{i\cdot}$	i -th row of matrix \mathbf{X}
$\mathbf{X}_{\cdot j}$	j -th row of matrix \mathbf{X}
x_{ij}	element (i, j) of matrix \mathbf{X}
$\ \mathbf{x}\ _1$	ℓ_1 norm of vector \mathbf{x}
$\ \mathbf{x}\ _2$	ℓ_2 norm of vector \mathbf{x}
$\frac{\partial x}{\partial y}$	partial derivative of y with respect to x
$\nabla_{x,y}$	gradient of y with respect to x
$G(V, E)$	graph with sets of \mathbb{V} vertices and \mathbb{E} edges
$\mathcal{N}(m, s^2)$	a Gaussian distribution with mean m and variance s^2
\mathcal{M}	Riemannian manifold

CHAPTER 1

INTRODUCTION

In the era of big data, most of the data generated every day is represented directly as: 1) *structured data* - ND vectors, or as 2) *unstructured data* - embedded in ND Euclidean space (e.g. pictures, graphs, text). It is common for popular deep learning approaches to use data augmentation to satisfy the need to train a huge number of parameters without overfitting, the growing amount of data requires some key data reduction methods for different motivations. In general, dimensionality reduction (DR) is useful for: 1) better storage efficiency, 2) shorter computation time, 3) creating better data representation, 4) removing outliers, and 5) better recognition performance. Data reduction approaches are divided into two categories (Fig. 1.1), that is, dimensionality reduction and numerosity reduction, which reduce the dimensionality and sample size of the data, respectively.

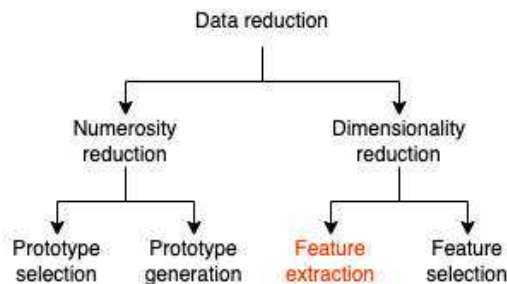


Figure 1.1: Taxonomy of data reduction. This dissertation tackles the red part.

1.1. Motivation

In this work, we investigate the ways in which dimensionality reduction (DR) methods can benefit from a common framework that would allow us to determine the reliability of low-dimensional mapping (embedding). We focus on visualization using the scatter plot technique. This type of data visualization is useful in illustrating the relationships that exist between variables and can be used to identify trends or correlations

in the data. Other types of data visualization (such as charts, heat maps, graphs) have been omitted from our considerations. Furthermore, to visualize the structured data, we only need a distance matrix between consecutive objects. For unstructured data, we first need to create a vector representation in Euclidean space, which allows for: 1) not using complicated metrics to calculate distances, and 2) such data can be further used to train a neural network. Additionally, it would allow one to trace the changes in embedding quality for modern algorithms used in the visualization of high-dimensional information, which is obtained by reducing the amount of information each individual object (that is being embedded) has about the entire data set and reducing the computational complexity of modern DR algorithms. In particular, we focus on two main areas where, as we have found, DR methodologies have the biggest procedural holes.

First, we attempt to establish a common framework that would allow for a proper quality evaluation of different embeddings. We show how this can be useful in the problem of interpreting data that are represented by sparse, unstructured, high-dimensional feature vectors. This type of data often arises in fields such as social networks, web indexing, gene sequencing, and biomedical analysis. Interactive visualization allows for: 1) instant verification of a number of hypotheses, 2) precise matching of data mining tools to the properties of the data investigated, 3) adapting the optimal parameters to machine learning algorithms, and 4) selecting the best data representation.

Second, we investigate how to considerably decrease the time&memory complexity of visualization of high-dimensional data¹ with minimal decrease of the embedding quality. This would enable one to analyze visually radically larger datasets than those of the state-of-the-art visualization algorithms. The 2D data embedding would allow for both insight into the large data structure and its interactive exploration through direct manipulation of all or part of the data set. In this way, the shapes and mutual locations of single data and classes can be observed, irrelevant data samples can be removed, and outliers can be identified. The multiscale structure can be explored visually by changing data embedding strategies and visualization modes (e.g., the type of the loss function) and zooming in and out selected fragments of 2D (3D) data mapping. We also investigated a centroid meta-procedure, which utilizes state-of-the-art clustering algorithms [115], allowing for even better parameterization of the final visualization in terms of its local and global properties.

Finally, using the research described in the first two paragraphs, our objective is to solve another well-recognized challenge, namely to interpret a model prediction when training and analyzing deep learning models [153]. Although these techniques are useful, few works have been published [110, 125, 150] to explain how model predictions are developed during the training process. Obtaining training information, which evolves over time, can be useful, but it is difficult to abstract the evolving part of the underlying model. The following subsidiary questions arise: 1) How does the training process incrementally improve the model? 2) How does the model gradually make trade-offs to fit some samples while sacrificing others? 3) How does the model handle matching and learning difficult samples? To answer these questions, we visualize

¹While the term "high-dimensional" is sometimes used to refer to data described by at least four features, here we consider a feature vector to be high-dimensional when its dimensionality is on the order of at least $10^2 - 10^4$

the relationship between learned representations of observations and the relationship between artificial neurons. We show how visualization can provide highly valuable feedback for network designers.

1.2. The thesis and goals

The overarching purpose of this dissertation is to investigate an optimal DR method that could effectively and interactively visualize truly large and high-dimensional datasets, when very strict time&memory performance regimes are implemented. To this end, we follow these crucial assumptions.

1. We focus on embedding high-dimensional data into low-dimensional spaces. For this purpose, each object is represented as a point in order to visualize the structure of the dataset, choose the appropriate representation of the data, choose the appropriate metric to represent the manifold, and match meta-parameters of machine learning algorithms.
2. The method is expected to enable direct analysis of data (through interactive intervention in the data set, selection of loss functions, etc.).
3. We are not interested in preprocessing the data beforehand, i.e., we work on the final vector representation of the data and the defined distance metric.

In the dissertation, we also refer to supervised embedding methods and present the application of the proposed method to the analysis and interpretation of the performance of neural networks.

The thesis of the dissertation is as follows.

Computational complexity of modern DR algorithms for embedding N high-dimensional data vectors into low-dimensional spaces might be reduced to $O(kN)$. It can be achieved by: a) drastically reducing the neighbor information of each object, b) introducing the binary distance between objects, and c) using an efficient loss function minimization method, which does not drastically degrade the embedding quality.

The main contributions of this dissertation are as follows:

- An extensive overview and analysis of existing DR methods and the challenges they face.
- Highlight the advantages and disadvantages of the simplest method (IVHD) in a variety of possible contexts compared to far more sophisticated approaches.
- GPU implementation, which allowed one to visualize large high-dimensional datasets ($N \sim 10^{6+}$) in a reasonable amount of time.
- Introducing significant improvements to IVHD that improve the global and local properties of visualization.
- Implementation of the DR library [92], which allows us to efficiently measure the reliability of low-dimensional (embedded) data representations.

- Comparison of methods for complex low-dimensional data (in addition to medium and large datasets).
- The proposition of a centroid-based meta-procedure that allows any unsupervised method to be used in supervised fashion.
- Application of the IVHD method for inspection and interpretation of inter-epoch and inter-layer DNN behavior.

All the research described above has made it possible to develop an optimal method that can be used to visualize very large datasets and investigate neural networks in an interactive way.

1.3. The structure of dissertation

The dissertation is organized as follows.

Chapter 2 introduces the state-of-the-art division of DR methods and describes the classical ones. Then, we focus on two currently dominant DR methods: UMAP and t-SNE. Both are currently the fundamental algorithms that are employed as the basis for many new ideas in high-dimensional data visualization.

In Chapter 3 we focus on reducing the dimensionality reduction problem to a graph visualization. We introduce the IVHD method, which is a modification of the classical MDS method and verify how binary and Euclidean distances affect the quality of DR algorithms. Additionally, optimization methods are presented and compared to force-directed method implemented in IVHD. In the last subsection, we present the improvements that have been made to the IVHD method that have improved its local and global properties.

Chapter 4 introduces the common framework for conducting DR experiments. We define the quality measures that are used for all of the methods. Furthermore, we describe the baseline methods used for IVHD comparisons. The experiments are divided into two subsections, depending on the size of analyzed datasets.

In Chapter 5, IVHD-CUDA is introduced. It is a CUDA implementation of the IVHD method. The chapter includes benchmarks for the IVHD-CUDA and SOTA methods implemented in a CUDA environment (AtSNE and BH-SNE-CUDA).

Chapter 6 presents meta-platform for supervised visualization of high-dimensional data. It allows any unsupervised DR method to be used in supervised fashion.

In Chapter 7 we evaluate our research by applying it to inspection and interpretation of Artificial Neuron Networks (ANNs). We investigate how to visualize the relationships between learned representations and between neurons in networks.

Finally, in Chapter 8 we conclude the dissertation and discuss further directions for research.

CHAPTER 2

DIMENSIONALITY REDUCTION

In this chapter, we introduce several important baseline DR methods and explain the technical background. It also explains the open problems in data reduction which are tackled in this thesis. In Section 2.1 we focus on structured and unstructured data that can be embedded with DR methods. In Section 2.2 we introduce the reasons and related work for dimensionality reduction. We present baseline methods for specific subcategories of dimensionality reduction types. Sections 2.4.1 and 2.4.2 introduce stochastic neighbor embedding heuristics and uniform manifold approximation and projection (UMAP) [90], which are the state-of-the-art methods for data visualization.

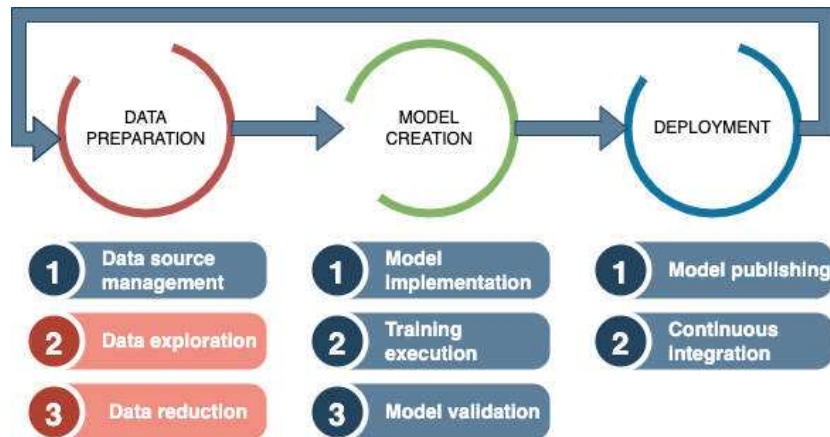


Figure 2.1: A taxonomy of AI systems infrastructure. This dissertation focuses on two key stages of the data preparation step: data exploration and data reduction, which are shown as the red parts of the diagram. Data reduction taxonomy was described in the first chapter of this thesis.

Dimensionality reduction methods can be divided into three categories: 1) spectral methods, 2) probabilistic methods and 3) methods based on neural networks [49], which have a geometric, probabilistic, and information-theoretic viewpoint on dimensionality reduction. These categories are based on the generalized eigenvalue problem, latent variables, and neural networks, respectively. In feature extraction for dimension-

ality reduction, required for interactive visualization of high-dimensional data, which this thesis focuses on, a new set of features is found for better representation or data discrimination.

Definition 1 Let $\mathbf{X} = \{x_i\}_{i=1}^M$ and $\mathbf{Y} = \{y_i\}_{i=1}^M$ be, respectively, sets of M instances in a high-dimensional space ($\mathbf{X} \subset \mathbb{R}^N$) and corresponding embeddings in a low-dimensional space ($\mathbf{Y} \subset \mathbb{R}^n$), where $N \gg n$.

Definition 2 In supervised learning, there are $\{(x_i, l_i)\}_{i=1}^M$ meaning that every instance x_i has a corresponding label $l_i \in \mathbb{R}^\ell$, where ℓ is the dimensionality of the label. We can then form the label matrix $L = [l_1, \dots, l_i] \in \mathbb{R}^{M \times \ell}$. In classification, each instance belongs to one of $|C|$ where C is the set that includes labels from classes. The cardinality of the set of instances in class c is denoted by n_c .

Definition 3 Dimensionality reduction (DR) is defined as a mapping:

$$\mathbf{B}: \mathbf{X} \rightarrow \mathbf{Y}$$

\mathbf{B} can be perceived as a lossy compression of the data. It is carried out by minimizing a loss function $E(\|\mathbf{X} - \mathbf{Y}\|)$, where $\|\cdot\|$ is a measure of the topological dissimilarity between \mathbf{X} and \mathbf{Y} . Due to the high complexity of the low-dimensional manifold, immersed in the ND feature space and occupied by data samples x_i , perfect embedding of \mathbf{X} in the nD space is possible only for trivial cases.

Dimensionality reduction (DR) tools in data visualization is a double-edged sword in understanding the geometric and neighborhood structures of data sets. Having the ability to efficiently visualize data sets can provide an understanding of the cluster structure and provide an intuition of distributional characteristics. However, it is well-known that DR results can be misleading, showing cluster structures that are simply not present in the original data, or showing that the observations are far apart in the projection space when they are close together in the original space. Thus, if we were to run several DR algorithms, we could get different results. It is not clear how we would determine which of these results, if any, give a reliable representation of the original data distribution.

2.1. Structured and unstructured data

From the point of view of interactive data visualization, the following properties of the datasets are important because they determine the calculation of the distance matrix, which is later used as input to the dimensionality reduction algorithms [94].

Data dimensionality. A popular and intuitive way to represent a given data set is to use a vector space model [121]. In a vector space model, observations are represented by a matrix $M \times N$ called a design matrix, in which each of the rows M corresponds to an observation that is described by attributes N (also called characteristics or variables). Interpreting the attributes depends, of course, on the nature of the data set. In the case of an image collection, an observation refers to an image that is defined by a list of

pixel intensities or higher-order features, while text, for example, is often represented as a multi-collection of its words, the so-called bag-of-words (*BOW*) representation. Regardless of the interpretation of the features, their N number or the dimensionality of the data play an important role in determining the applicability of machine learning, data mining and data embedding methods. High-dimensionality is an ubiquitous property of modern datasets. Data with hundreds or even millions of features appear in various application domains such as 2D/3D digital image processing, bioinformatics, e-commerce, web crawling, social networks, mass spectrometry, text analytics, and speech processing.

Data sparsity. It is a common property of many high-dimensional data sets. It is defined as the number of elements with zero values in a matrix $M \times N$ divided by the total number of elements MN . However, when working with highly sparse datasets, a more convenient term is data density, which is equal to one minus sparsity [57]. The zeros in the computational matrix may simply represent missing measurements, also denoted as null values or "NaN" values.

Data structure. Structured data are data that have been predefined and formatted to a set structure before being placed in data storage, which is often called schema-on-write. The best example of structured data is the relational database: the data have been formatted into precisely defined fields, such as credit card numbers or addresses. Unstructured data are data stored in its native format and not processed until it is used, which is known as a schema-on-read. They come in a myriad of file formats, including emails, social media posts, presentations, chats, IoT (Internet of Things) sensor data, audio, video, and satellite imagery. The big data industry is growing rapidly, and so are the data it produces. The biggest challenge is to effectively harness the latent power of unstructured data, especially with the speed of information creation continuing to accelerate.

Distance matrix. It is a non-negative, square matrix with elements $-\frac{1}{2}d_{ij}^2$ corresponding to estimates of some pairwise distance between the sequences in a set. The distance matrix \mathbf{D} is Euclidean, when the $\frac{1}{2}N(N-1)$ quantities d_{ij} can be generated as distances between a set of M points, $\mathbf{X}(M \times N)$, in an Euclidean space of dimension N . The dimensionality of \mathbf{D} is defined as the least value of $p = \text{rank}(\mathbf{X})$ of any generating \mathbf{X} . The most important properties of the Euclidean distance matrix follow from the fact that the Euclidean distance is a metric. Thus, the Euclidean matrix \mathbf{D} has the following properties:

- all elements on the diagonal $d_{ii} = 0$, hence $\text{trace}(\mathbf{D}) = 0$,
- \mathbf{D} is symmetric ($d_{ij} = d_{ji}$),
- $\sqrt{d_{ij}} \leq \sqrt{d_{ik}} + \sqrt{d_{kj}}$ (by the triangle inequality),
- $d_{ij} \geq 0$,
- in dimension k , an Euclidean distance matrix has rank less than or equal to $k + 2$.

Additionally, it is worth mentioning that distances do not have to meet the formal definition of distance. Dimensionality reduction methods often use the so-called proximity matrix, which measures similarity or dissimilarity between data.

2.2. Manifold learning

Manifold learning methods [117] play a prominent role in non-linear dimensionality reduction and other tasks involving high-dimensional data sets with low intrinsic dimensionality. Many of these methods are graph-based: they associate a vertex with each data point and a weighted edge with each pair.

A manifold is a generalization of curves and surfaces to higher dimensions. It is locally Euclidean (Def. 5) in that every point has a neighborhood, called a chart, homeomorphic (Def. 4) to an open subset of \mathbb{R}^N . The coordinates on a chart allow one to perform computations as if in Euclidean space, so many concepts from \mathbb{R}^N , such as differentiability, point derivations, tangent spaces, and differential forms, carry over to a manifold [143]. A good example of a manifold is the Earth. Locally, at each point on the surface of the Earth, we have a 3-D coordinate system: two for location and the last one for altitude. Globally, it is a 2-D sphere in a 3-D space.

Definition 4 *A continuous map $F : X \rightarrow Y$ is a homeomorphism if it is bijective and its inverse F^{-1} is also continuous. If two topological spaces admit a homeomorphism between them, we say that they are homeomorphic: they are essentially the same topological space.*

Definition 5 *A topological space \mathcal{M} is locally Euclidean of dimension N if every point p in \mathcal{M} has a neighborhood \mathcal{U} such that there is a homeomorphism ϕ from \mathcal{M} onto an open subset of \mathbb{R}^N . We call the pair $(\mathcal{U}, \phi : \mathcal{U} \rightarrow \mathbb{R}^N)$ a chart, \mathcal{U} a coordinate neighborhood or a coordinate open set, and ϕ a coordinate map or a coordinate system in \mathcal{U} . We say that a chart (\mathcal{U}, ϕ) is centered at $p \in \mathcal{U}$ if $\phi(p) = 0$.*

Geometric and topological relationships are fundamental to essentially every data analysis and machine learning task, since we use geometry to identify similarities and distinctive characteristics in the data. In classification, for example, data points that are similar (close to each other) are assigned to the same class, and data points that are significantly different (i.e., far apart in one or many of the feature dimensions) are assigned to different labels. We usually consider data as a finite set of M -dimensional vectors \mathbf{X} , sometimes called a point cloud. However, geometric and topological structures, such as metric spaces and manifolds, are continuous, not discrete. To discover any geometric or topological properties of the data, we fit a continuous shape to the data, and we must make certain assumptions about the underlying mathematical space we are working in. It is often simply assumed that our data lie in the standard Euclidean space with the typical Euclidean metric, and the data is analyzed by referencing a global, external coordinate system. However, many interesting and important structures that arise are actually non-Euclidean, and by fitting a continuous shape (such as a manifold) to the data, we can translate our data analysis task from an external, global coordinate system (possibly having very high dimensionality) into the intrinsic coordinate system defined by the assumed manifold structure itself. Manifolds offer a powerful framework for DR and there are several motivations for manifold learning and interactive data visualization as a result.

- According to the manifold hypothesis [36], the data usually exist in a subspace or submanifold (unless it is a random noise). Therefore, the entire N -dimensional space is not required and a large part of it is redundant information. We can find the best n -dimensional subspace to represent the data with the smallest possible reconstruction error.
- Manifold learning methods can provide more efficient feature extraction later used for classification, representation, clustering, or revealing patterns in data.

2.3. Spectral dimensionality reduction

Spectral DR methods come down to eigenvalue decomposition and the generalized eigenvalue problem [48]. They use a geometric approach and unfold the manifold into a lower-dimensional subspace. The most well-known unsupervised methods with the spectral approach are the following: PCA [60], classical MDS [39], Sammon mapping [122], LLE [116], Isomap [134].

2.3.1. Unsupervised methods

Unsupervised visualization refers to visualization based only on input data without corresponding output variables, or labels. The goal is for the system to generate its own model of the underlying structure or distribution.

Principal Component Analysis (PCA)

PCA [60], one of the most widely used tools in data analysis and data mining, is also one of the most popular linear-dimensionality reduction methods.

Assume that we have a matrix of centered data observations:

$$\mathbf{X} = [x_1 - \mu, \dots, x_N - \mu] \quad (2.1)$$

where μ denotes the mean vector. $\mathbf{X} \in \mathbb{R}^{N \times M}$, where N is the number of dimensions (dimensionality) and M is the number of observations (numerosity). Their covariance matrix is given by:

$$\mathbf{S}_t = \frac{1}{M} \sum_{i=1}^M (x_i - \mu)(x_i - \mu)^T = \frac{1}{M} \mathbf{X} \mathbf{X}^T \quad (2.2)$$

In Principal Component Analysis (PCA), we aim to maximize the variance of each dimension by:

$$\begin{aligned} \mathbf{W}_0 = \arg \max_{\mathbf{W}} \quad & tr(\mathbf{W}^T \mathbf{S}_t \mathbf{W}) \\ \text{subject to} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I} \end{aligned} \quad (2.3)$$

The solution of Eq. 3 can be derived by solving:

$$\mathbf{S}_t \mathbf{W} = \mathbf{W} \mathbf{\Lambda} \quad (2.4)$$

Thus, we need to perform an eigenanalysis on \mathbf{S}_t . If we want to keep d principal components, the computational cost of the above operation is $O(dN^2)$.

Lemma 1 *Let us assume that $\mathbf{B} = \mathbf{X}\mathbf{X}^T$ and $\mathbf{C} = \mathbf{X}^T\mathbf{X}$. It can be proven that \mathbf{B} and \mathbf{C} have the same positive eigenvalues Λ and, assuming that $M < N$, then the eigenvectors \mathbf{U} of \mathbf{B} and the eigenvectors \mathbf{V} of \mathbf{C} are related as $\mathbf{U} = \mathbf{X}\mathbf{V}\Lambda^{-\frac{1}{2}}$.*

Using Lemma 1 we can compute the eigenvectors \mathbf{U} of \mathbf{S}_t in $O(N^3)$. The eigenanalysis of $\mathbf{X}^T\mathbf{X}$ is denoted by:

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\Lambda\mathbf{V}^T \quad (2.5)$$

Given that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^T \neq \mathbf{I}$ the covariance matrix of \mathbf{Y} is:

$$\mathbf{Y}\mathbf{Y}^T = \mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U} = \Lambda \quad (2.6)$$

The final solution of Eq. 2.3 is given as the projection matrix:

$$\mathbf{W} = \mathbf{U}\Lambda^{-\frac{1}{2}} \quad (2.7)$$

Algorithm 1: Principal Component Analysis scheme.

Input: data matrix \mathbf{X} .

procedure PCA:

1. Compute dot product matrix: $\mathbf{X}^T\mathbf{X}$.
 2. Eigenanalysis: $\mathbf{X}^T\mathbf{X} = \mathbf{V}\Lambda\mathbf{V}^T$.
 3. Compute eigenvectors: $\mathbf{U} = \mathbf{X}\mathbf{V}\Lambda^{-\frac{1}{2}}$.
 4. Keep specific numbers of first components: $\mathbf{U}_d = [u_1, \dots, u_d]$.
 5. Compute d features: $\mathbf{Y} = \mathbf{U}_d^T\mathbf{X}$.
-

PCA has been applied to a variety of applied problems, such as image processing, statistics, text mining, and facial recognition. However, there are obvious drawbacks to the method, clearly, one being that the centralized data are required to lie on a linear subspace or something very close to it. This is a very strong assumption, since it assumes that the variables of the data are correlated in a linear fashion, which is not true in many applications. The most common variations of PCA are the following:

- Kernel PCA [44], which increases the dimensionality of the data by mapping them to the feature space with a higher dimensionality in hopes that they fall on a linear manifold in that feature space.
- Supervised PCA [44], which utilizes information concerning classes.

Furthermore, it is worth mentioning SiMultaneous PCA (SMPCA) [82], Progressive PCA (PGPCA) [82], Successive PCA (SCPCA) [82] and PRioritized PCA (PRPCA) [82].

Multidimensional Scaling (MDS)

MDS [39] aims to preserve the similarity (later distances) of the data points in the embedding space and in the input space. There are different types of multidimensional scaling technique, and in this section we will examine the classical MDS that was first introduced by Torgerson and Gower [146]. Classical MDS measures similarity using the Euclidean distance and the mapping $\mathbf{B} : \mathbf{Y} \rightarrow \mathbf{X}$ by minimizing the following cost function (stress), where for cMDS $k = 1$ and $m = 2$.

$$E(\|\mathbf{D} - \mathbf{d}\|) = \sum_{ij}^M w_{ij} (\delta_{ij}^k - d_{ij}^k)^m = \|\mathbf{D}^{\mathbf{X}} - \mathbf{D}^{\mathbf{Y}}\|_F^2 = \|(-\frac{1}{2}\mathbf{H}\mathbf{D}^{\mathbf{X}}\mathbf{H}) - \mathbf{Y}^T\mathbf{Y}\|_F^2, \quad (2.8)$$

which represents the error between the dissimilarities D and the corresponding distances d , where: $i, j = \{1, \dots, M\}$, w_{ij} are weights, $\|\cdot\|_F$ is the Frobenius norm, and k, m are the parameters.

Algorithm 2: Multidimensional Scaling scheme.

Input: proximity matrix $\mathbf{D} = [d_{ij}^2]$.

procedure MDS:

1. Apply double centering: $\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}\mathbf{H}$ using the centering matrix $\mathbf{H} = \mathbf{I} - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^T$.
 2. Find the n largest eigenvalues $\lambda_1, \dots, \lambda_n$ and corresponding eigenvectors (v_1, \dots, v_n) of \mathbf{B} .
 3. $\mathbf{Y} = \mathbf{U}_n \Lambda_n^{\frac{1}{2}} = (\sqrt{\lambda_1}v_1, \dots, \sqrt{\lambda_n}v_n) = (y_1, \dots, y_M)^T$.
-

The MDS algorithm described here is used in a wide variety of applications, such as surface matching [13], psychometrics [130], and more [11]. The major drawback of classical MDS is its sensitivity to noise and the fact that it does not work well when the underlying structure of the data is non-linear. For these reasons, many variations of MDS were developed later. Few of those are:

- Generalized classical MDS [46].
- Metric MDS [46] tries to preserve the distances of the points in the embedding space rather than similarities.
- Non-metric MDS [46], which rather than using a distance metric, $d_y(x_i, x_j)$, for the distances between points in the embedding space, uses $f(d_y(x_i, x_j))$ where $f(\cdot)$ is a non-parametric monotonic function.
- Sammon mapping [46] is a special case of metric MDS. Introduce changes to the MDS optimization formulation.

Additionally, MDS has inspired the non-linear manifold learning technique Isomap, which we will cover next.

Isometric mapping (Isomap)

Isomap [134] is a special case of the generalized classical MDS, which gives a closed form solution to the dimensionality reduction problem and uses the Euclidean

distance as the similarity metric, while Isomap uses an approximation of geodesic distances.

Algorithm 3: Isomap scheme.

Input: k -nearest neighbor graph $G(X, E)$.

procedure Isomap:

1. Compute the shortest path distances between all pairs of vertices x_i and x_j and store them in \mathbf{S}_{ij} .
 2. Create dissimilarity matrix $\mathbf{D}=\mathbf{S}_{ij}^2$.
 3. Apply the MDS algorithm using \mathbf{D} as input.
-

The geodesic distance is the length of the shortest path between two points on the possibly curvy manifold. It is ideal to use the geodesic distance; however, calculating the geodesic distance is very difficult because it requires traversing from one point to another point on the manifold. Isomap approximates the geodesic distance by pairwise Euclidean distances and finds the k -nearest neighbors (k NN) graph of the dataset. Then, the shortest path between two points, through their neighbors, is found using a shortest-path algorithm such as for example the Dijkstra algorithm [22] or the Floyd-Warshall algorithm [22] ($O(N^3)$).

Most Isomap variants are heavily focused on technical aspects of implementation to increase computation speed [79, 109].

Laplacian eigenmaps

Laplacian eigenmaps [6, 5] attempt to capture information about the local geometry and reconstruct the global geometry from the local information. The locality-preserving character of the Laplacian eigenmap algorithm makes it relatively insensitive to outliers and noise. It is also not prone to short circuits, as only the local distances are used. The algorithm constructs a weighted graph with k nodes, one for each point, and a set of edges that connect neighboring points. The embedding map is provided by computing the eigenvectors of the graph Laplacian. There are two options for choosing the edge weights:

1. **Heat kernel:** If x_i and x_j are connected by an edge, then set the edge weight as:

$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{t}\right) \quad (2.9)$$

otherwise, set $W_{ij}=0$.

2. **Combinatorial:** $W_{ij}=1$ if the vertex i and j are connected by an edge and $W_{ij}=0$ otherwise. This choice of weights avoids the need to choose a parameter t , making it easier to apply.

Let $G(V, E)$ be the graph constructed according to the previous two steps. Furthermore, assume that G is a connected graph; otherwise, apply the current step to each connected component of G . Compute eigenvalues and eigenvectors of the generalized eigenvalue problem $\mathbf{L}f = \lambda\mathbf{D}f$, where \mathbf{D} is a diagonal weight matrix called a degree matrix, and its entries are $D_{ii} = \sum_{j \in N(x_i)} W_{ji}$. We call the Laplacian matrix graph

Algorithm 4: Laplacian eigenmaps scheme.**Input:** data matrix \mathbf{X} .**procedure** Laplacian eigenmaps:

1. Constructs a weighted graph $G(X, E)$.
2. Choose weights of the edges E .
3. Compute eigenvalues and eigenvectors $\mathbf{L}\phi^i = \lambda_i\mathbf{D}\phi^i$.
4. Calculate embedding $\mathbf{Y}: y_i \mapsto (\phi_1^1, \dots, \phi_n^n)$.

$\mathbf{L}=\mathbf{D}-\mathbf{W}$. By the spectral theorem, we know that the eigenvalues are real. We order the eigenvalues in increasing order $\lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_n$, and let ϕ^i be the corresponding eigenvectors such that $\mathbf{L}\phi^i = \lambda_i\mathbf{D}\phi^i$. We leave out the zeroth eigenvector (since it is constant) and proceed with the embedding using the following map:

$$y_i \mapsto (\phi_1^1, \dots, \phi_n^n). \quad (2.10)$$

where ϕ_i^j stands for i -th component of the j -th eigenvector. The idea behind the method is to map close together points to close together points in the new dimension reduced space.

There have been recent developments on the basic Laplacian eigenmap. Some examples are Laplacian Eigenmaps Latent Variable Model (LELVM) [15], robust Laplacian eigenmap [117] and Laplacian forest [83].

Locally Linear Embedding (LLE)

LLE [116] consists of three steps. First, it finds the k -nearest neighbors (k NN) graph of all training points. Then, it tries to find weights for reconstructing every point by its neighbors, using linear combination. Using the same weights, it embeds every point using a linear combination of its embedded neighbors. The main idea of LLE is to use the same reconstruction weights in the low-dimensional embedding space as in the high-dimensional input space.

1. A k NN graph is formed using pairwise Euclidean distance between the data points
2. Linear reconstruction by the neighbors: We find the weights for the linear reconstruction of every point by its k NN. The optimization for this linear reconstruction in the high-dimensional input space is formulated as follows:

$$\epsilon(\tilde{W}) := \sum_{i=1}^M \|x_i - \sum_{j=1}^k \tilde{w}_{ij}x_{ij}\|_2^2 \quad (2.11)$$

where $\mathbb{R}^{M \times k} \ni \tilde{W} := [\tilde{w}_1, \dots, \tilde{w}_n]^T$ includes the weights, $\mathbb{R}^k \ni \tilde{w}_i := [\tilde{w}_{i1}, \dots, \tilde{w}_{ik}]^T$ includes the weights of linear reconstruction of the i -th data point using its k neighbors, and $x_{ij} \in \mathbb{R}^N$ is the j -th neighbor of the i -th data point.

3. Linear embedding: We embed the data in the low-dimensional embedding space using the same weights as in the input space. This linear embedding can be formulated as the following optimization problem:

$$\text{minimize } \sum_{i=1}^M \|y_i - \sum_{j=1}^M w_{ij}y_j\|_2^2 \quad (2.12)$$

Algorithm 5: LLE scheme.**Input:** k -nearest neighbor graph $G(X, E)$.**procedure** Laplacian eigenmaps:

1. Constructs a weighted graph $G(X, E)$.
2. Linear reconstruction of each point by its k neighbors.
3. Linear embedding.

The most known variations of LLE are the following:

- Kernel LLE [45],
- Incremental LLE [70] to handle online data by embedding new received data using the already embedded data.
- Landmark LLE [141] used in big data, that approximates the embedding of all points. using the embedding of some landmark.

Currently, the most widely used nonlinear dimension reduction algorithms (described in Section 1.4) are: 1) t-distributed Stochastic Neighbor Embedding (t-SNE) and 2) Uniform Manifold Approximation and Projection (UMAP). UMAP produces similar or better representations, as it preserves more global features of the data, and the performance of the algorithm itself, measured by the Procrustean measure [53] (a form of statistical shape analysis), is more stable. Furthermore, UMAP, in terms of both dimensionality and data size, is more efficient than t-SNE. However, to gain a good understanding of the modern algorithms that have emerged in recent years, we need to go back to the classical algorithms. An advantage of probabilistic methods is that they are relatively robust to noise because of their stochastic behavior.

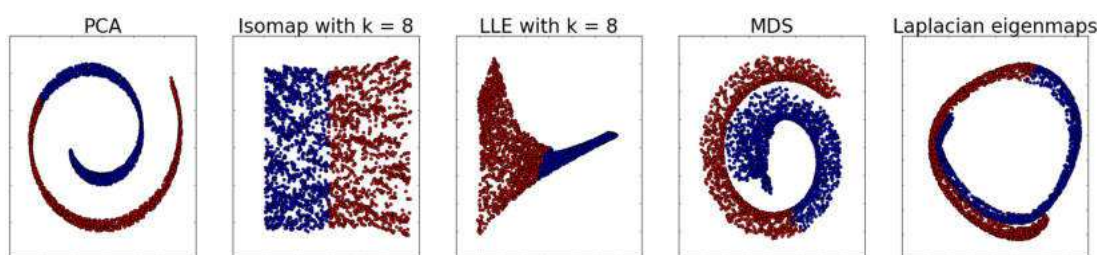


Figure 2.2: Comparison of baseline state-of-the-art spectral methods using "swiss roll" dataset [89].

Figure 2.2 presents visualizations for the *swiss roll* dataset created by the unsupervised spectral methods mentioned above. As shown, LLE and Isomap completely distort the overall structure of the dataset while reducing to 2-D. Best results are achieved by MDS and PCA.

2.3.2. Supervised methods

Unlike unsupervised visualization, supervised variants visualize the underlying structure or distribution of the data using corresponding output variables, or labels (algorithms know both input and output).

Linear Discriminant Analysis (LDA)

LDA [135], unlike PCA, is a supervised method and computes linear directions that maximize separation between multiple classes. This is mathematically expressed as maximizing

$$\begin{aligned} \mathbf{W}_0 = \arg \max_{\mathbf{W}} \quad & tr(\mathbf{W}^T \mathbf{S}_b \mathbf{W}) \\ \text{subject to} \quad & \mathbf{W}^T \mathbf{S}_w \mathbf{W} = \mathbf{I} \end{aligned} \quad (2.13)$$

\mathbf{S}_w being scatter matrix within-class and \mathbf{S}_b being the scatter matrix between classes.

The solution of Eq. 2.8 is given from the generalized eigenvalue problem:

$$\mathbf{S}_b \mathbf{W} = \mathbf{S}_w \mathbf{W} \mathbf{\Lambda} \quad (2.14)$$

Algorithm 6: Linear Discriminant Analysis scheme.

Input: data matrix \mathbf{X} .

procedure LDA:

1. Find eigenvectors of \mathbf{S}_w that correspond to non-zero eigenvalues by performing eigen-analysis to $(\mathbf{I} - \mathbf{M})\mathbf{X}^T \mathbf{X}(\mathbf{I} - \mathbf{M})$ and computing.
 2. $\mathbf{U} = \mathbf{X}(\mathbf{I} - \mathbf{M})\mathbf{V}_w \mathbf{\Lambda}_w^{-1}$.
 3. Project the data as $\tilde{\mathbf{X}}_b = \mathbf{U}^T \mathbf{X} \mathbf{M}$.
 4. Perform PCA on $\tilde{\mathbf{X}}_b$ to find \mathbf{Q} .
 5. The total transform is $\mathbf{W} = \mathbf{U} \mathbf{Q}$.
-

Given the properties of the scatter matrix [102], the objective function of Eq. 2.8 can be expressed as:

$$\begin{aligned} \mathbf{W}_0 = \arg \max_{\mathbf{W}} \quad & tr(\mathbf{W}^T \mathbf{X} \mathbf{M} \mathbf{M}^T \mathbf{X}^T \mathbf{W}) \\ \text{subject to} \quad & \mathbf{W}^T \mathbf{X}(\mathbf{I} - \mathbf{M})(\mathbf{I} - \mathbf{M})\mathbf{X}^T \mathbf{W} = \mathbf{I} \end{aligned} \quad (2.15)$$

The optimization of this problem involves a procedure called Simultaneous Diagonalization. Let us assume that the final transform matrix has the form:

$$\mathbf{W} = \mathbf{U} \mathbf{Q} \quad (2.16)$$

Supervised Multidimensional Scaling (SMDS)

SMDS [145] compared to its unsupervised variant only changes the criterion, which is minimized. It includes two steps: 1) evaluating pairwise distances among entities based on their labels and constructing a new space based on the distance matrix using a projection strategy similar to MDS, and 2) establishing an explicit linear relationship between the feature space and the new space.

The first step aims to construct a new space in which the distances among the entities approximate the distances among their labels. Unlike classic MDS, which establishes a low-dimensional new space based on the distance matrix generated from the features, SMDS establishes a new space based on the distance matrix generated from the labels. By replacing the distance matrix $\mathbf{D}^{\mathbf{X}}$ in Eq. 2.8 with the distance matrix $\mathbf{D}^{\text{Label}}$, a new space represented by $\mathbf{Y}^{\text{Label}}$ can also be obtained as follows:

$$\min_{\mathbf{Y}^{\text{Label}}} \left(\left\| -\frac{1}{2} \mathbf{H} \mathbf{D}^{\text{Label}} \mathbf{H} - ((\mathbf{Y}^{\text{Label}})^T \mathbf{Y}^{\text{Label}}) \right\|_F^2 \right) \quad (2.17)$$

The second step aims to establish a linear model $\mathbf{Y}^{\text{Label}} = \mathbf{W} \mathbf{X}^{\text{Feature}}$, which projects the high-dimensional features of each training entity (indicated by $\mathbf{X}^{\text{Feature}}$) into the new space obtained above (indicated by $\mathbf{Y}^{\text{Label}}$).

$$\min_{\mathbf{W}} \left(\left\| \mathbf{Y}^{\text{Label}} - \mathbf{W} \mathbf{X}^{\text{Feature}} \right\|_F^2 \right) \quad (2.18)$$

Enhanced Supervised Isomap (ES-Isomap)

The ES-Isomap [112] is based on the dissimilarity matrix constructed as:

$$D(x_i|x_j) = \begin{cases} \left(\frac{a-1}{a}\right)^{\frac{1}{2}}, & \text{if } c_i = c_j, \\ a^{\frac{1}{2}} - d_0, & \text{if } c_i \neq c_j \end{cases}, \quad (2.19)$$

where $a = 1/e^{-d_{ij}^2/\sigma}$, σ is a smoothing parameter (set according to the 'density' of the data), d_0 is a constant ($0 \leq d_0 \leq 1$) and c_i, c_j are the labels of the data.

Algorithm 7: ES-Isomap scheme.

Input: data matrix \mathbf{X} .

procedure ES-Isomap:

1. Compute dissimilarity matrix using class labels in the distance matrix.
 2. Run Isomap using dissimilarity matrix from step 1.
 3. Learn the embedded mapping.
 4. SVM testing on new points data.
-

Supervised Laplacian Eigenmaps

Similarly as in ES-Isomap, a crucial difference between supervised and unsupervised variants of Laplacian eigenmaps is how the neighbor graph is being calculated. In S-LapEig [63] the dissimilarity distance between two points x_i and x_j is defined as:

$$D(x_i|x_j) = \begin{cases} \sqrt{1 - \exp(-d_{ij}^2)/\beta}, & \text{if } c_i = c_j \\ \sqrt{\exp(d_{ij}^2)}, & \text{if } c_i \neq c_j \end{cases}, \quad (2.20)$$

where d_{ij} denotes the Euclidean distance between x_i and x_j , β is set to the average Euclidean distance between all pairs of data points and c_i, c_j are the labels of the data.

Supervised LLE (SLLE)

SLLE [113] was introduced to deal with datasets containing multiple (often disjoint) manifolds, corresponding to classes. For fully disjoint manifolds, the local neighborhood of a sample x_i from class c ($1 \leq c \leq C$) should be composed of samples belonging to the same class only. This can be achieved by artificially increasing the precalculated distances between samples belonging to different classes, but leaving them unchanged if samples are from the same class:

$$G' = G + \alpha \max(G)\Lambda, \quad \alpha \in [0, 1], \quad (2.21)$$

where G is the square distance matrix, $\max(G)$ is the maximum entry of G , $\Lambda_{ij}=1$ if x_i and x_j belong to the same class and 0 otherwise. When $\alpha = 0$, one obtains unsupervised LLE; when $\alpha = 1$, the result is fully supervised.

2.4. Probabilistic dimensionality reduction

Probabilistic dimensionality reduction methods assume that there is a low-dimensional embedding influenced and caused by its high-dimensional representation. Probabilistic methods try to infer and discover this link. The advantage of the probabilistic approach over spectral methods is the handling of missing data.

2.4.1. Uniform Manifold Approximation and Projection (UMAP) as a general approach to data embedding and visualization

UMAP [90] is a dimension reduction technique that can be used for visualization and also for general non-linear dimension reduction. The algorithm is founded on three crucial assumptions about the data:

1. The data is uniformly distributed on a Riemannian manifold [119].
2. The Riemannian metric is locally constant (or can be approximated as such).
3. The manifold is locally Euclidean (Def. 5).

On the basis of these assumptions, it is possible to model the manifold using a fuzzy topological structure. The embedding is found by searching for a low-dimensional projection of the data that has the closest possible equivalent fuzzy topological structure. UMAP uses local approximations of manifolds and combines their local representations of fuzzy simplicial sets to construct a topological representation of high-dimensional data. Given some low-dimensional representation of the data, a similar process can be used to construct an equivalent topological representation. UMAP then optimizes the layout of the data representation in a low-dimensional space to minimize the cross-entropy between the two topological representations. The construction of fuzzy topological representations can be divided into two problems: the approximation of the manifold on which the data are inherently based and the construction of a fuzzy representation of the simplicial sets of this approximated manifold.

Uniform distribution of data on a manifold and geodesic approximation

The first step of the UMAP algorithm is to approximate the manifold in which the data lie (approximately). The manifold may be known a priori (simply \mathbb{R}^N) or may need to be inferred from the data. Suppose that the manifold is not known in advance and that we wish to approximate the geodesic distance on it. Let the input data be $X = [x_1, \dots, x_N] \in \mathbb{R}^{M \times N}$, where M is the sample size, and N is the dimensionality. As shown in the work of Belkin and Niyogi on Laplacian eigenmaps [5, 6], for theoretical reasons, it is beneficial to assume that the data are uniformly distributed in the manifold. Additionally, if we assume that the manifold has a Riemannian metric not inherited from the ambient space, we can find a metric such that the data are approximately uniformly distributed regarding that metric. Formally, let \mathcal{M} be the data manifold on which to lie, and let g be the Riemannian metric on \mathcal{M} . For each point $p \in \mathcal{M}$, we have g_p , which is an inner product in the tangent space $\mathcal{T}_p\mathcal{M}$.

Lemma 2 *Let (\mathcal{M}, g) be a Riemannian manifold in ambient \mathbb{R}^N , and let $p \in \mathcal{M}$ be a point. If g is locally constant about p in an open neighborhood U such that g is a constant diagonal matrix in the ambient coordinates, then in a ball $B \subseteq U$ centered on p with volume $\frac{\pi^{n/2}}{\Gamma(n/2+1)}$ with respect to g , the geodesic distance from p to any point $q \in B$ is $\frac{1}{r}d_{\mathbb{R}^N}(p, q)$, where r is the radius of the ball in the ambient space and $d_{\mathbb{R}^N}$ is the existing metric in the ambient space.*

If we assume that the data are uniformly distributed on \mathcal{M} (with respect to g), then away from any boundaries, any ball of fixed volume should contain approximately the same number of points of X regardless of where it is centered in the manifold.

Data Graph in the Input Space UMAP inspired by t-SNE [86] uses the Gaussian or Radial Basis Function (RBF) kernel to measure the similarity between points in the input space. The probability that a point x_i has the point x_j as its neighbor can be calculated by the similarity of these points:

$$p_{ji} = \begin{cases} \exp\left(-\frac{\|x_i, x_j\|_2 - \rho_i}{\sigma_i}\right), & \text{if } x_j \in \mathcal{N}_i \\ 0, & \text{otherwise} \end{cases}, \quad (2.22)$$

where $\|\cdot\|_2$ denotes the norm ℓ_2 . The ρ_i is the distance from x_i to its nearest neighbor:

$$\rho_i = \min\|x_i - x_{i,j}\|_2 \mid 1 \leq j \leq k. \quad (2.23)$$

σ_i is the scale parameter calculated so that the total similarity of the point x_i to its nearest neighbors k is normalized. By binary search, we find σ_i to satisfy:

$$\sum_{j=1}^k \exp\left(-\frac{\|x_i - x_{i,j}\|_2 - \rho_i}{\sigma_i}\right) = \log_2(k). \quad (2.24)$$

The Eq. 2.22 is a measure of directional similarity. To have a symmetric measure with respect to i and j , UMAP symmetrizes it as:

$$\mathbb{R} \ni p_{ij} := p_{ji} + p_{ij} - p_{ji}p_{ij}. \quad (2.25)$$

Data Graph in the Embedding Space Let the embeddings of the points be $\mathbf{Y}=[y_1, \dots, y_M] \in \mathbb{R}^{n \times M}$, where n is the dimensionality of the embedding space. In the embedding space, the probability that a point y_i has the point y_j as its neighbor can be calculated by the similarity of these points:

$$\mathbb{R} \ni q_{ij} := (1 + a\|y_i - y_j\|_2^{2b})^{-1}, \quad (2.26)$$

which is symmetric with respect to i and j . The variables $a > 0$ and $b > 0$ are hyper-parameters determined by the user.

Fuzzy topological representation

UMAP uses functors between the relevant categories to convert metric spaces to fuzzy topological representations. This will provide a means of merging incompatible local views of the data. The topological structure of choice is that of simplicial sets, which are a means to construct topological spaces out of simple combinatorial components. This allows one to reduce the complexity of dealing with the continuous geometry of topological spaces to the task of relatively simple combinatorics and counting. This method of taking geometry and topology is fundamental to UMAP topological data analysis.

The first step is to provide some simple combinatorial building blocks called *simplices*. Geometrically, a simplex is a very simple way to build a k -dimensional object. A k dimensional simplex is called a k simplex and is formed by taking the convex hull of $k+1$ independent points. Thus, a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron. Such a simple construction allows for easy generalization to arbitrary dimensions and provides a basic building block. Formally, simplicial sets are most easily defined purely abstractly in the language of category theory.

Definition 6 *The category Δ has as objects the finite-order sets $[n] = \{1, \dots, n\}$ with morphisms given by (nonstrictly) order-preserving maps. Following the standard category-theoretic notation, Δ^{op} denotes the category with the same objects as Δ and the morphisms given by the morphisms of Δ with the direction (domain and codomain) reversed.*

Definition 7 *A simplicial set is a functor from Δ^{op} to **Sets**, the category of sets; that is, a contravariant functor from Δ to **Sets**.*

However, to construct complex topological spaces, we need to be able to combine simplices. A *simplicial complex* \mathcal{K} is a set of simplices such that any face of any simplex in \mathcal{K} is also in \mathcal{K} and the intersection of any two simplices in \mathcal{K} is a face of both simplices. To construct a *simplicial complex* from a topological space, UMAP uses the Čech or Vietoris-Rips complex [56] given an open cover of a topological space. The key difference between two complexes is that Vietoris-Rips is entirely determined by the 0 and 1 simplices.

Optimizing a low-dimensional representation

In contrast to the source data, where we want to estimate a manifold on which the data are uniformly distributed, a target manifold for Y is chosen a priori (usually this will simply be \mathbb{R}^n itself, but other choices such as n -spheres or n -tori are certainly possible). Therefore, we know the manifold and the manifold metric a priori, and we can compute the fuzzy topological representation directly. In particular, we still want to incorporate the distance to the nearest neighbor according to the local connectivity requirement. This can be achieved by providing a parameter that defines the expected distance between the nearest neighbors in the embedded space. Given fuzzy simplicial set representations of X and Y , a means of comparison is required. If we consider only the 1-skeleton of fuzzy simplicial sets, we can describe each as a fuzzy graph, or, more specifically, a fuzzy set of edges. To compare two fuzzy sets, UMAP uses the fuzzy set cross-entropy.

Definition 8 *The cross entropy C of two fuzzy sets (A, μ) and (A, ν) is defined as:*

$$c_1 := \sum_{i=1}^n \sum_{j=1, j \neq i}^n (p_{ij} \ln(\frac{p_{ij}}{q_{ij}}) + (1 - p_{ij}) \ln(\frac{1 - p_{ij}}{1 - q_{ij}})), \quad (2.27)$$

The first term in Eq. 2.27 is the attractive force that attracts the embeddings of neighboring points toward each other. This term should only appear when $p_{ij} \neq 0$, which means that x_j is a neighbor of x_i , or x_i is a neighbor of x_j , or both. The second term in Eq. 2.27 is the repulsive force that repulses the embeddings of non-neighbor points away from each other. As the number of all permutations of non-neighbor points is very large, computation of the second term is non-tractable in big data.

Inspired by Word2Vec [91] and LargeVis [131], UMAP uses negative sampling [90] where, for every point x_i , m points are randomly sampled from the training dataset and treated as non-negative (negative) points for x_i . As the dataset is usually large, the sampled points will be actual negative points with high probability. The summation over the second term in Eq. 2.27 is computed only on these negative samples, rather than on all negative points. UMAP changes the data graph in the embedding space to make it similar to the data graph in the input space. Eq. 2.27 is the cost function minimized in UMAP where the optimization variables are $\{y_i\}_{i=1}^n$:

$$\begin{aligned} \min_{\{y_i\}_{i=1}^n} c_1 &:= \min_{\{y_i\}_{i=1}^n} \sum_{i=1}^n \sum_{j=1, j \neq i}^n (p_{ij} \ln(p_{ij}) - p_{ij} \ln(q_{ij})) \\ &+ (1 - p_{ij}) \ln(1 - p_{ij}) - (1 - p_{ij}) \ln(1 - q_{ij}) \\ &= \min_{\{y_i\}_{i=1}^n} - \sum_{i=1}^n \sum_{j=1, j \neq i}^n (p_{ij} \ln(q_{ij}) + (1 - p_{ij}) \ln(1 - q_{ij})) \end{aligned} \quad (2.28)$$

According to Eqs. 2.22, 2.25, and 2.26, in contrast to q_{ij} , p_{ij} is independent of the optimization variables $\{y_i\}_{i=1}^n$. Hence, we can drop the constant terms to revise the cost function:

$$c_2 := - \sum_{i=1}^n \sum_{j=1, j \neq i}^n (p_{ij} \ln(q_{ij}) + (1 - p_{ij}) \ln(1 - q_{ij})), \quad (2.29)$$

which should be minimized.

Similarly to t-SNE, UMAP can optimize the embedding Y with respect to the cross entropy of the fuzzy set C using stochastic gradient descent. However, this requires a differentiable fuzzy singular set functor. If the expected minimum distance between points is zero, the fuzzy singular set functor is differentiable for these purposes; however, for any nonzero value, we need to make a differentiable approximation (chosen from a suitable family of differentiable functions). The complete algorithm presents as follows: By using manifold approximation and patching together local fuzzy simplicial set representations, UMAP constructs a topological representation of the high-dimensional data. Then, it optimizes the layout of data in a low-dimensional space to minimize the error between the two topological representations. The whole process can be extended to the comparison of \mathcal{L} -skeleta with fuzzy simplicial sets instead of the 1-skeleton. Then, the cost function is defined as follows:

$$C_{\mathcal{F}}(X, Y) = \sum_{i=1}^{\mathcal{L}} \lambda_i C(X_i, Y_i), \quad (2.30)$$

where X_i denotes the fuzzy set of i -simplices of X and λ_i are appropriately chosen real-valued weights. While such an approach captures the overall topological structure more accurately, it comes at a non-negligible computational cost due to the increasingly large number of higher-dimensional simplices. For this reason, current implementations restrict themselves to the 1-skeleton.

A computational view of UMAP

UMAP can be ultimately described in terms of weighted graph construction and operations. In particular, this situates UMAP in the class of k -neighbor based graph learning algorithms such as Laplacian Eigenmaps, Isomap and t-SNE. As with other k -neighbor graph-based algorithms, UMAP can be described in two phases. In the first phase, a particular weighted k -neighbor graph is constructed. In the second phase, a low-dimensional layout of this graph is computed. The differences between all algorithms in this class amount to specific details in how the graph is constructed and how the layout is computed. From previous sections, UMAP assumes these axioms to be true:

1. There exists a manifold on which the data would be uniformly distributed.
2. The underlying manifold of interest is locally connected.
3. The primary goal is to preserve the topological structure of this manifold.

Any algorithm that attempts to use a mathematical structure similar to a k -neighbor graph to approximate a manifold must follow a similar basic structure.

- Graph Construction

1. Construct a weighted k NN graph.
2. Apply some transform on the edges to the ambient local distance.

3. Deal with the inherent asymmetry of the k NN graph.
- Graph Layout
 1. Define an objective function that preserves the desired characteristics of this k NN graph.
 2. Find a low-dimensional representation that optimizes this objective function.

Algorithm 8: UMAP scheme.

Input: k NN graph $G(X, E)$.

procedure UMAP:

Initialize \mathbf{Y} using Laplacian eigenmap.

Calculate $p_{ij}, q_{ij} \forall i, j \in \{1, \dots, n\}$ (Eqs. 2.25 and 2.26).

$\eta \leftarrow 1, \nu \leftarrow 1$

while not converged **do**

for i from 1 to M **do**

for j from 1 to M **do**

$u \sim U(0, 1)$

if $u \leq p_{ij}$ **then**

$y_i \leftarrow y_i - \eta \frac{\delta c_{i,j}^a}{\delta y_i}$

$y_j \leftarrow y_j - \eta \frac{\delta c_{i,j}^a}{\delta y_j}$

for it iterations **do**

$l \sim U\{1, \dots, n\}$

$y_i \leftarrow y_i - \eta \frac{\delta c_{i,l}^a}{\delta y_i}$

$\eta \leftarrow 1 - \frac{\nu}{\nu_{max}}$

return embedding \mathbf{Y} .

2.4.2. Stochastic Neighborhood Embedding heuristics

It was originally developed in 2002 by Sam Roweis and Geoffrey Hinton [59]. SNE starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. The similarity of the data point x_j to the data point x_i is the conditional probability $p_{j|i}$, which x_i would choose x_j as its neighbor if the neighbors were chosen proportionally to their probability density under a Gaussian center at x_i . For nearby datapoints, $p_{j|i}$ is relatively high, while for widely separated datapoints, $p_{j|i}$ will be almost infinite (for reasonable values of Gaussian variance). Mathematically, the conditional probability $p_{j|i}$ is given by:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \quad (2.31)$$

where σ_i is the Gaussian variance that is centered on the data point x_i . For the low-dimensional counterparts y_i and y_j of the high-dimensional data points x_i and x_j , it is possible to compute a similar conditional probability, which is denoted by $q_{j|i}$. The similarity of the map point y_j to the map point y_i is given by the following:

$$q_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} \quad (2.32)$$

If the map points y_i and y_j correctly model the similarity between the high-dimensional data points x_i and x_j , the conditional probabilities p_{ji} and q_{ji} will be equal. Motivated by this observation, the SNE aims to find a low-dimensional data representation that minimizes the mismatch between p_{ji} and q_{ji} . A natural measure of the faithfulness with which q_{ji} models p_{ji} is the Kullback-Leibler divergence. SNE minimizes the sum of Kullback-Leibler divergences over all data points using the gradient descent method. The cost function C is given by:

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right), \quad (2.33)$$

in which P_i represents the conditional probability distribution over all other data points given the data point x_i , and Q_i represents the conditional probability distribution over all other map points given the map point y_i . Because the Kullback-Leibler divergence is not symmetric, different types of error in the pairwise distances in the low-dimensional map are not weighted equally.

The remaining parameter to be selected is the Gaussian variance σ_i that is centered on each high-dimensional datapoint, x_i . It is not likely that there is a single value of σ_i that is optimal for all data points in the data set, because it is likely that the density of the data will vary. In dense regions, a smaller value of σ_i is generally more appropriate than in sparse regions. Any particular value of σ_i induces a probability distribution, P_i , on all other data points. This distribution has an entropy that increases as σ_i increases. SNE performs a binary search for the value of σ_i that produces a P_i with a fixed perplexity specified by the user. Perplexity is defined as follows:

$$Perp(P_i) = 2^{H(P_i)}, \quad (2.34)$$

where $H(P_i)$ is the Shannon entropy of P_i measured in bits:

$$H(P_i) = - \sum_j p_{ji} \log_2 p_{ji} \quad (2.35)$$

Perplexity can be interpreted as a smooth measure of the effective number of neighbors.

Minimizing the cost function C is performed using a gradient descent method:

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{ij} - q_{ij} + p_{ji} - q_{ji})(y_i - y_j) \quad (2.36)$$

Physically, the gradient can be interpreted as the resultant force created by a set of springs between the point on the map y_i and all other points on the map y_j . All springs exert a force along the direction $y_i - y_j$. The spring between y_i and y_j repels or attracts the points on the map, depending on whether the distance between the two on the map is too small or too large to represent the similarity between the two high-dimensional data points. The force exerted by the spring between y_i and y_j is proportional to its

length and proportional to its stiffness, which is the mismatch ($p_{ji} - q_{ji} + p_{ij} - q_{ij}$) between the pairwise similarities of the data points and the points on the map.

Symmetric SNE, t-SNE, bh-SNE

In symmetric SNE [86], we consider a Gaussian probability around every point x_i . The probability that the point $x_i \in \mathbb{R}^N$ takes $x_j \in \mathbb{R}^N$ as its neighbor is:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_k - x_i\|^2/2\sigma_k^2)} \quad (2.37)$$

Note that the denominator of Eq. 2.37 for all points is fixed and, thus, it is symmetric for i and j . Compare this with Eq. 2.31 which is not symmetric. The σ_i^2 is the variance that we consider for the Gaussian distribution used for the x_i . It can be set to a fixed number or by a binary search to make the distribution entropy a specific value [59]. The Eq. (15) has a problem with outliers. If the point x_i is an outlier, its p_{ij} will be extremely small because the denominator is fixed for every point and the numerator will be small for the outlier. However, if we use Eq. 2.31 for p_{ij} , the denominator for all points is not the same, and therefore the denominator for an outlier will also be large, excluding the problem of a large numerator. For this problem mentioned, we do not use Eq. 2.37 and rather we use:

$$p_{ji} = \frac{p_{ij} + p_{ji}}{2n} \quad (2.38)$$

where:

$$p_{ji} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_k - x_i\|^2/2\sigma_i^2)} \quad (2.39)$$

is the probability that $x_i \in \mathbb{R}^N$ takes $x_j \in \mathbb{R}^N$ as its neighbor. In the low-dimensional embedding space, we consider a Gaussian probability distribution for the point $x_i \in \mathbb{R}^N$ to take $y_i \in \mathbb{R}^n$ as its neighbor and make it symmetric (fixed denominator for all points):

$$q_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_k - y_i\|^2)} \quad (2.40)$$

Note that the Eq. 2.40 does not have the problem of outliers as in Eq. 2.37 because even for an outlier, the embedded points are initialized close to each other and not far away.

We want the probability distributions in both the input and embedded spaces to be as similar as possible; therefore, the cost function to be minimized can be a summation of the Kullback-Leibler (KL) divergences:

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right), \quad (2.41)$$

Minimizing the cost function C is performed using a gradient descent method:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) \quad (2.42)$$

t-distributed Stochastic Neighbor Embedding (t-SNE)

The SNE cost functions previously presented are difficult to optimize and are the root cause of why visualizations are not resistant to the phenomenon *crowding*. In 2008, to address this, Laurens van der Maaten created t-SNE [86]. Implement a modified cost function such that (1) uses a symmetrized version of the SNE cost function with simpler gradients and (2) uses a student t-distribution rather than a Gaussian to compute the similarity between two points in the low-dimensional space.

Let $\mathbf{D} = [D_{ij}]$ be the distance table in Y and D_{ij} be the distances between the feature vectors i and j of y_i and y_j , while $\mathbf{d} = [d_{ij}]$ is the respective distance matrix in X . Then the loss function $C = E(\mathbf{D}, \mathbf{d})$ is defined by the Kullback–Leibler (KL) divergence:

$$C = E(\mathbf{D}, \mathbf{d}) = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right), \quad (2.43)$$

where, for the t-SNE algorithm, p_{ij} is approximated by Gaussian $\mathcal{N}(y_i, \sigma)$, while q_{ij} is defined by the Cauchy distribution [86]. Then p_{ij} and q_{ij} are defined as follows:

$$p_{ij} = \frac{\exp(-D_{ij}^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-D_{ik}^2/2\sigma_i^2)} \quad (2.44) \quad q_{ij} = \frac{(1 + d_{ij}^2)^{-1}}{\sum_{k \neq i} (1 + d_{ik}^2)^{-1}} \quad (2.45)$$

To minimize KL divergence, t-SNE uses modern optimal gradient descent optimization schemes [118]. The gradient of the loss function $C(\cdot)$ (Eq. 2.43) is as follows:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j). \quad (2.46)$$

Algorithm 9: t-SNE scheme.

Input: data matrix \mathbf{X} .

procedure t-SNE:

1. Compute pairwise affinities p_{ij} with defined *perplexity*.
2. Set $p_{ij} = \frac{p_{ji} + p_{ij}}{2}$.
3. Sample initial solution $\mathbf{Y}^{(0)}$.

for t from 1 to T **do**

Compute low-dimensional affinities q_{ij} .

Compute gradient $\frac{\delta C}{\delta \mathbf{Y}}$.

Set $\mathbf{Y}^{(t)}$.

Barnes-Hut-SNE (BH-SNE)

This variant [85] of SNE uses metric trees to approximate P by a sparse distribution in which only the values of $O(uN)$ are non-zero and approximate the gradients $\frac{\delta C}{\delta y_i}$ using a Barnes-Hut algorithm.

As input similarities are computed using a (normalized) Gaussian kernel, the probabilities p_{ij} corresponding to dissimilar input objects i and j are (nearly) infinitesimal. Therefore, a sparse approximation of probability p_{ij} can be used without a substantial negative effect on the quality of the final embeddings. In particular, bh-SNE computes

the sparse approximation by finding the nearest neighbors [3u] of each of the N data objects and redefining pairwise similarities p_{ij} as:

$$p_{ji} = \begin{cases} \frac{\exp(-d(x_i, x_j)^2/2\sigma_i^2)}{\sum_{k \in \mathcal{N}_i} \exp(-d(x_i, x_k)^2/2\sigma_i^2)}, & \text{if } j \in \mathcal{N}_i \\ 0 & \text{otherwise} \end{cases} \quad (2.47)$$

where \mathcal{N}_i represents the set of nearest neighbors [3u] of x_i , and σ_i is set so that the perplexity of the conditional distribution is equal to u . The nearest-neighbor sets are found in $O(uN \log N)$ time by building a vantage point tree on the data set.

Some examples of probabilistic dimensionality reduction are factor analysis, whose non-linear extension is the variational autoencoder [67], probabilistic PCA [137], probabilistic LDA [62]. Some other examples are SNE [59] and t-SNE [86], where the Gaussian and Student-t distributions are considered for the embedded space, respectively. A recent successful method is UMAP [90], which optimizes the probability of closeness of the graphs in the input and embedded spaces.

2.4.3. State-of-the-art supervised and unsupervised improvements and simplifications. UMAP and t-SNE.

Both t-SNE [86] and UMAP [90] have become very popular among the DR community. This has motivated the design of their variants, such as, for example, parametric extensions [23, 120]. In addition, both algorithms comprise the same two broad steps: 1) construct a graph of local relationships between datasets, 2) optimize an embedding in a low-dimensional space that preserves the structure of the graph.

t-SNE improvements

t-SNE requires the user to choose an approximation to adjust the width of its Gaussian HD neighborhoods. Although such a single-scale method does a good job of preserving neighborhood sizes close to perplexity, but without achieving similar performance for other neighborhoods, multi-scale approaches typically recover both local and global HD structures much better [76]. Additionally, in its original formulation, t-SNE [86] is a non-parametric manifold learner. The primary limitation of non-parametric manifold learners is that they do not provide a parametric mapping between the high-dimensional data space and the low-dimensional latent space, making it impossible to embed new data points without retraining the model. For this reason, there is plenty of room for improvement.

Parametric t-SNE

The parametric t-SNE [84] parameterizes the nonlinear mapping $f : \mathbf{X} \rightarrow \mathbf{Y}$ by means of a feed-forward neural network with weights W . It uses a deep neural network (DNN) because a neural network with sufficient hidden layers (with non-linear activation functions) is capable of parametrizing arbitrarily complex non-linear functions. The neural network is trained in such a way as to preserve the local structure of

the data in the latent space. The DNN weights are learned to minimize the Kullback-Leibler (KL) divergence C_{t-SNE} . Since t-SNE optimization requires normalization over the embedding distribution in the projection space, gradient descent can only be performed after calculating the edge probabilities over the entire dataset. However, projecting the entire dataset onto a neural network; between each gradient descent step, would be too computationally expensive for optimization. The trick that parametric t-SNE proposes for this problem is to divide the dataset into large batches (e.g., 5000 data points in the original paper [84]), which are used to compute separate graphs that are independently normalized and used continuously during training, meaning that the relationships between elements in different batches are not explicitly preserved.

Perplexity-free Parametric t-SNE

The perplexity-free Parametric t-SNE updates the original parametric t-SNE neural network using σ_{ij} to compute high-dimensional similarities, in a multi-scale fashion. Moreover, it replaces logistic activation functions with piecewise-linear ones (i.e., ReLUs), which do not saturate during training. This simple architectural choice allowed one to simplify the training procedure by eliminating the unsupervised pre-training step introduced in the original implementation.

q-Gaussian SNE

The q-SNE [1] in theory leads to a more powerful and flexible visualization of the 2 dimension mapping than t-SNE and SNE using a q-Gaussian distribution as the low-dimensional data distribution. The q-Gaussian distribution includes the Gaussian distribution and the t-distribution as special cases with $q=1$ and $q=2$. Therefore, q-SNE can also express t-SNE and SNE by changing the parameter q , which allows the best visualization by choosing the parameter q .

The q-Gaussian distribution is derived by maximizing the Tsallis entropy under appropriate constraints and is a generalization of the Gaussian distribution. Let s be a 1-dimensional observation. The q-Gaussian distribution for the observation s is defined as follows:

$$P_q(s; \mu, \sigma^2) = \frac{1}{Z_q} \left(1 + \frac{q-1}{3-q} \frac{(s-\mu)^2}{\sigma^2} \right)^{-\frac{1}{q-1}}, \quad (2.48)$$

where μ and σ are the mean and variance, respectively. The normalization factor Z_q is given by:

$$Z_q = \begin{cases} \sqrt{\frac{3-q}{q-1}} \cdot B\left(\frac{3-q}{2(q-1)}, \frac{1}{2}\right) \cdot \sigma, & 1 \leq q < 3 \\ \sqrt{\frac{3-q}{1-q}} \cdot B\left(\frac{2-q}{1-q}, \frac{1}{2}\right) \cdot \sigma, & q < 1 \end{cases}, \quad (2.49)$$

where B is the beta function. It is known that the q-Gaussian distribution defined by Equation 2.48 always satisfies the inequality:

$$1 + \frac{q-1}{3-q} \frac{(s-\mu)^2}{\sigma^2} \geq 0. \quad (2.50)$$

Similarly to symmetric SNE or t-SNE, q-SNE uses the local Gaussian distribution in a high-dimensional void. The joint probability in the low-dimensional space is defined as:

$$r_{ij} = \frac{(1 + \frac{q-1}{3-q} \|y_i - y_j\|^2)^{-\frac{1}{q-1}}}{\sum_l^N \sum_{k \neq l}^N (1 + \frac{q-1}{3-q} \|y_l - y_k\|^2)^{-\frac{1}{q-1}}}, \quad (2.51)$$

where q is the hyperparameter, $r_{ii} = 0$ and $r_{ij} = r_{ji} \forall i, j$. The Kullback-Leibler divergence and the optimization update rule are the same as in the SNE equations 2.33 and 2.36. The gradient for y_i is given as

$$\frac{\delta C}{\delta y_i} = \frac{4}{3-q} \sum_j^N (p_{ij} - r_{ij})(y_i - y_j) (1 + \frac{q-1}{3-q} \|y_i - y_j\|^2)^{-1}. \quad (2.52)$$

Since the q-Gaussian distribution is an extension of the Gaussian distribution and the t-distribution with the parameter q , q-SNE can generate the same low-dimensional embedded space with SNE or t-SNE by changing the parameter q .

Tree-SNE

Tree-SNE[114] is a hierarchical clustering method based on a one-dimensional t-SNE with decreasing values of α and perplexity at each level. It allows us to visualize and elucidate high-dimensional hierarchical structures by creating t-SNE embeddings with increasingly heavy tails to reveal increasingly fine-grained structures and then stacking these embeddings to create a tree-like structure. Then, it performs spectral clustering on each one-dimensional embedding, computationally determining the number of distinct clusters in the embedding. The number of clusters will increase as the value of α decreases. Tree-SNE defines alpha-clustering of data as the cluster assignment that is stable over the largest range of α values.

The method starts with a standard one-dimensional t-SNE embedding ($\alpha=1$) with a high perplexity, by default equal to the square root of the number of data points. A high initial perplexity increases the effective number of neighbors used by t-SNE, which means that larger clusters will tend to form, capturing more global structures in the data. Using a high error count at the start, a t-SNE tree can show the entire spectrum of data organization, from global structures at the base of the tree to very fine structures. The exact initial value of the error count does not appear to be important in many trials, as long as it is high, because t-SNE is quite robust to small changes in the error count, and most of the interesting features of the data emerge in further embeddings from adjustments in the error count and α .

Local Interpolation with Outlier CoNtrol t-SNE

The LION-tSNE[12] uses a random sampling method for the design of the tSNE model, creating an initial visual environment. Then, new data points are added to this environment using the local-IDW [127] (inverse distance weighting) interpolation method.

Algorithm 10: LION-tSNE scheme.

Input: data point x .
procedure LION-tSNE:
 Find neighbors in radius r_x .
 if found neighbors > 1 **then**
 Perform *IDW*.
 else if found neighbors $== 1$ **then**
 Perform single neighbor placement.
 else
 Perform outlier placement.
return y .

The randomly selected sample data often suffer from non-representativeness of the entire data, which creates inconsistency in the t-SNE environment. To overcome this problem, two new sampling methods were proposed in [12], which are based on graph update properties of k -NN. It has been empirically shown that the proposed methods outperform the existing LION-tSNE method with 0.5 to 2% more precision in k -NN and the results are more consistent. LION t-SNE combines outlier detection and locality into a single approach: it can use local interpolation when a new sample x has neighbors in a certain radius r_x .

Capacity Preserving Mapping

In CPM [144], the low-dimensional embedding \mathbf{Y} is the minimizer of the following optimization problem with the well-defined Capacity Adjusted Distance $\tilde{\mathbf{D}}$:

$$\mathbf{Y} = \arg_{\mathbf{Y}_i} \min \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}, \quad (2.53)$$

where

$$p_{ij} = \frac{(\epsilon + \tilde{\mathbf{D}}^2 \|x_i - x_j\|)^{-1}}{\sum_{k,l,k \neq l} (\epsilon + \tilde{\mathbf{D}}^2 \|x_k - x_l\|)^{-1}}, \quad q_{ij} = \frac{(1 + \|y_i - y_j\|)^{-1}}{\sum_{k,l,k \neq l} (1 + \|y_k - y_l\|)^{-1}}, \quad (2.54)$$

where $\|\cdot\|$ is the distance measure in high-dimensional space, and $\epsilon > 0$ is some small constant used to avoid taking the reciprocal of 0. In addition to the KL divergence, one can use other (dis)similarity measures and obtain different formulations of the optimization problems. Optimization Eq. 2.53 is trying to match the KL divergence between the network probabilities of the original and embedded graphs. The small positive constant ϵ in the expression of p_{ij} and the 1 in the expression of q_{ij} are used to avoid dividing by 0. Although the formulation of Eq. 2.53 looks similar to that of t-SNE (Eq. 2.43), their performances are completely different. The biggest upside of CPM compared to t-SNE is that it: 1) lowers over-stretching, 2) does not have tuning parameters, and 3) preserves more geometry.

Class-aware t-SNE

Cat-SNE [10] explicitly accounts for the class labels in t-SNE to improve the accuracy of KNN in the LD embedding. For this purpose, it modifies the t-SNE adjustment of the individual radius of the normalized Gaussian neighborhood around each datum. Instead of targeting a fixed neighborhood entropy provided by the user through the perplexity, it adjusts the neighborhood radius for neighbors with the same class to cumulate a dominant fraction of the probability distribution. This results in smaller HD neighborhoods near class boundaries than in their bulk, and therefore tends to stretch the former and shrink the latter.

Let c_i be the class associated with x_i (x_i being a point in the HD space X , y_i in the LD space Y). Cat-SNE defines a condition on the weighted proportion t_i of neighbors x_j that share the same class as x_i , that is,

$$t_i = \sum_{j \in Y\{i\}|c_j=c_i} \sigma_{ij} > 0. \quad (2.55)$$

where $t_i \in [0, 1]$ as $\sum_{j \in Y\{i\}|c_j=c_i} \sigma_{ij} = 1$ and the hyper-parameter θ is in $[0.5, 1]$ to ensure the majority of the class c_i . Therefore, the perplexity metaparameter in t-SNE is replaced by the threshold θ . No other changes are made to t-SNE.

Supervised t-SNE

St-SNE [18] incorporates the outcome information, while calculating the $KL(P||Q)$ by defining a similarity measure o_{ij} in the outcome space O as follows:

$$o_{ji} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} \quad (2.56)$$

The low-dimensional representation by minimizing the following cost function C_S :

$$C_S = \rho KL(P||Q) + (1 - \rho) KL(O||Q) = \rho \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - \rho) \sum_{i \neq j} o_{ij} \log \frac{o_{ij}}{q_{ij}}, \quad (2.57)$$

where $\rho \in [0, 1]$ controls the level of supervision in the learning process. Larger ρ reflect less supervision.

Algorithm 11: St-SNE scheme.

Input: data matrix \mathbf{X} .

procedure St-SNE:

Generate $z_1^{(0)}, \dots, z_n^{(0)}$ using multivariate normal distribution with mean zero.

Calculate p_{ij}, q_{ij} .

for t from 1 to T **do**

 Calculate gradient $\frac{\delta C_S}{\delta z}$.

$z^{(t)} = z^{(t-1)} - \eta \frac{\delta C_S}{\delta z} + \alpha \delta^{(t-1)}$.

$\delta^{(t)} = z^{(t)} - z^{(t-1)}$.

return $z = z^T$.

Other improvements

Here, we just list some of the recent improvements of t-SNE and do not explain them in detail for the sake of brevity.

- Dense t-SNE [99], which tackles the problem of local density information for points in denser regions, which have a smaller σ_i^2 .
- Parametric kernel t-SNE [51].
- Fast Interpolation-based t-SNE [81], which accelerates the t-SNE procedure.
- d-SNE [148] used for domain adaptation in neural network training, which uses SNE and a novel modified-Hausdorff distance.
- Approximated and User Steerable tSNE (A-tSNE) [108] is a controllable tSNE approximation, which trades off speed and accuracy, to enable interactive data exploration.

UMAP improvements and simplifications.

UMAP is an emerging dimensionality reduction technique that offers better versatility and stability than t-SNE. Although UMAP is also more efficient than t-SNE, it still suffers from an initial delay of several minutes to produce the first projection, which limits its use in interactive data mining. It was developed in 2018, so not as many new variants were created as in the case of the t-SNE method.

Parametric UMAP

In general, parametric approaches use deep neural networks to preserve the structure of the dataset. The parametric form of UMAP [120], through the use of negative sampling, can in principle be trained on batches as small as a single edge, making it suitable for training the mini-patches needed for memory-intensive neural networks trained on full graphs on large data sets, as well as for online learning. Given these design features, UMAP loss can be used as regularization in typical deep learning paradigms using stochastic gradient descent, without the batch trick on which parametric t-SNE relies (described in previous Section).

Progressive UMAP

Progressive UMAP [68] allows users to feed small batches of data points incrementally into UMAP to obtain the desired latency between intermediate projection outputs. To this end, it identifies sequential procedures in the original UMAP algorithm and transforms them into progressive procedures.

Computing \mathcal{N}_i : To build and maintain the k -nearest neighbor graph, Progressive UMAP leverages the k -NN lookup table from PANENE [103]. PANENE uses randomized kd-trees [97] to approximate and update the k -nearest neighbors of an increasing number of data points. PANENE accepts a parameter called *ops*, which indicates the allowed number of tasks per iteration that can be controlled to find the balance between latency and accuracy.

Computing ρ_i and σ_i : For every data point in $\mathbf{X}_{updated}$ and \mathbf{X}_{new} , Progressive UMAP recomputes ρ_i and σ_i . For space efficiency, UMAP used the Coordinate List (COO), which stores only row, column, and value information as a list of tuples. Progressive UMAP updates the COO by recalculating $v_{j|i}$ for the selected points – $\mathbf{X}_{updated}$ and \mathbf{X}_{new} – and changing the corresponding values if there is a change from the previous ones.

$$\rho_i = \min_{j \in \mathcal{N}_i} \{d(x_i, x_j) | d(x_i, x_j) > 0\}, \quad (2.58)$$

$$\sum_{j \in \mathcal{N}_i} \exp\left(\frac{-\max(0, d(x_i, x_j)) - \rho_i}{\sigma_i}\right) = \log_2(k). \quad (2.59)$$

Using ρ_i and σ_i it can calculate (the same as UMAP) $v_{j|i}$, the weight of the edge from a point x_i to another point x_j :

$$v_{j|i} = \exp\left(\frac{-\max(0, d(x_i, x_j)) - \rho_i}{\sigma_i}\right). \quad (2.60)$$

Layout Initialization: Although spectral embedding produces an effective initial projection, its quadratic time complexity causes severe delay. Progressive UMAP initializes the positions of newly inserted points in two stages. For the first batch of points, 1) it runs the algorithm with a large value of ops (e.g., 15000), using the same spectral embedding technique as UMAP. Because it starts with a relatively small number of points, this would take much less time than the original UMAP spectral embedding. Hereafter, 2) we lower the value of ops (e.g., 1000) not to focus on the appending process, but to obtain an optimized projection output fast.

Layout Optimization: Similarly, Progressive UMAP goes through two stages for layout optimization. As it affects the overall convergence time and increases the stability of the final output, it is very important to position the first batch of points well so that the clusters are unambiguously separated. To this end, 1) it runs more iterations (e.g., 40) in the first batch so that each cluster can settle its position. Afterwards, 2) it runs fewer iterations (e.g., 4) to focus on attaining the projection result fast.

Other improvements

Here, we just list some of the recent improvements of UMAP and do not explain them in detail for the sake of brevity.

- Hierarchical UMAP [88], preserves the mental map while requesting more cluster details, while balancing the trade-off between global and local relationships.
- UMAP [90] by design allows for using categorical label information to do supervised dimension reduction.
- DensMAP [98] computes estimates of the local density and uses those estimates as a regularizer in the optimization of the low-dimensional representation.

Algorithm 12: Progressive UMAP scheme.**Input:** New batch of training data \mathbf{X}_{new} .**procedure** Progressive UMAP:Update k NN graph by *PANENE* method $\rightarrow \mathbf{X}_{new}, \mathbf{X}_{updated}$.**if** is *initial batch* **then**Initialize \mathbf{Y} using Laplacian eigenmap.**else****foreach** $\mathbf{x}_i \in \mathbf{X}_{new}$ **do**

Find nearest neighbor to previously accumulated data.

Initialize \mathbf{y}_i to the embedding of nearest neighbor point.**foreach** $\mathbf{x}_i \in \mathbf{X}_{new}, \mathbf{X}_{updated}$ **do**Calculate ρ_i, σ_i (Eqs. 2.23 and 2.24).Calculate $p_{ij}, q_{ij} \forall i, j \in \{1, \dots, n\}$ (Eqs. 2.25 and 2.26).**while** not converged **do****foreach** $\mathbf{x}_i \in \mathbf{X}_{new}, \mathbf{X}_{updated}$ **do****for** j from 1 to M **do** $u \sim U(0, 1)$ **if** $u \leq p_{ij}$ **then**

$$y_i \leftarrow y_i - \eta \frac{\delta c_{i,j}^a}{\delta y_i}$$

for it iterations **do**

$$l \sim U\{1, \dots, n\}$$

$$y_i \leftarrow y_i - \eta \frac{\delta c_{i,l}^r}{\delta y_i}$$

return embedding \mathbf{Y} for new and updated points.

2.5. Neural network based dimensionality reduction

The category of DR based on neural networks has an approach based on information theory, where the center of a neural network or autoencoder is viewed as an information bottleneck that retains only useful and important information. Some examples are the Restricted Boltzmann machine (RBM) and the Deep Belief Network (DBN) [58], which are fundamental DR methods in a network structure. They have been proposed to avoid the problem of vanishing gradient [77]. Another example is autoencoder, in which the latent embedding space is encoded by the middle layer of a possibly deep autoencoder. There is also deep-metric learning [66], which encodes data in the embedding space trained by a deep neural network and attempts to increase and decrease the interclass and intraclass variances of the data, respectively [49]. Note that metric learning can be seen as a DR, as it can be viewed as a linear or non-linear projection onto the embedding space and then applying the Euclidean distance to that space. In a variational autoencoder [67], the latent space has a specific distribution, such as a Gaussian distribution. Another example is the adversarial autoencoder [87], which uses game theory to optimize the encoding.

More recently, several deep learning metric methods have been proposed that focus on maximizing and minimizing interclass and intraclass variance in data [49]. A

Siamese network is a collection of several networks (usually two or three) that share weights with each other [124]. The weights are trained using loss based on anchor, neighbor (positive) and distant (negative) samples, where anchor and neighbor belong to the same class, but anchor and distant instances belong to different classes. The loss functions used to train a Siamese network typically use anchor, neighbor and distant samples, trying to attract the anchor and neighbor to each other while pushing the anchor and distant tiles away from each other. Two different loss functions that are used to train Siamese networks are triple loss [124] and contrast loss [54] for networks with three and two subnets, respectively.

Autoencoders

Autoencoders [58] can learn non-linear mappings that are required to embed highly non-linear real-world data in the latent space. They primarily focus on maximizing the variance of the data in the latent space and as a result, autoencoders are less successful in retaining the local structure of the data in the latent space compared to manifold learners.

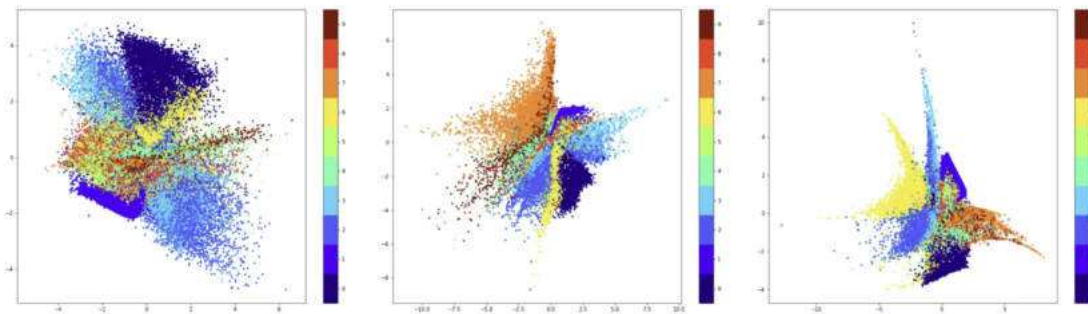


Figure 2.3: Visualization of the latent spaces generated by three different variational autoencoder (VAE) instances trained on the same MNIST data.

Autoencoders can be used to reduce the dimensionality of data, but in our work, we do not consider cases where the dataset has been preprocessed with an autoencoder. Such datasets already have a structure, which contains global information about itself, thus it will strongly affect the final visualization generated by the embedding method (e.g. t-SNE or UMAP).

In Figure 2.3 we can observe that in all cases VAE tends to produce a very smooth latent space and preserves global structure very well. In contrast, the strong clustering of UMAP, for example, very clearly separates similar digits, while in the case of VAE they smoothly transition into each other.

CHAPTER 3

TOWARDS OPTIMAL NN GRAPH VISUALIZATION

Many successful techniques have been reported recently that first compute a similarity structure of the data points and then project them into a low-dimensional space with the structure preserved [16]. These two steps suffer from considerable computational costs, preventing state-of-the-art methods (such as the SNE variants) from scaling to large-scale and high-dimensional data (e.g., millions of data points and hundreds of dimensions).

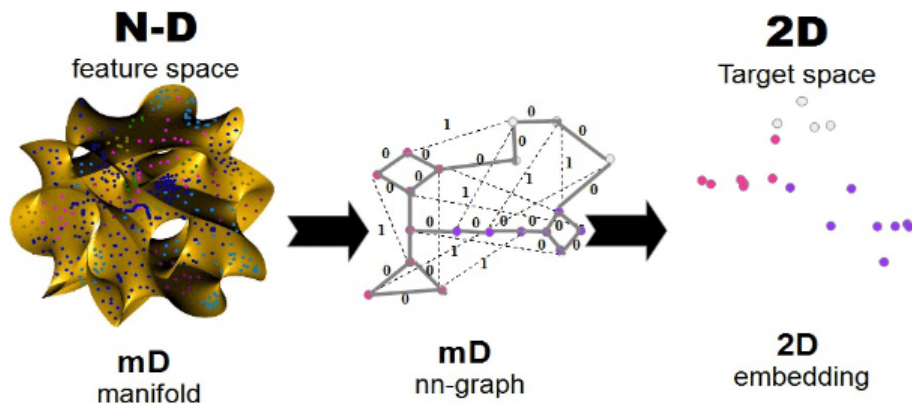


Figure 3.1: The general idea of data embedding by means of the nearest neighbor graph. Source: [32].

The construction of k NN graphs from high-dimensional data is critical for many applications, such as similarity search, collaborative filtering, manifold learning, and network analysis. Although the exact computation of a k NN has a complexity of $O(M^2N)$ (with the number of data points M and the number of dimensions N) that is too costly, existing approaches use roughly three categories of techniques: space partition trees [8, 40, 128, 25], locality-sensitive hashing techniques [26, 154], and neighbor exploration techniques [29]. Space partitioning methods divide the entire space into different regions and organize the regions into different tree structures, e.g., k -d trees [8, 40], vp -trees [151], cover trees [9] and random projection trees [25]. Once the trees have been constructed, the nearest neighbors of each data point can be found by traversing

the trees. Locality-sensitive hashing techniques [26] deploy multiple hashing functions to map data points to different buckets, and data points in the same buckets are likely to be similar to each other. Neighbor-exploring techniques, such as NN descent [29], are built on the intuition that "my neighbors neighbor are likely to be my neighbors as well". Starting from an initial nearest-neighbor graph, the algorithm iteratively refines the graph by exploring the neighbors of neighbors defined according to the current graph.

3.1. LargeVis

Technique that first constructs an accurately approximated k NN graph from the data and then layouts the graph in a low-dimensional space. Compared to t-SNE, LargeVis significantly reduces the computational cost of the graph construction step and employs a principled probabilistic model for the visualization step, the objective of which can be effectively optimized through asynchronous stochastic gradient descent with linear time complexity. The whole procedure thus easily scales to millions of high-dimensional data points. Existing approaches generally first compute the similarities of all pairs of x_i, x_j and then preserve the similarities in the low-dimensional transformation. As computing pairwise similarities is too expensive (i.e. $O(N^2d)$), approaches such as the t-SNE or UMAP construct a graph of k -nearest neighbor instead and then project the graph into the $2D$ space. LargeVis follows this procedure, but uses a very efficient algorithm for the construction of the k -nearest neighbor graph and a principled probabilistic model for graph visualization. It is worth mentioning that LargeVis visualizes high-dimensional information and is not used to visualize graphs.

3.1.1. k NN graph construction

LargeVis uses the most common Euclidean metric to compute the graph of k NN in a high-dimensional space. The same approach is used by t-SNE. The algorithm starts by partitioning the entire space and building a tree. Specifically, for each non-leaf node of the tree, the algorithm chooses a random hyperplane to divide the subspace corresponding to the non-leaf node into two, which become the children of that node. The hyperplane is selected by randomly sampling two points from the current subspace and then taking a hyperplane that is equally distant from those two points. This process continues until the number of nodes in the subspace reaches a threshold. After building a random projection tree, each data point can traverse the tree to find the corresponding leaf node. The points in the subspace of this leaf node will be treated as candidates for the nearest neighbors of the input data point. In practice, multiple trees can be built to improve the accuracy of the nearest neighbors. After finding the nearest neighbors of all data points, a k NN graph is built. However, to construct a very accurate k NN graph requires the construction of many trees, which significantly affects the efficiency. This dilemma has been a bottleneck in applying random projection trees to visualization. LargeVis proposes a new solution: instead of building a large number of trees to obtain a highly accurate k NN graph, it uses neighbor exploration techniques to improve the

accuracy of a less accurate graph [29]. Specifically, it builds a few random projection trees to construct an approximate k NN graph, whose accuracy may not be that high. Then, for each node of the graph, it searches for neighbors of its neighbors, which are also likely to be candidates of its nearest neighbors. This process can be repeated for multiple iterations to improve the accuracy of the graph. For the edge weights in the k NN graph, LargeVis uses the same approach as t-SNE. The conditional probability of data x_i and x_j is first calculated as:

$$p_{ji} = \frac{\exp(-d(x_i, x_j)^2/2\sigma_i^2)}{\sum_{(i,k) \in E} \exp(-d(x_i, x_k)^2/2\sigma_i^2)}, \quad p_{ii} = 0 \quad (3.1)$$

where the parameter σ_i is chosen by setting the perplexity of the conditional distribution $p_{\cdot|i}$ equal to the perplexity u . The graph is then symmetrized by setting the weight between x_i and x_j as:

$$w_{ij} = \frac{p_{ji} + p_{ij}}{2N}. \quad (3.2)$$

3.1.2. Probabilistic model for graph visualization

After constructing the k NN graph, LargeVis only needs to project the graph nodes into a 2D / 3D space to visualize the data. For this purpose, it introduces an essential probabilistic model for this purpose. The idea is to preserve the similarity of the vertices in a low-dimensional space. In other words, it wants to keep similar vertices close to each other and dissimilar vertices far apart in a low-dimensional space. Given a pair of vertices (v_i, v_j) , the probability of observing a binary edge $e_{ij} = 1$ between v_i and v_j is first determined as follows:

$$P(e_{ij} = 1) = f(\|y_i - y_j\|) \quad (3.3)$$

where y_i is the embedding of the vertex v_i in the low-dimensional space, $f(\cdot)$ is a probabilistic function with respect to the distance of the vertex y_i and y_j , that is, $d = \|y_i - y_j\|$. When y_i is close to y_j in low-dimensional space (that is, d is small), there is a high probability of observing a binary edge between the two vertices. In reality, many probabilistic functions can be used, such as $f(x) = \frac{1}{1+ax^2}$ or $f(x) = \frac{1}{1+\exp(x^2)}$. Different probabilistic functions are compared in [131].

Equation 3.3 only determines the probability of observing a binary edge between a pair of vertices. To further extend it to general weighted edges, LargeVis defines the probability of observing a weighted edge $e_{ij} = w_{ij}$ as follows:

$$P(e_{ij} = w_{ij}) = P(e_{ij} = 1)^{w_{ij}} \quad (3.4)$$

With the above definition, given a weighted graph $G = (V, E)$, the likelihood of the graph can be calculated as follows:

$$O = \prod_{(i,j) \in E} (p(e_{ij} = 1))^{w_{ij}} \prod_{(i,j) \in \bar{E}} (1 - (p(e_{ij} = 1)))^y \quad (3.5)$$

in which \tilde{E} is the set of pairs of vertices that are not observed and γ is the unified weight assigned to the negative edges. The first part of the above equation models the probability of observed edges, and by maximizing this part, similar data points will be kept close together in the low-dimensional space; the second part models the probability of all pairs of vertices without edges, i.e. negative edges. By maximizing this part, the dissimilar data will move away from each other.

Optimization

Direct optimization of Eq. 3.5 is computationally expensive because the number of negative edges is squared to the number of nodes. Inspired by negative sampling techniques, instead of using all negative edges, LargeVis randomly selects all negative edges to optimize the model. For each vertex i , it randomly selects some vertices j according to the noisy distribution $P_n(j)$ and treats (i, j) as negative edges. It used the noisy distribution in the field: $P_n(j)^{0.75}$, where d_j is the degree of the vertex j . Let M be the number of negative samples for each positive edge; the objective function can be redefined as:

$$O = \sum_{(i,j) \in E} w_{ij} (\log p(e_{ij} = 1) + \sum_{k=1}^M E_{jk \sim P_n(j)} \gamma \log(1 - p(e_{ijk} = 1))) \quad (3.6)$$

A simple approach to optimizing the above equation is to use the stochastic gradient descent, although it is problematic. This happens because when the edges (i, j) are sampled to update the model, the edge weight w_{ij} will be multiplied by the gradient. When the values of the weights are divergent (e.g., from 1 to thousands), the norms of the gradient are also divergent, and in this case it is very difficult to choose an appropriate learning rate. Therefore, LargeVis adopts the edge sampling approach proposed in [132]. It randomly samples edges with a probability proportional to their weights and then treats the sampled edges as binary edges. With this edge sampling technique, the objective function remains the same and the learning process will not be affected by the variance of the edge weights.

3.1.3. Differences between UMAP, SNE and LargeVis methods

As an aid to illustrate the similarities of UMAP with other neighbor embedding methods (t-SNE and LargeVis), we review the main equations used in these methods and then present the equivalent UMAP expressions. Other

Probabilities comparison

First, consider how the probabilities between two objects i and j are defined in the high-dimensional input space X and the low-dimensional space Y . These are normalized and symmetrized in various ways. In a typical implementation, these pairwise quantities are stored and manipulated as (potentially sparse) matrices. Quantities with the subscript ij are symmetric, that is, $v_{ij} = v_{ji}$. Extending it to the conditional probability notation used in SNE, $j|i$ indicates an asymmetric similarity, that is, $v_{j|i} \neq v_{i|j}$. In t-SNE, the probabilities in the input and embedding spaces (Eqs. 2.44, 2.45) can be computed for the k NN graph, where p_{ji} is set to zero for non-neighbor points in the

k NN graph. LargeVis uses the same p_{ij} probabilities as t-SNE but approximates the k NN to compute it very quickly and become more efficient. In LargeVis (described in detail in Chapter 4), the probability in the embedding space is:

$$q_{ij} := (1 + \|y_i - y_j\|_2^2)^{-1}. \quad (3.7)$$

Comparing Eqs. 2.22 and 2.39 show that UMAP, t-SNE, and LargeVis all use a Gaussian or RBF kernel for probabilities in the input space. Comparing Eqs. 2.25 and 2.38 show that UMAP and t-SNE/LargeVis use different approaches to symmetrize probabilities in the input space. Comparing Eqs. 2.26, 2.45, and 3.7 show that, in contrast to t-SNE, UMAP, and LargeVis do not normalize the probabilities in the embedding space by all pairs of points. This advantage makes UMAP much faster than t-SNE and also makes it more suitable for mini-batch optimization in deep learning [47]. Comparing Eqs. 2.26, 2.45, and 3.7 also show that all three methods use the Cauchy distribution for probabilities in the embedding space.

Cost functions comparison

LargeVis uses a similar approach to Barnes-Hut t-SNE when approximating p_{ij} , but further improves efficiency by only requiring approximate nearest neighbors for each point. For low-dimensional coordinates, it completely abandons the normalization of w_{ij} . Rather than using the Kullback-Leibler divergence, it optimizes a likelihood function, and hence is maximized, not minimized:

$$C_{LV} = \sum_{i \neq j} p_{ij} \ln(q_{ij}) + \gamma \sum_{i \neq j} \ln(1 - q_{ij}) \quad (3.8)$$

p_{ij} and q_{ij} are defined as in Barnes-Hut t-SNE (apart from the use of approximate nearest neighbors for p_{ij} , and the fact that, in implementation, LargeVis does not normalize p_{ij} by N) and γ is a positive constant chosen by the user that weights the strength of repulsive contributions (second term) relative to the attractive contribution (first term).

Ignoring the two constant terms, the UMAP cost function has a very similar form to that of LargeVis, but without a γ term to weight the repulsive component of the cost function and without requiring matrix-wise normalization in the high-dimensional space. The cost function for UMAP can be optimized (in this case, minimized) with stochastic gradient descent in the same way as LargeVis. Comparison of Eqs. 2.43, 3.8, and 2.27 show that UMAP, t-SNE, and LargeVis have similar but not exactly equal cost functions. The first term in all these cost functions is responsible for the attractive forces and the second term is for the repulsive forces; therefore, they can all be considered as neighbor embedding methods [14, 24].

As shown in Fig. 3.2, an experiment was conducted for different datasets in which the results were averaged to show that UMAP clearly leads the way when it comes to embedding time. It is almost an order of magnitude faster than its competitors. On the contrary, all methods have comparable memory occupancy, which is entirely dependent on the size of the dataset.

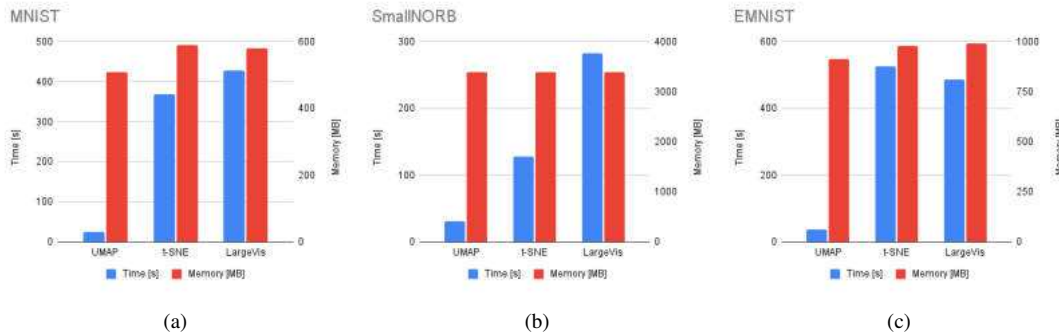


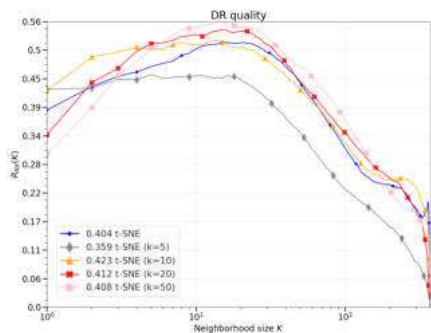
Figure 3.2: Time efficiency and memory load of UMAP methods compared to t-SNE and LargeVis.

3.2. t-SNE with Euclidean and binary distances

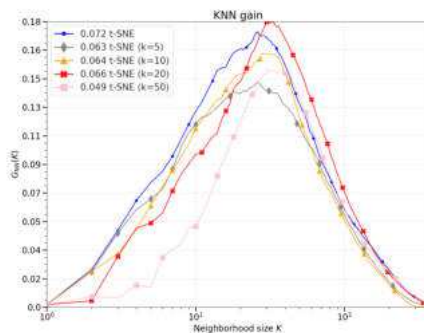
To evaluate whether the data visualization method can be perceived as embedding of an undirected graph, we create a variation of t-SNE that uses Euclidean and binary distances instead of the probability matrix p_{ij} . With this operation, we would be able to parameterize the t-SNE method by the number of nearest neighbors k instead of the perplexity. Other t-SNE calculations remain unchanged. All visualizations were created for a 10% subset of the MNIST dataset ($M=7 \cdot 10^3$).



(a) Original t-SNE. (b) Binary distances (k=10). (c) Binary distances (k=20). (d) Binary distances (k=50).



(e) DR quality.



(f) KNN gain.

Figure 3.3: Visualizations and metrics obtained for binary distances instead of probabilities. Parameterized by the number of nearest neighbors k .

In case of binary distance, we simply create a binary matrix. We insert 1 where we have determined the nearest neighbors and 0 otherwise. The same mechanism is used

for the Euclidean distances, but instead of 1 we insert the actual Euclidean distance into the *probability* matrix.

In both cases, we show that for $k=\{10, 20\}$, our t-SNE embedding variant achieves the same data reduction (DR) quality as the original visualization of t-SNE. It is verified in Figures 3.3 and 3.4, where *DR quality* is presented for binary and Euclidean distances. In both cases, we want the curves to at least overlap, which means that the visualization quality is comparable to the original t-SNE implementation. We can observe this effect for $k = 10$ using binary distances and for $k = 10$ and $k = 20$ for Euclidean distances. Intuitively, the more neighbors that are used, the better the quality of visualization, which is reflected by the increasingly higher position of the curves in both graphs (resulting in a higher AUC value).

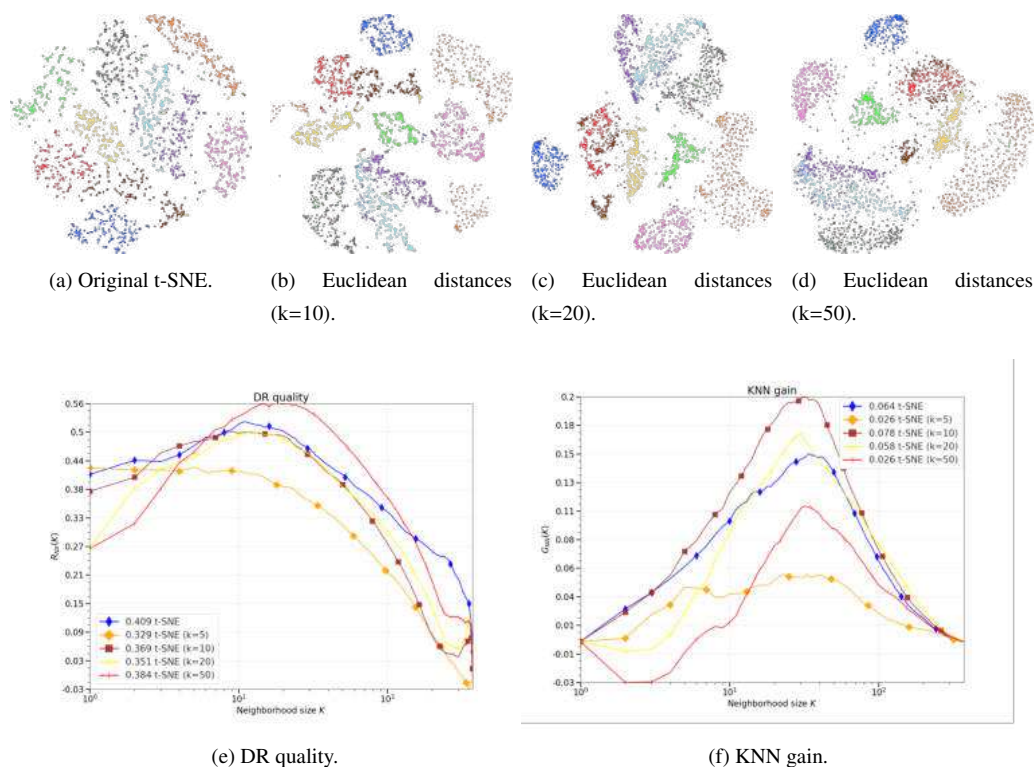


Figure 3.4: Visualizations and metrics obtained for Euclidean distances instead of probabilities. Parameterized by the number of nearest neighbors k .

This observation confirms that we are able to modify the t-SNE method step by step in a way that will ultimately result in the obtaining of the simplest IVHD (Interactive visualization of high-dimensional data) method described in the next Section. The next step would be to swap the part of the algorithm responsible for optimizing the Kullback-Leibler divergence. This is an important feature of data visualization algorithms that rely on neighbor embedding. Their cost function is defined in a way that defines two members. The first member is responsible for the forces of attraction and the second for the forces of repulsion. For this reason, we can think of them as methods described using a common mathematical apparatus imposed by UMAP.

3.3. Interactive visualization of high-dimensional data

As in previous chapters, we assume that each ND feature vector $x_i \in \mathbf{X} \subset \mathbb{R}^N$ is represented in a 2-D Euclidean space by a corresponding point $y_i \in \mathbf{Y} \subset \mathbb{R}^2$ and $i = \{1, \dots, M\}$. We define two distance matrices: $\mathbf{D} = \{\delta_{ij}\}_{M \times M}$ and $\mathbf{d} = \{d_{ij}\}_{M \times M}$, in the spaces *source* \mathbb{R}^N and *target* \mathbb{R}^2 , respectively.

The value of δ_{ij} is a measure of dissimilarity (proximity) between the feature vectors x_i and x_j while in the 2D Euclidean space: $d_{ij} = \sqrt{\|y_i - y_j\|}$. In general, the source space does not have to be Cartesian and can be solely defined by a proximity matrix \mathbf{D} between any two data objects of a (possibly) different data representation from the vector one (e.g., shapes, graphs, etc.).

Classical MDS performs the mapping as described in Eq. 2.8. For IVHD, we assume that $m = 2$ and $k = 1$. However, there are many other forms of this stress function, which are defined, for example, in [149, 138, 39]. One can easily adopt our approach to particular (k, m) parameters choice. However, finding the global minimum of this multidimensional and multimodal cost function (2.8) with respect to Y is not a trivial problem. To this end, we use the force-directed approach presented in [30, 104, 105, 106] and described in detail in a later section of this chapter.

We assume that the set of points \mathbf{Y} is treated as an ensemble of interacting particles y_i . Particles evolve in \mathbb{R}^2 space with discrete time t scaled by time step Δt , according to Newtonian equations of motion discretized using the *leapfrog* numerical scheme. We use their simplified form [32]:

$$\Delta y_i \leftarrow a \cdot \Delta y_i + b \cdot f_i, \quad (3.9)$$

$$f_i^t = -\nabla \left(\sum_{j=1}^M (\delta_{ij}^t - d_{ij}^t)^2 \right), \quad (3.10)$$

$$y_i \leftarrow y_i + \Delta y_i, \quad (3.11)$$

what resembles well-known momentum minimization method. Aforementioned Equations and differences to momentum method are described in details in Section 3.3.2. The value of $a \in [0, 1]$ represents friction and is equal to 1 in the absence of energy dissipation. Meanwhile, b parameterizes the forces between the particles. The proper balance of a and b is crucial for the convergence speed of the particle system to a stable and good (close to the global one) minimum of the stress function (Eq. 2.8). It is demonstrated in [30, 104, 105, 106] that this formulation of the classical MDS algorithm produces acceptable embeddings for low-dimensional ($N < 10$) and rather small datasets ($M \sim 10^3$) in sublinear time complexity, that is, similar to widely used stochastic gradient descent (SGD) algorithms and its clones employed, e.g., in the original implementation of t-SNE [86]. The main problems with MDS for $N > 10$ are both the effect of *curse of dimensionality*, which produces poor quality embeddings, and the high computational and storage complexity for $M \sim 10^{5+}$.

The linear time and memory complexity of the embedding of the data can be achieved by using only a limited number of distances from D and the corresponding

distances from d (such as in [30, 104, 105, 106, 61]). In the context of the so-called theory of structural rigidity [136], all the distances between the samples in a n -D dataset are not needed to retain the rigidity of its shape in a n -D space. The term *rigidity* can be understood as a property of a n -D structure made of rods (distances) and joints (data vectors) such that it does not bend or flex under an applied force. Therefore, to ensure the rigidity of X (and its 2-D embedding Y), only a fraction of the distances (joints) from D (and d) is required. What is the minimum number of distances for which the original X and target Y data sets remain rigid?

As shown in [136], a minimal n rigid structure, which consists of $M \geq n$ joints (vectors) in a n -dimensional space, requires at least:

$$L(n)_{\min} = n \cdot M - \frac{n \cdot (n + 1)}{2} \quad (3.12)$$

rigid rods (distance). This means that in the source space, the number of distances $L(N)_{\min}$ can ensure a lossless reconstruction of the data structure in N -D. In particular, in 2-D the structural rigidity can be preserved for $L(2)_{\min} = 2 \cdot M - 3$. However, L_{\min} defines only the lowest bound of L required to maintain structural rigidity. Therefore, IVHD answers the following questions:

1. What minimum number of distances in X should be known to obtain a reasonable reconstruction of the data in Y , simultaneously, preserving the structural rigidity of Y ? Is it closer to N or rather to 2?
2. What distances should be retained?

In general, \mathbf{D} cannot be an Euclidean matrix. It may represent the proximity of samples in an abstract space. In particular, samples $x_i \in \mathbf{X}$ can occupy a complicated n -D manifold for which $n \ll N$ (see Figure 3.1). Then we can assume that only the distances of each sample x_i from its nearest neighbors (nn) are Euclidean. As shown in [134, 32], the k -nearest neighbor graph (k NN graph), in which the vertices represent the data samples and the edges represent the connections to their nearest neighbors, can be treated as an approximation of this low-dimensional manifold, immersed in a high-dimensional feature space (see Figure 3.5). In the CDA and Isomap [149, 134] DR algorithms, the proximities of the more distant graph vertices (samples) are calculated as the lengths of the shortest paths between them in a respective k NN graph. However, although in this way we can more precisely approximate the real distances in the low-dimensional manifold, the full distance matrix D has to be calculated by the very time-consuming $O(M^3)$ Dijkstra algorithm (or Floyd-Warshall [37]) and, even more demanding, has to be stored in operational memory. Calculating the complete D is not necessary, and the problem of data embedding can be replaced by the congruent problem of visualization of the k NN graph.

Graph visualization (GV) and DR methodology share many common concepts. For example, the metaphor of the particle system and the force-directed method were used to minimize the cost function, which was introduced independently into GV and DR [30, 41]. In summary, as shown in Figure 3.1, the first step of the **IVHD** DR method consists of the construction of a k NN graph, which approximates the n non-Cartesian

dimensional manifold immersed in \mathbb{R}^N . Then, we can use the fast procedure for graph visualization in the 2-D space presented in [32].

Let us assume that $G(V, E)$ is the k NN graph for the high-dimensional dataset \mathbf{X} . The data samples $x_i \in \mathbf{X}$ correspond to the graph vertices $v_i \in V$, while E is the set of edges connecting each x_i with their k nearest neighbors. Because we have assumed (see Figure 3.1) that this graph is an approximation of the n -dimensional manifold, consequently, we assume that its topology-preserving embedding in 2-D Euclidean space will produce similar results as DE algorithms. According to the definition formulated in [126]:

Definition 9 *The topology is preserved if a connectivity algorithm, such as the k NN, can easily recover the edges of the input graph from the coordinates of the nodes after embedding.*

This means that unlike DR algorithms, which tend to retain the order of neighbors, the ordering of nearest neighbors (usually a few) is irrelevant in this case. The requirement of preserving the order of a small number of nearest neighbors nn for each $x_i \in \mathbf{X}$, which are subject to uncertainty and measurement errors, is secondary in terms of visualization of large data. The crucial problem in formulating hypotheses *ad hoc* and deciding on the use of particular machine learning tools in further data analysis is revealing the data topology, that is, its multiscale cluster structure.

3.3.1. Improving classical MDS

In addition to the assumptions of the previous chapter, we assume that by imposing a high contrast on these two types of distance, we will be able to preserve in \mathbf{Y} the multiscale cluster structure of \mathbf{X} , by employing MDS mapping only on those binary distances L . This way we could decrease both the computational complexity of data embedding to $O(a \cdot M)$ with $a \sim n_{vi}$ and its computational load to only $nn \cdot M$ integers.

To obtain the 2-D embedding \mathbf{Y} of $G(V, \mathbb{E})$, IVHD minimize the following stress function:

$$E(\|\mathbf{D} - \mathbf{d}\|) = \sum_i \sum_{j \in O_{nn}(i) \cup O_m(i)} b (\delta_{ij} - d_{ij})^2. \quad (3.13)$$

Consequently, the interparticle force f_i from Eq. (3.10) in $E(\cdot)$ minimization procedure simplifies to:

$$f_i^n = - \sum_{j \in O_{nn}(i)} y_{ij}^n - c \sum_{k \in O_m(i)} (1 - d_{ik}^n) \cdot \frac{y_{ik}^n}{d_{ik}^n}, \quad y_{ik}^n = y_i^n - y_k^n. \quad (3.14)$$

To explain this concept in terms of data embedding, let us use a toy example.

Assume that \mathbf{X} consists of $M = 38$ samples located in the vertices of two identical but translated and mutually rotated regular 18-dimensional hypertetrahedrons. In fact, the data create an unconnected graph where the vertices are joined by the hypertetrahedron edges. As shown in Figure 3.5a, by employing classical MDS with the cost function (2.8) and the data set defined by the complete distance matrix \mathbf{D} ($L = 703$

distances), we have obtained the embedding \mathbf{Y} shown in Figure 3.5a. One can observe that the result is not satisfactory. Although it shows the separability of the data well, the local structure of \mathbf{Y} remains very unclear. Now, assume that we know only $rn = 10$ distances from \mathbf{D} to random neighbors for every $x_i \in \mathbf{X}$ ($L = 300$ distance). As expected, the final embedding from Figure 3.5b shows that data separation is still visible, but lacks any fine-grained structure. As shown in Figure 3.5c, this flaw can be partially eliminated by also taking into account the distances to the nearest neighbors nn of x_i , that is, in this particular case $nn = 2$ and $rn = 1$ were established ($L = 103$ distances). Furthermore, we have reduced the strong bias caused by long-range interactions between the x_i samples and their random neighbors, by decreasing the scaling factor of the respective *forces* ($c = 0.1$ in Eq. (3.14)) 10 times. Unlike in Figure 3.5b, in Figure 3.5c we observe a much better reconstruction of the local structure Y , but at the cost of worse data separation. The embeddings of two regular hypertetrahedrons overlap each other.

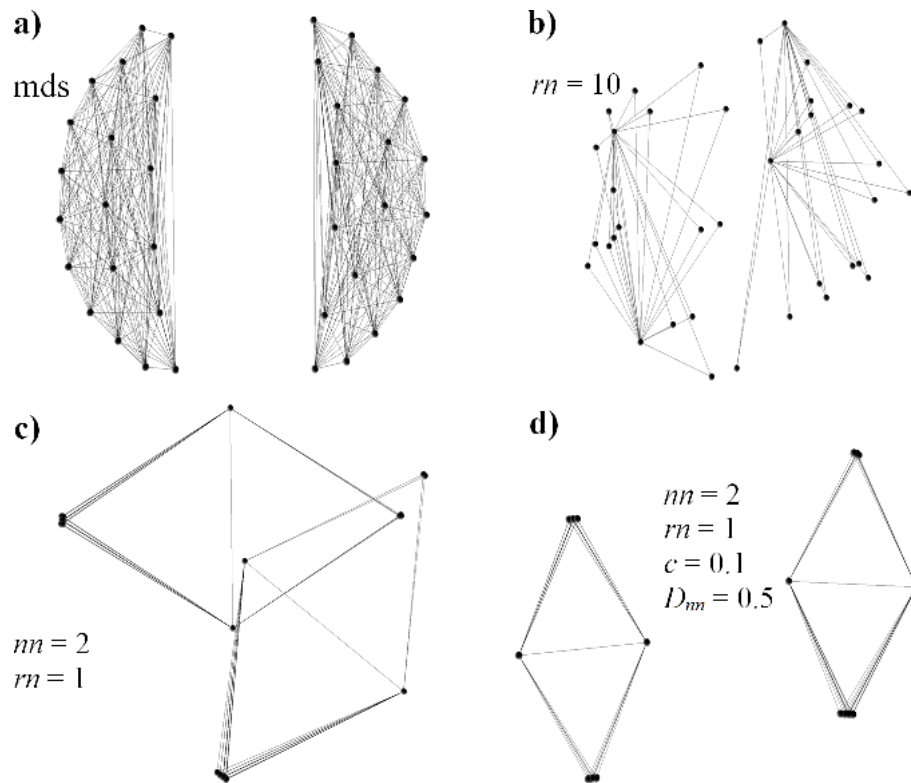


Figure 3.5: The results of 2-D embedding of two identical regular hypertetrahedrons for (a) original MDS setting and (b-d) highly reduced number of distances in the cost function (1). Source: [34].

Now, assume that all the distances between all the samples M are equal to 1. To increase the contrast between the nn nearest and rn random neighbors of y_i , we assume that $D_{nn} = 0.5$ and $D_{rn} = 1$. Since now we use binary distances $\{0.5, 1\}$, we do not need to store floating point matrix \mathbf{D} but just remember the nn indices (integers) of each y_i to their nearest neighbors (here 70 integers in total). As shown in Figure 3.5d, the 2-D embedding clearly reconstructs the local and global structure of the nn -graph

representing the original 18-dimensional data. Because M and N were small in this example, the value of $D_{nn} = 0.5$, was sufficient to properly contrast the distances nn and rn . However, as we show above (see Eq. (3.13)), for larger M and N , D_{nn} should be equal to 0 to reduce the effect of *curse of dimensionality*. It is worth mentioning here that the neighborhood relation is not symmetrical; thus, $(n = nn + rn) \cdot M$ can be smaller than $L(2)_{\min}$ for $n = 2$. Consequently, it may produce non-rigid and deformed (2-D) embeddings. Furthermore, assuming that $nn = 1$, we can obtain meaningless 2-D embeddings, although rn is much greater than 2. In summary, we can state the following conjectures.

1. The number of distances from \mathbf{D} , sufficient to obtain the 2-D embedding \mathbf{Y} of the original N -D dataset \mathbf{X} , which preserves the local and global properties of \mathbf{X} , can be surprisingly small and close to $\sim 3 \cdot M$.
2. For each $x_i \in \mathbf{X}$, its vicinity consisting of a few nearest neighbors nn should be preserved, while the global cluster structure is controlled by a few (often one) rn randomly selected neighbors.
3. The distances between $x_i \in \mathbf{X}$ and their nearest and random neighbors, and the respective forces in the optimization procedure, should be properly contrasted. For high-dimensional data, the binary distance can be used. This can drastically reduce the memory load from $M \cdot (M - 1)$ floating points (two distance arrays \mathbf{D} and \mathbf{d}) to $nn \cdot M$ integers, where $nn \sim 3$.

Many types of force-directed methods and algorithms have been used for both vector visualization and data embedding problems [50, 131]. Neglecting technical details, all of them are based on N-body dynamics, where the low-dimensional embedding of a graph (or a data set) is represented by an ensemble of interacting particles evolving in 2-D (or 3-D) Euclidean space. However, to adapt it directly to graph visualization, a precise definition of the graph in terms of a dissimilarity matrix is required. The stress function, similar to that of MDS, could then be applied as the cost function in graph visualization. However, the number and kind of distances that IVHD employs are radically different from those of the classical MDS algorithm.

In Fig. 3.6, we show the envelopes of the distance histograms \mathbf{D} and \mathbf{d} for the MNIST dataset before and after IVHD embedding. Although the MNIST dataset has a varied structure, the envelope of the \mathbf{D} histogram in linear coordinates (Fig.3.6a) is perfectly bell-shaped, while that of \mathbf{d} is more deformed, but still resembles the Cauchy distribution. To increase distance diversification (see Fig.3.6b), instead of all $M(M - 1)/2$ floating point distances, we can consider binary distances for only a few (nn) nearest neighbors and just one (rn) randomly selected neighbor. This is because most of the real distances (95%) from the nearest and random neighbors are located in separated and rather distant intervals (darker blue and red boxes in Fig.3.6b). For higher dimensions, the random neighbors are almost equidistant from y_i due to the "curse of dimensionality" effect. The overlap region (green) contains only 0.3% distances. Therefore, we can also assume that $O_{nn}(i) \cap O_{rn}(i) = \emptyset$. However, this assumption is superfluous because the probability of choosing the nearest neighbor as a random

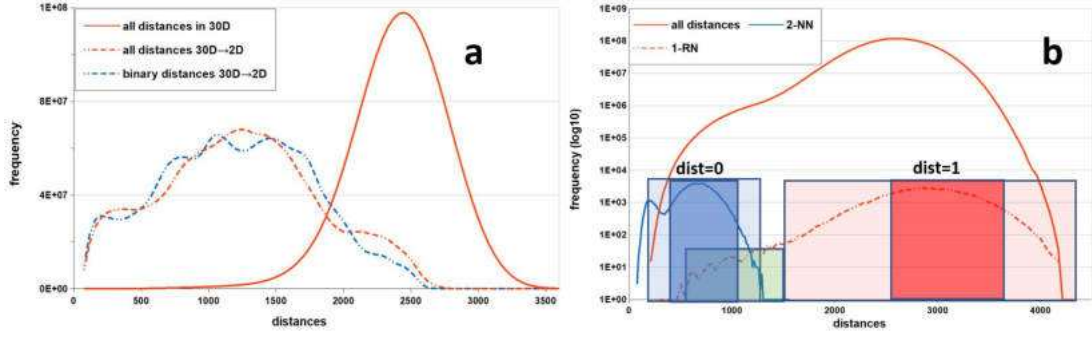


Figure 3.6: The envelopes of histograms for MNIST dataset (after PCA transformation $784D \rightarrow 30D$). *Red solid line*: all \mathbf{D} distances (a - linear, b - logarithmic scale); a) *dashed lines*: all \mathbf{d} distances; b) *blue solid line*: \mathbf{D} distances only between samples and their 2-NNs, and *red dashed line*: \mathbf{D} distances only between samples and one random neighbor. Source: [93].

neighbor is negligibly small for large N . As shown in Fig.3.6a, for non-binarized and binarized source distances, their histograms for respective 2D embeddings are very similar. Thus, let $O_{nn}(i)$ and $O_{rn}(i)$ will be the sets of indices of the nn nearest (connected) neighbors and the rn (unconnected) random neighbors of a feature vector y_i in k NN-graph, respectively. We define the binary dissimilarity measure as follows:

$$\forall x_i \in \mathbf{X} : D_{ij} = \begin{cases} 0 & \text{if } j \in O_{nn}(i) \\ 1 & \text{if } j \in O_{rn}(i) \end{cases}. \quad (3.15)$$

Thus, unlike the UMAP and t-SNE algorithms, IVHD is not interested in even an approximate ordering of k NN for each $x_i \in \mathbf{X}$. This is justified for small nn , because the distances to the first few NNs, in general, cannot differ too much (see the blue plot in Fig. 4.5b), and the ordering of NN can result from measurement errors. We assume that the number of neighbors nn must meet two conditions. First, the k NN-graph should be fully connected (or approximately, that is, the size of the largest component should be comparable to the size of the entire graph). Second, the k NN-graph augmented with approximately rn edges should be at least a minimal n -rigid graph (in 2D: 2-rigid). The lower band of the number of connections L , required to make the 2-rigid augmented k NN-graph, is $L \sim 2 \cdot M$. Meanwhile, the augmented k NN-graph has approximately $L \sim n_{vi} \cdot M$ edges, where $n_{vi} = nn + rn > 2$ [33]. As our experience shows, the probability that the largest connected component is rigid (or approximately rigid) is very high. In summary, to obtain the largest connected component approximately equal to the full k NN-graph, the number of nearest neighbors nn can be very low (mostly $nn = 2$, but for some specific datasets with very similar samples, it can be a bit larger). Assuming additionally that $rn = 1$, we can obtain a stable and rigid 2-D embedding of the k NN-graph.

This way, instead of the $O(M^2)$ floating point \mathbf{D} matrix, we have as input data $O(nn \cdot M)$ integers - the list of edges of k NN graph. The indices of rn random neighbors can be generated *ad hoc* during the embedding process. Thus, embedding high-dimensional data reduces to embedding of the corresponding sparse k NN graph. To this end, we minimize the following stress function:

$$E(\|\mathbf{D} - \mathbf{d}\|) = \sum_i \sum_{j \in O_m(i) \cup O_m(i)} \begin{cases} d_{ij}^2 & \text{if } j \in O_m(i) \\ c \cdot (1 - d_{ij})^2 & \text{if } j \in O_m(i) \end{cases}, \quad (3.16)$$

which represents the error between the dissimilarities $D_{ij} \in \{0, 1\}$ and the corresponding Euclidean distances d_{ij} , where: $i, j = \{1, \dots, M\}$, and $c \in (0, 1)$ is the scaling factor for random neighbors.

3.3.2. Optimization methods

Consider the continuous cost function f , such that $f : \Omega \subset \mathbb{R}^N \rightarrow \mathbb{R}$. Then, the general solution to the optimization problem is to find $x_{min} \in \Omega$ such that:

$$f(x_{min}) \leq f(x) \quad \forall x \in \Omega. \quad (3.17)$$

Stochastic Gradient Descent

The direction opposite to the direction of the gradient of the function f , ∇f , is equal to the direction of the steepest descent. Therefore, minimizing f is useful because it tells us how to change the argument to obtain a small improvement in the value of f . We can formulate this idea using an iterative algorithm that starts from a selected point x_0 and updates its value according to the formula:

$$x_i = x_{i-1} - \alpha \cdot \nabla f(x_{i-1}), \quad \alpha \in \mathbb{R}, \quad (3.18)$$

where α is *learning rate*. This is a gradient descent (GD) method of optimization. Stochastic gradient descent (SGD) is a probabilistic approximation of the above method. At each step, the algorithm calculates the gradient for one observation picked at random, rather than calculating the gradient for the entire dataset. It is much faster and more suitable for large-scale datasets.

Adadelta

Adadelta tries to improve *Gradient Descent*, by calculating *learning rate* in each iteration based on the previous values x and $\nabla f(x)$.

$$x_i = x_{i-1} - \frac{RMS(\Delta x_{i-1})}{\epsilon + RMS(\nabla f(x_{i-1}))} \nabla f(x_{i-1}) \quad (3.19)$$

By $RMS(x)$ we refer to the decaying root mean square of the time series of x [69].

Momentum

It is a heuristic algorithm. Its name derives from an analogy to physics. The matrix v can be interpreted as a matrix of velocities, the gradient as gravity, and β as a coefficient of friction.

$$v_i = \beta v_{i-1} + \alpha \nabla f(x_{i-1}) \quad (3.20)$$

$$x_i = x_{i-1} + v_i \quad (3.21)$$

Adam

The name *Adam* derives from the phrase *adaptive moments* [52]. It is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. The iteration of *Adam* consists of five steps:

$$v_i = \gamma_v v_{i-1} + (1 - \gamma_v) \nabla f(x_{i-1}) \quad (3.22)$$

$$s_i = \gamma_s s_{i-1} + (1 - \gamma_s) (\nabla f(x_{i-1}) \cdot \nabla f(x_{i-1})) \quad (3.23)$$

$$\hat{v}_i = \frac{v_i}{(1 - \gamma_v)^{i-1}} \quad (3.24)$$

$$\hat{s}_i = \frac{s_i}{(1 - \gamma_s)^{i-1}} \quad (3.25)$$

$$x_i = x_{i-1} - \alpha \frac{\hat{v}_i}{\epsilon + \sqrt{\hat{s}_i}} \quad (3.26)$$

It can be seen as a combination of the *RMSProp* and *Momentum* algorithms and is known for its computational efficiency, low memory requirements, and invariant to diagonal gradient rescaling.

Nesterov

Algorithm that is very similar to the *Momentum* (in fact it is sometimes called *Nesterov Momentum*). The only difference between the two is that *Nesterov* uses a gradient at the *future* position.

$$v_i = \beta v_{i-1} + \alpha \nabla f(x_{i-1} + \beta v_{i-1}) \quad (3.27)$$

$$x_i = x_{i-1} + v_i \quad (3.28)$$

3.3.3. Force directed

To find the minimum of $E(\cdot)$ IVHD uses mainly the force-directed approach in the form of the interacting particle method presented in [31] (although it can also utilize state-of-the-art optimization methods). The set of vertices v_i in \mathbb{R}^2 is treated as an ensemble of interacting particles. Particles evolve in \mathbb{R}^2 space with time according to Newtonian equations of motion starting from a random setup. Newtonian equations are discretized using the leap-frog numerical scheme with a timestep length equal to Δt . The force $f_i^n = -\nabla E_i$ coming from the particles corresponding to the vertices of the neighborhoods $O_{nn}(i)$ and $O_m(i)$ is calculated for each particle i at each time step n . Furthermore, the friction force proportional to the velocity of the particles $-\lambda v_n^i$ and acting against the direction of movement of the particles is used to dissipate the energy of the particle system. The following equations present the discretized version of the

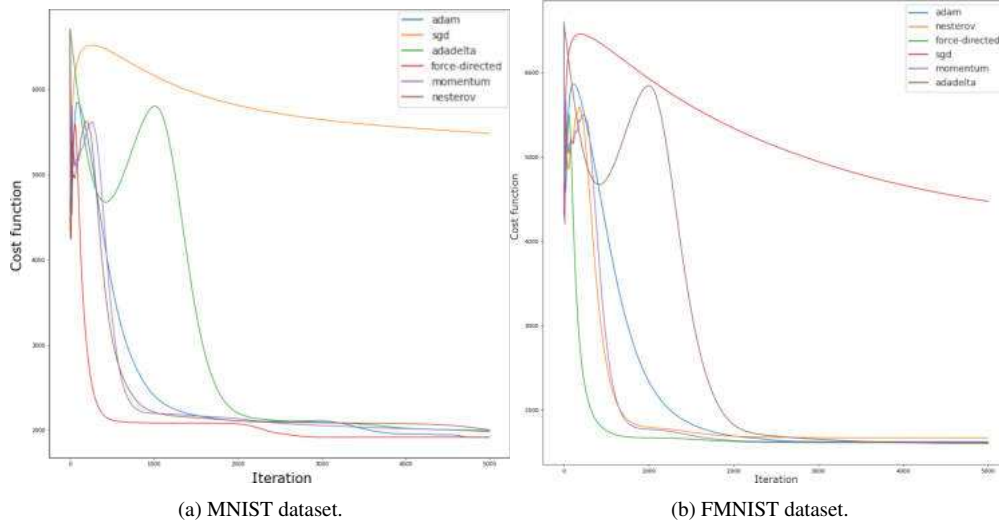


Figure 3.7: Speed of convergence to solution (measured as the value of the cost function in successive iterations) for different optimization methods used in IVHD.

Newtonian equations, which is used to calculate the particle velocities v_i^n and their positions r_i^n :

$$v_i^{n+\frac{1}{2}} = v_i^{n-\frac{1}{2}} + (2k_{mn}f_i^n - \lambda v_i^n) \cdot \Delta t \quad (3.29)$$

$$r_i^{n+1} = r_i^n + v_i^{n+\frac{1}{2}} \cdot \Delta t \quad (3.30)$$

$$v_i^n = \frac{v_i^{n+\frac{1}{2}} + v_i^{n-\frac{1}{2}}}{2} \quad (3.31)$$

$$f_i^n = - \sum_{j \in O_{mn}(i)} r_{ij}^n - c \sum_{k \in O_m(i)} (1 - d_{ik}^n) \cdot \frac{r_{ik}^n}{d_{ik}^n}, \quad r_{ik}^n = r_i^n - r_k^n \quad (3.32)$$

These equations can be simplified by the following substitutions:

$$a \leftarrow \frac{1 - \frac{\lambda \Delta t}{2}}{1 + \frac{\lambda \Delta t}{2}}, \quad b \leftarrow \frac{2k_{mn} \Delta t^2}{1 + \frac{\lambda \Delta t}{2}}, \quad \Delta r_i^{n+1} = r_i^{n+1} - r_i^n \quad (3.33)$$

Finally, we obtain the following scheme:

$$v_i = \Delta r_i \leftarrow a \cdot \Delta r_i + b \cdot f_i = a \cdot v_i + b \cdot f_i \quad (3.34)$$

$$r_i \leftarrow r_i + \Delta r_i \quad (3.35)$$

The value of $a \in [0, 1]$, assuming slow dissipation, should be close to 1. At the beginning of embedding, it can be slightly greater for faster dissipation of large initial fluctuations. Consequently, the value of c should be very small, not to overwhelm the influence of the forces that come from the closest neighborhood $O_{mn}(i)$. By comparing Eqs. 3.34 and 3.20, one can conclude that the methods are identical, but what is

important is that force-directed imposes additional constraints for a and b parameters derived from 3.33:

$$\frac{a}{b} = \frac{1 - \frac{\lambda\Delta t}{2}}{2k_{nn}\Delta t} \quad (3.36)$$

To ensure simulation speed and system convergence to the stable solution, we used the following self-adaption scheme to determine the value of parameter b :

$$\begin{aligned} \text{if } \text{abs}(\Delta T = \sum_{i=1}^N [|\Delta r_i^{n+1}|^2 - |\Delta r_i^n|^2]) > \tau \\ \text{then } b \leftarrow (\gamma = \begin{cases} \gamma_1 > 1 & \text{if } \Delta T < -\tau \\ \gamma_2 < 1 & \text{if } \Delta T > \tau \end{cases}) \cdot b \wedge r_i^{n+1} = r_i^n \end{aligned} \quad (3.37)$$

The quality of visualization depends on the relationship between $rn(i)$ and $nn(i)$ and the value of the parameter c (Eq. 3.32) i.e. the factors influencing the balance between the stretching and contracting forces. In the implementation of the method, for simplicity, we define constant input values of nn and rn as randomly chosen numbers of connected and disconnected vertices for each i , respectively. We assume that a neighbor connected to i cannot be its random neighbor. If the number of connected neighbors is less than nn , all of them are selected. The correct choice of nn , rn and c is crucial for the quality of 2D reconstruction and strongly depends on the type and size of visualized dataset. Experiments show that nn should be greater than or, at most, equal to rn [93]. Otherwise, long-range interactions start to dominate, producing a sphere of particles in 2D.

3.3.4. Binary and Euclidean distances comparison

As shown in Equation 3.15 - IVHD omits ordering nn for each $y_i \in \mathbf{Y}$ to obtain a good approximation of the data manifold. Thus, for simplicity, the distances between the connected vertices of the nn -graph should be the same and as close to 0 as possible. Furthermore, the nn -graph should be enhanced with additional edges (*random neighbors*) to ensure rigidity. In this section, we compare the results obtained by embedding with IVHD using Euclidean and binary distances for our calculations.

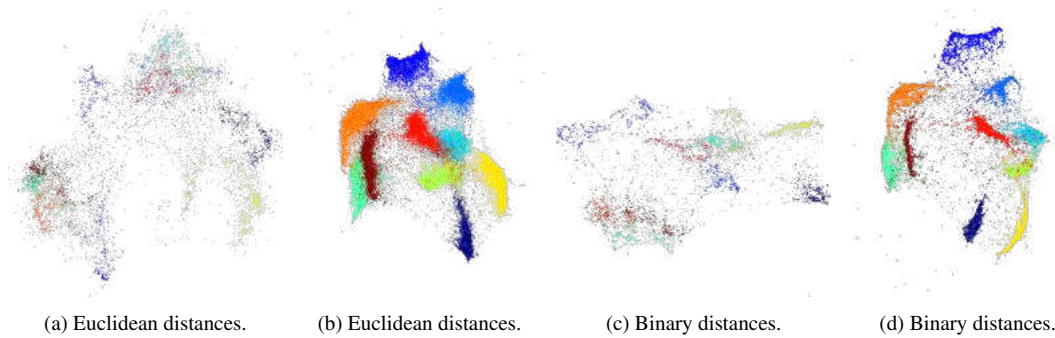


Figure 3.8: Visualizations for MNIST dataset and its sub-set ($M=7 \cdot 10^3$), for both euclidean and binary distances.

We compare the Euclidean and binary distances for the subset of the MNIST dataset ($7k$ samples) and for MNIST after PCA to 100D. The procedure to obtain the appro-

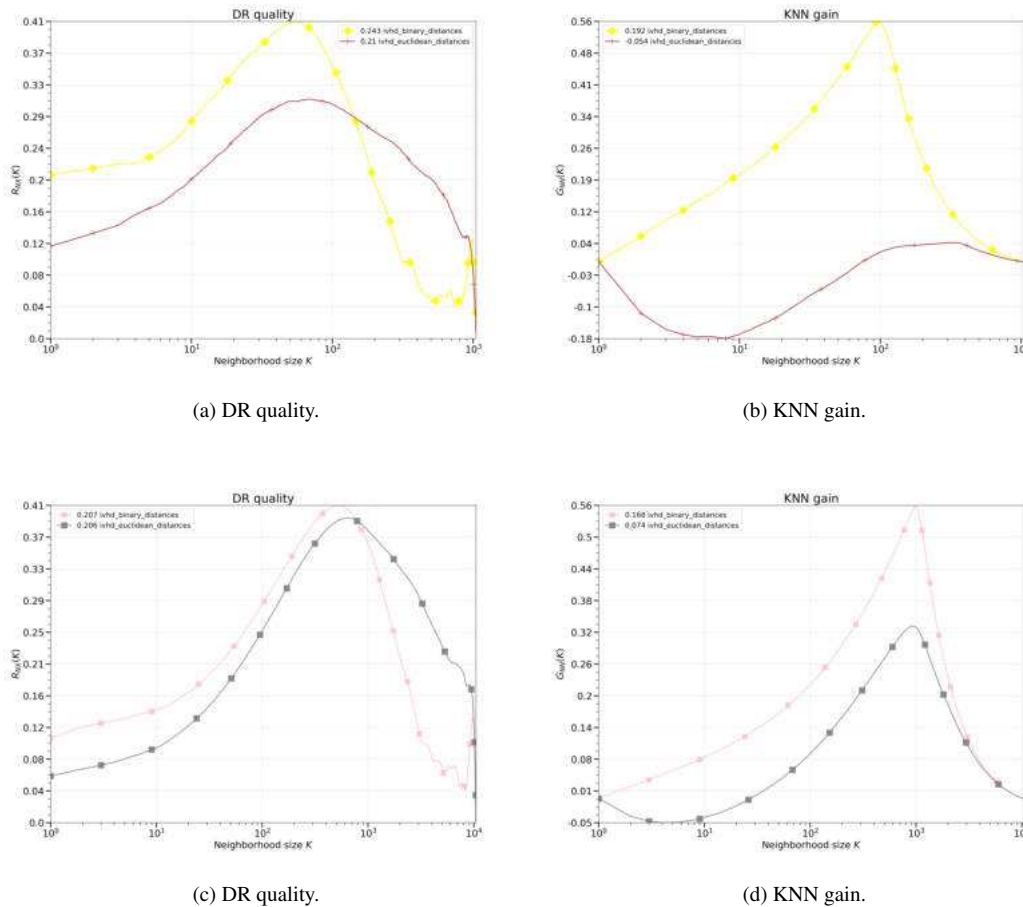


Figure 3.9: Metrics obtained for MNIST dataset and its sub-set ($M=7 \cdot 10^3$), for both euclidean and binary distances.

In the visualizations in Fig. 3.8, we see that for binary distances, IVHD generates visualizations in which the clusters are more compact. This is intuitive, because for the case of normal distances, the interactions between clusters would be more balanced, resulting in more noise between classes. In addition, this fact is confirmed by the metrics. Both DR quality and KNN gain reach higher values in most of the spectrum of the neighborhood under consideration. For these reasons, we only consider binary distances in further experiments, since IVHD then works more efficiently and better quality visualizations are achieved.

3.4. Improvements in IVHD algorithm

As shown in the next chapter, IVHD outperforms, in terms of computational time, the state-of-the-art DR algorithms by more than one order of magnitude in the standard DR benchmark data. The data embeddings obtained by IVHD are also very effective in reconstructing data separation in large and high-dimensional datasets. This is the principal requirement for knowledge extraction, because only multi-scale clusters of

data represent basic data *granules of knowledge*. Due to the simplicity of the IVHD algorithm, which is, in fact, a clone of the classical MDS algorithm, its efficiency could be further increased by implementing its parallel version in a GPU environment [93] (described in Section 5). It should be mentioned that IVHD produces less impressive results than SNE clones in reconstructing the local neighborhood. The *crowding problem*, similar to that in SNE, causes most of the data points in 2-D embeddings to gather in the center of the clusters.

However, an accurate reconstruction of nearest neighbors is a secondary requirement in interactive visualization of big data. The exact nn lists (or their approximates) are well known prior to data embedding. They can be visually presented at any time, for example, as the data points are arranged in \mathbf{Y} . Furthermore, the exact lists of the nearest neighbors are not reliable in the context of both measurement errors and the *curse of dimensionality's* principle. If a more accurate reconstruction of the location of the data is required, IVHD can be used to generate the initial configuration for DR algorithms based on a much slower SNE. Unlike SNE competitors, IVHD has only three free parameters that must be adapted to the data: nn , rn , and c . Because the method is fast, they can be easily matched interactively, although the *universal set*, that is, $nn = 3$, $rn = 1$, and $c = 0.1$ (or $c = 0.01$), works very well for most data (and networks).

Knowing all of these does not mean that further improvements cannot be made to the IVHD method. In this chapter, we will provide a description of the mechanisms that improve the quality of IVHD embeddings. This will be done mainly by modifying the metrics used and the interactions between neighbors at a certain step of the simulation.

3.4.1. L1 norm in late stages of embedding

Using the ℓ_1 norm in the late stages of embedding provides a suction effect, which pulls outliers closer to the centers of the clusters. The number of steps performed with the ℓ_1 metric cannot be too large, because the embedding would collapse in the centers of the clusters, completely distorting the global embedding structure.

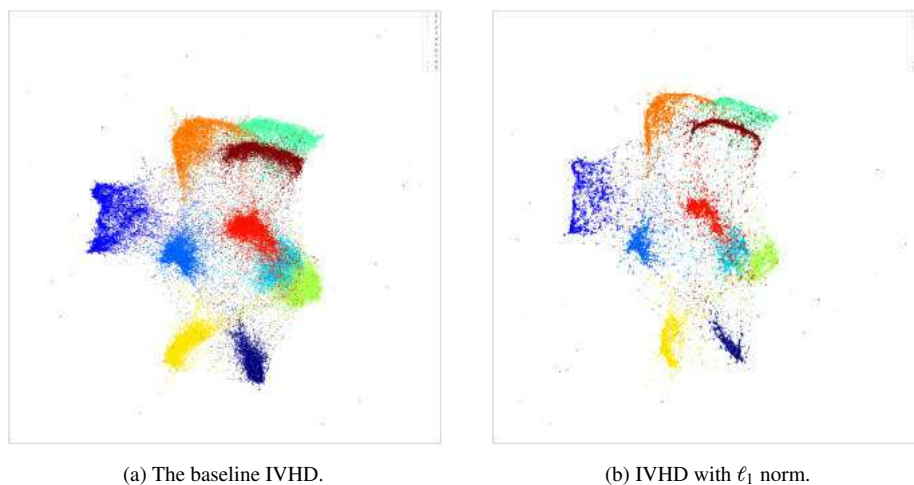


Figure 3.10: IVHD comparison on MNIST dataset, when ℓ_1 norm is used for last 50 steps of embedding procedure.

Both in Figures 3.10 and 3.11 *suction effect* are clearly visible. Most of the noise is moved to the center of the clusters from the space between. Additionally, the global structure is still maintained and is not distorted (the mutual positioning of classes in relation to each other does not change). This is reflected in the metrics shown in Fig. 3.14.

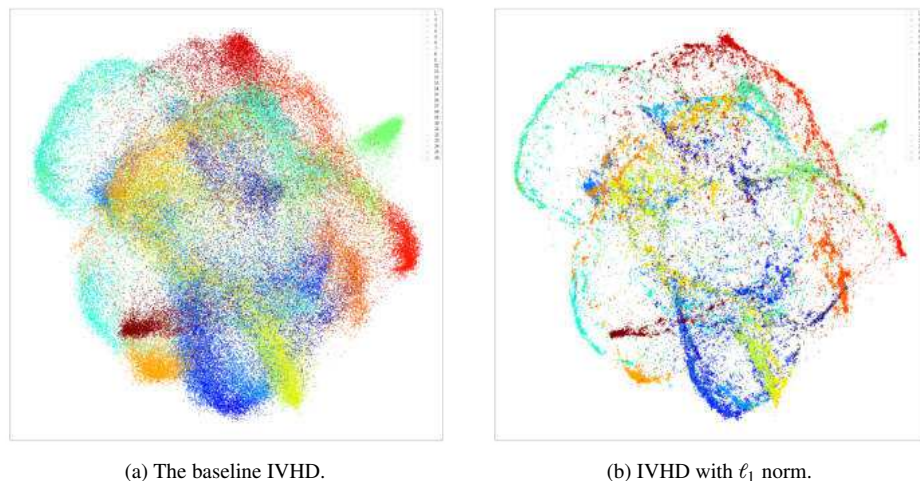


Figure 3.11: IVHD comparison on EMNIST dataset, when ℓ_1 norm is used for last 50 steps of embedding procedure.

3.4.2. Reverse neighbors mechanism

Let \mathbf{X} be a high-dimensional dataset, and let q be a point in this high-dimensional space. A reverse nearest neighbor (RNN) query retrieves all points $p \in X$ that have q as their nearest neighbor. The set $RNN(q)$ of the reverse nearest neighbors of q is called the influence set of q . Specifically, $RkNN(q) = \{p \in P \mid dist(p, q) \leq dist(p, p_k), \text{ where } p_k \text{ is the } k\text{-th nearest neighbor to } p\}$. Figure 3.12 shows four 2D points, where each point p is associated with a circle that covers its two nearest neighbors. For example, the two NNs of p_4 (p_2, p_3) are in the circle centered at p_4 . Consequently, $p_4 \in R2NN(p_2)$ and $p_4 \in R2NN(p_3)$. Let $kNN(p)$ be the set of k nearest neighbors to the point p . It is important to note that $p \in kNN(q)$ does not necessarily imply $p \in RkNN(q)$ and vice versa. For example, $2NN(p_4) = \{p_2, p_3\}$, while $R2NN(p_4) = \emptyset$ (that is, p_4 is not contained in the circles of p_1, p_2 or p_3).

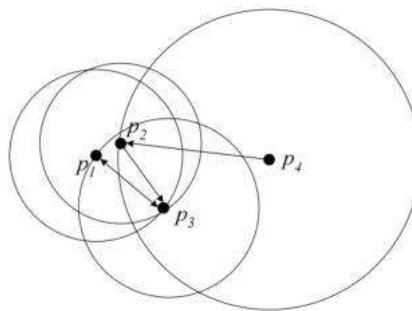


Figure 3.12: 2NN and R2NN examples. Source: [133].

From a technical point of view, the RNN mechanism employs a helper nearest neighbors graph, which contains a higher number of neighbors (default: $4 \cdot nn$, where nn is the number of *nearest neighbors* in the *calculation graph*). Using the procedure

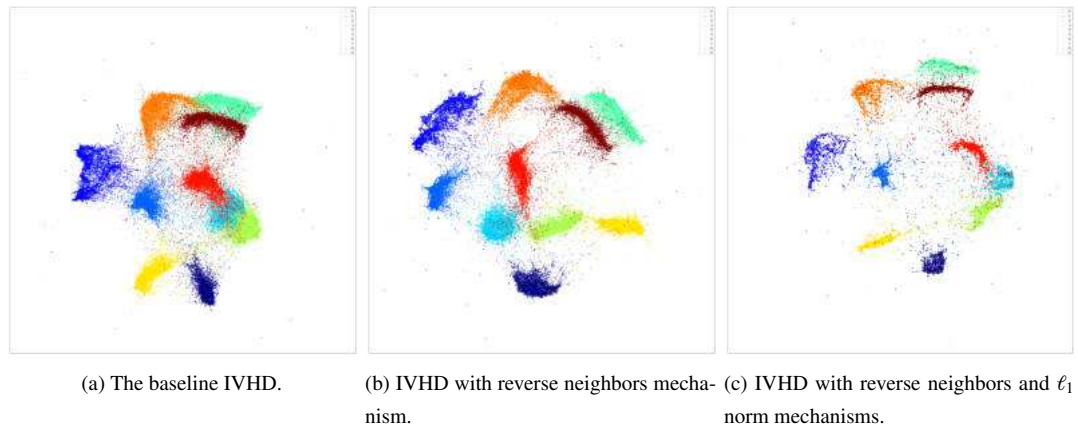


Figure 3.13: IVHD comparison on MNIST dataset, when reverse neighbors mechanism is used for last 200 steps of embedding procedure and when it is combined with ℓ_1 norm.

described in the first paragraph of this section and a helper graph, we determine the reverse nearest neighbors, and we retain only interactions in which the particles are their mutual neighbors or one is the reverse neighbor of the other. In this way, we decrease the number of neighbors (and, as a result, interactions) in the embedding *calculation*, causing the particles to interact with even more specific subsets of particles in the final stage of embedding. As shown in both Figs. 3.13 and 3.15, the reverse neighbors mechanism used for the last 200 embedding steps reduces the amount of noise as well (but in a different manner, than the norm ℓ_1). In the MNIST dataset (Fig. 3.13) the effect is slightly more noticeable.

We combined both mechanisms to provide the most generic way to deal with the noise that remains in the embedding (Figs. 3.13c and 3.15c). As the metrics in Fig. 3.14 indicate, the combined methods improve the quality of the DR and the kNN gain of the embedding. Changes do not distort the global properties of the visualization, which is confirmed by the proportional increase in both charts. Furthermore, in both cases, the best results were obtained by combining the norm ℓ_1 and the reverse neighbors mechanism. Importantly, the improvements do not introduce a large computational overhead. The helper graph is created simultaneously (or read from cache) with the main graph, which is used to calculate interactions between particles. The main operation that causes overhead is the search for reverse neighbors based on the two graphs, but the time of this procedure is negligible compared to the time of the entire embedding. It is worth mentioning that all the improvements added (and tested) to the IVHD method had a crucial assumption, which was that no additional and significantly higher computational load could be introduced into the algorithm.

In summary, the application of the above-described enhancements makes IVHD even more beneficial in terms of data reduction quality of created visualizations. Furthermore, the results have become more comparable to the baseline methods (Chapter 4.4) with no additional computational load.

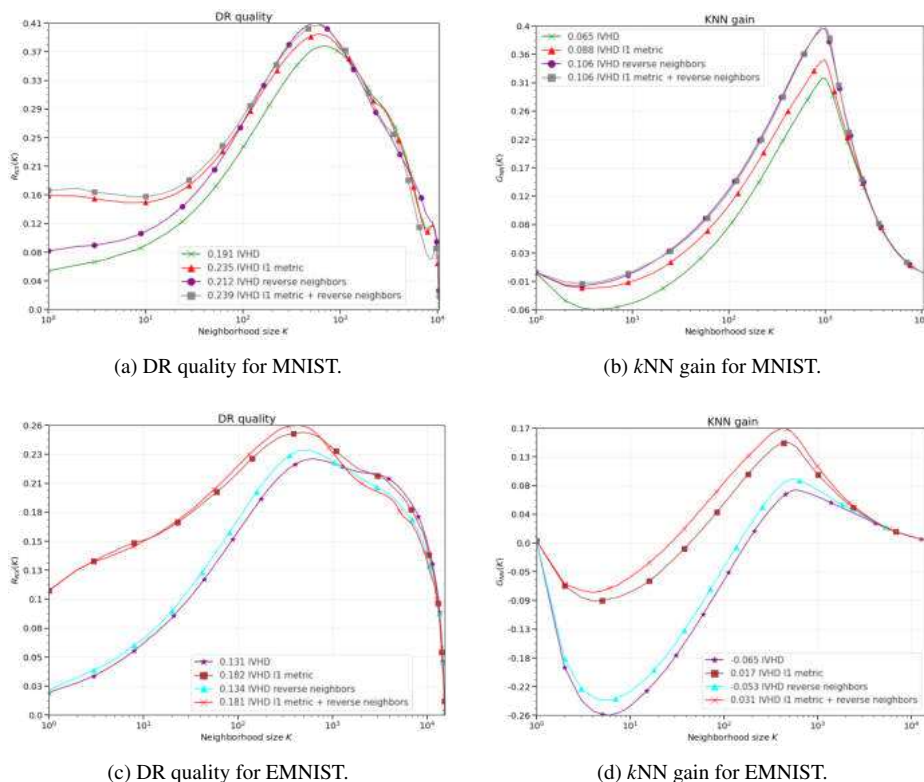


Figure 3.14: Metrics obtained for both MNIST and EMNIST dataset. It contains: 1) the baseline IVHD visualization, 2) employed ℓ_1 norm mechanism, 3) employed reverse neighbors mechanism and 4) combined ℓ_1 norm and reverse neighbors mechanisms.

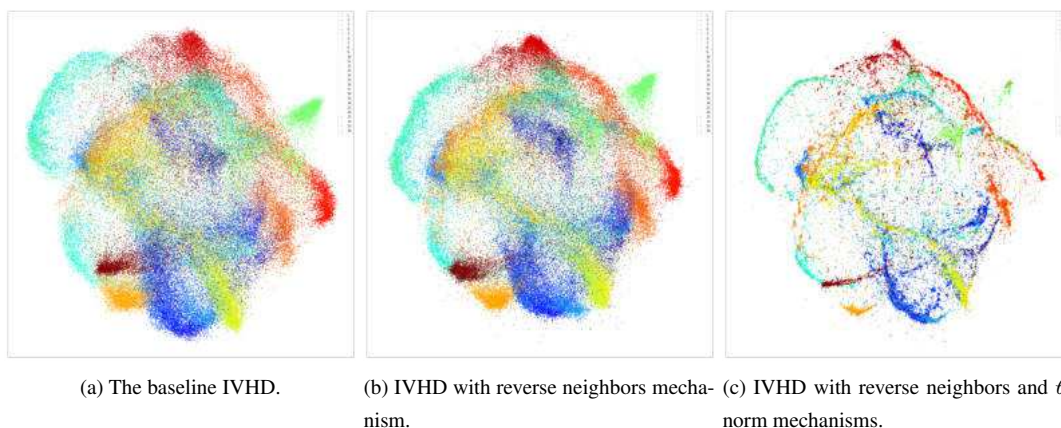


Figure 3.15: IVHD comparison on EMNIST dataset, when reverse neighbors mechanism is used for last 200 steps of embedding procedure and when it is combined with ℓ_1 norm.

CHAPTER 4

EXPERIMENTAL SETUP

The experimental setup consists of two key components: 1) definition of mutual quality assessment framework and 2) common architectures for running experiments. Multiple methods were implemented in the VisKit visualization library[92], which consists of the C++ and Python APIs. The components of the library are closely interrelated, but it is possible to use them independently. This was done to create the possibility to implement new visualization methods into this framework, without any additional requirements. This framework allows creating a stable environment for: 1) testing and evaluating embedding and optimization methods, 2) adjusting the parameters used by different methods, 3) formulate and instantly verify a number of hypotheses, and 4) integrate the methods in one visualization framework allowing for multiscale analysis of data, employing more accurate methods for finer scales.

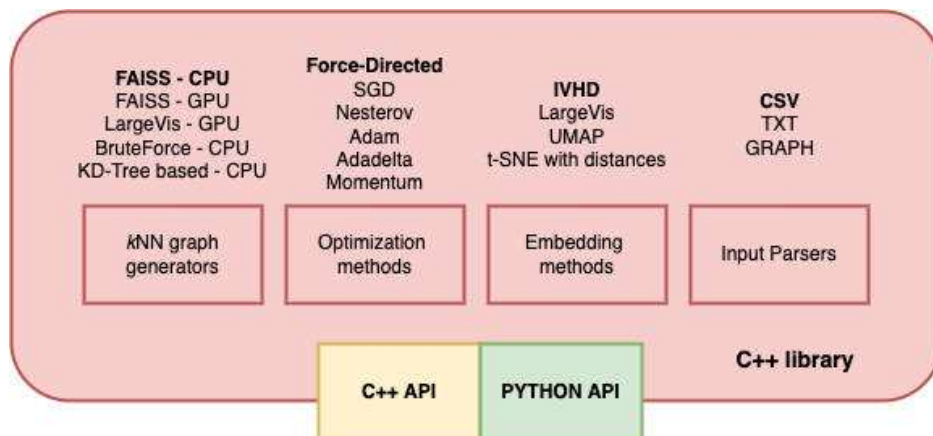


Figure 4.1: Viskit library architecture. It consists of multiple visualization methods, that can be freely modified (in terms of implementation) and tested. It also contains different optimization variants, that can be used in IVHD.

4.1. Quality assessment of dimensionality reduction

In practice, neither intrinsic dimensionality nor topological properties can be easily identified starting from a set of points. Therefore, the goal of non-linear dimensionality reduction (NLDR) is most often to preserve the structure of the dataset, as indicated, for example, by some kind of neighborhood relations such as proximity or similarity. The research community has focused on the design of new NLDR methods, and the question of a unified quality assessment framework remains largely unanswered. Since most NLDR methods optimize a given objective function, a simplified way to assess quality is to look at the value of the objective function after convergence. Of course, this allows for the comparison of several runs with, for example, different parameter values, but makes the comparison of different methods unfair. Another obvious criterion is the reconstruction error. The first purpose of this chapter is to review some SOTA-based and ranking-based criteria.

Table 4.1: Symbols used to define error measures.

$C_k(x_i)$	the set of k data vectors that are closest to x_i in the original data space
$\hat{C}_k(x_i)$	the set of k data vectors that are closest to x_i after projection
$U_k(x_i)$	the set of data vectors x_j for which $x_j \in \hat{C}_k(x_i) \wedge x_j \notin C_k(x_i)$ holds
$V_k(x_i)$	the set of data vectors x_j for which $x_j \notin \hat{C}_k(x_i) \wedge x_j \in C_k(x_i)$ holds
$r(x_i, x_j), i \neq j$	the rank of x_j when the data vectors are ordered based on their Euclidean distance from the data vector x_i in the original data space
$\hat{r}(x_i, x_j), i \neq j$	the rank of x_j when the data vectors are ordered based on their distance from the data vector x_i after projection

Trustworthiness and continuity

In principle, errors could simply be measured as the average number of data items that enter or leave the neighborhoods in the projection. Using slightly more informative measures: The trustworthiness of the neighborhoods [139] is quantified by measuring how far from the original neighborhood the new data points entering a neighborhood come. Distances are measured as rank orders; similar results have also been obtained with Euclidean distances. The trustworthiness of the projected result is defined as:

$$M_1(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^N \sum_{x_j \in U_k(x_i)} (r(x_i, x_j) - k), \quad (4.1)$$

where the term before the summation scales the values of the measure between zero and one. Preservation of the original neighborhoods (continuity) is measured by the following:

$$M_2(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^N \sum_{x_j \in V_k(x_i)} (\hat{r}(x_i, x_j) - k), \quad (4.2)$$

When interpreting the visualizations, it is particularly important that small neighborhoods, which have a small value of k , are trustworthy. Although preservation of the original neighborhoods is not as important as the trustworthiness of the projection, it gives insight into the trade-offs made in the projection.

Shepard and Co-rank diagrams

In the general situation of non-metric scaling, using the Shepard-Kruskal approach, we have data $Y = \{y_i, \dots, y_N\}$ and a model $f_i(\theta)$ with a certain number of free parameters θ . Often this is a non-metric multivariate scaling model in which the model values are dissimilarities. However, linear models and models with an inner product can and have been treated in the same way. We want to choose the parameters so that the rank order of the model best approximates the rank order of the data. To do this, we construct a loss function of the form:

$$\sigma(\theta, \hat{y}) = \sum_{i=1}^N w_i (\hat{y}_i - f_i(\theta))^2 \quad (4.3)$$

After we have found the minimum, we can make a scatter plot with the data Y on the horizontal axis and the model values f on the vertical axis. This is what we would also do in a linear or non-linear regression analysis. In summary, *Shepard diagram* shows the change in distances after dimensionality reduction for each pair of points. Similarly, we can define *Co-rank diagram*, which shows the change in mutual distance ranks after dimensionality reduction for each pair of points. The fitness of the co-rank diagram to the identity function is measured with a R^2 score. In a perfect scenario, the co-rank diagram should depict an identity function, meaning that the relative distance ranks from original dimensionality are accurately preserved in the space of reduced dimensionality.

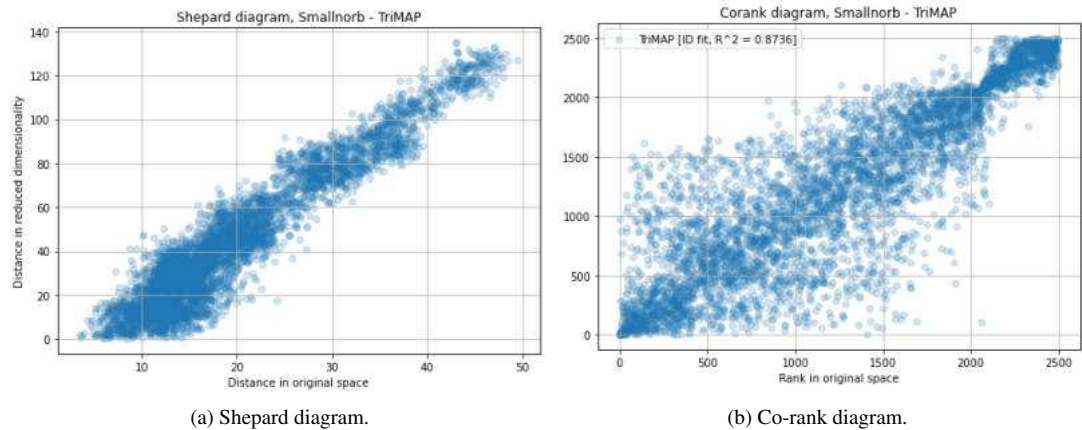


Figure 4.2: Example diagrams obtained for Trimap[2] visualization of smallNORB dataset (described in details in Appendix A).

Neighbor hit

Because of the high computational load required to calculate the precision / recall coefficients, to compare data separability and class purity, we define the following sim-

ple metrics, which utilize the mechanisms of neighbor hit (NH) and original neighbor hit (ONH):

$$cf_{nn} = \frac{\sum_{i=1}^M nn(i)}{nn \cdot M} \quad \text{and} \quad cf = \frac{\sum_{nn=1}^{nn_{\max}} cf_{nn}}{nn_{\max}}, \quad (4.4)$$

where $nn(i)$ is the number of nearest neighbors of y_i in the space Y , which belong to the same class as x_i . The value of cf is calculated for an arbitrarily defined value of nn_{\max} depending on the number of feature vectors in the class. To reflect a wide range of embedding properties, we use $nn_{\max}=100$. The value of $cf \sim 1$ for well separated and pure classes, while $cf \sim 1/K$ for random points in the K class. These simple metrics allow us to assess the quality of the embeddings by calculating several cf_{nn} values for a small, medium and greater number of nn . Differences in this criterion for confronting methods allow inferring their embedding quality for a very local ($nn=2$), local ($nn=10$) and medium ($nn=100$) reconstruction depth. The stability of cf_{nn} to increase nn means a more compact and circular shape of the classes. For the elongated and mixed classes, the values of cf_{nn} decrease faster with nn .

Rank-based criteria

Furthermore, we use state-of-the-art quality criteria [78, 76] for unsupervised DR methods, measuring the preservation of the high-dimensional neighborhood in the low-dimensional space. The general consensus that emerges from many publications [38, 17, 78, 140] is to use the average agreement rate between k -ary neighborhoods in high and low dimensions. The rank of x_j with respect to x_i in a high-dimensional space is written as $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leq k < j \leq N)\}|$, where $|A|$ denotes the cardinality of the set A . Similarly, the rank of y_j relative to y_i in the low-dimensional space is $r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq N)\}|$. Now, let $\mathbf{v}_i^k = \{j : 1 \leq \rho_{ij} \leq k\}$ and $\mathbf{n}_i^k = \{j : 1 \leq r_{ij} \leq k\}$ denote the sets of nearest neighbors k of x_i and y_i in the high-dimensional and low-dimensional space, respectively. $Q_{NX}(k)$ measures their averaged normalized agreement (also called *neighborhood preservation*):

$$Q_{NX}(k) = \frac{1}{kN} \sum_{i=1}^N |\mathbf{v}_i^k \cap \mathbf{n}_i^k| \in [0, 1] \quad (4.5)$$

It varies between 0 and 1, with 0 being an empty intersection and 1 being perfect agreement. Knowing that the random coordinates in \mathbf{Y} on average lead to $Q_{NX}(k) = \frac{k}{k-1}$, the useful range of $Q_{NX}(k)$ is $N - 1 - k$, which depends on k . Therefore, to compare or combine the values of $Q_{NX}(k)$ fairly for different neighborhood sizes, the criterion can be rescaled, as in [77]:

$$R_{NX}(k) = \frac{(N - 1)Q_{NX}(k) - k}{N - 1 - k} \quad (4.6)$$

This modified criterion indicates an improvement over a random embedding and has the same useful range between 0 (random) and 1 (perfect) for all k . $R_{NX}(k)$ is shown, with a logarithmic scale for k . This choice is justified by the fact that the size k and the radius R of small neighborhoods with uniform density in a P -dimensional

space are (locally) related by $k \propto R^P$. A logarithmic axis also reflects that errors in large neighborhoods are proportionally less important than in small ones. Eventually, a scalar score is obtained by computing the area under the $R_{NX}(k)$ curve in the logarithmic graph given by:

$$AUC(R_{NX}(k)) = \frac{\sum_{k=1}^{N-2} \frac{R_{NX}(k)}{k}}{\sum_{k=1}^{N-2} \frac{1}{k}}. \quad (4.7)$$

The AUC supposedly assesses the average quality of DR on all scales with the most appropriate weights. The higher the AUC, the better the result.

When the data come with class labels, unsupervised DR can also be assessed by its performance in classification tasks, reporting the accuracy of a KNN classifier in the LD space:

$$G_{NN}(k) = \frac{1}{N} \sum_{i=1}^N \frac{|j \in n_i^k \text{ s.t. } c_i = c_j| - |j \in v_i^k \text{ s.t. } c_i = c_j|}{k} \quad (4.8)$$

Average the gain (or loss, if negative) of neighbors of the same class around each point after DR. Therefore, a positive value is probably correlated with better KNN classification performance. $G_{NN}(k)$ can also be summarized by a global score provided by its Area Under Curve (AUC), where $AUC|G_{NN}(k)| \in [-1, 1]$.

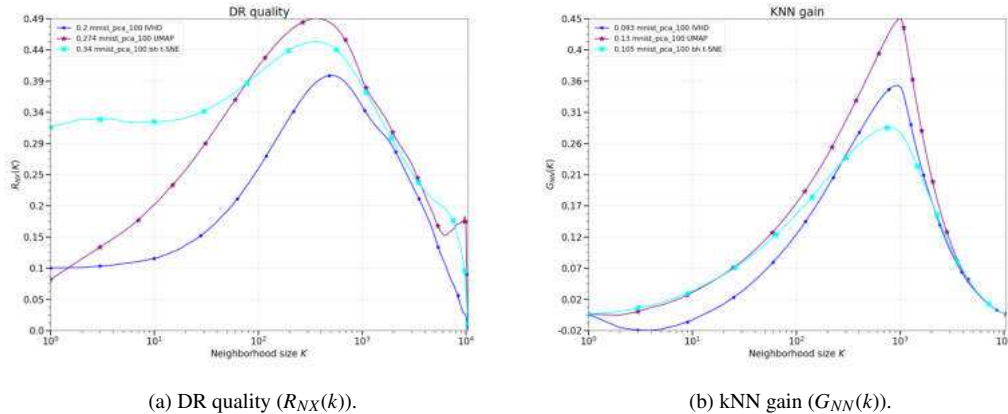


Figure 4.3: Example rank-based diagrams obtained for UMAP, t-SNE and IVHD visualizations of MNIST dataset.

As shown in Figure 4.3, the DR quality and kNN gain allow you to better observe the results of the data embedding depending on the size of the neighborhood that is being observed. This makes it possible to assess whether the method is better or worse in terms of locality (small neighborhood) or globality (large neighborhood size). However, it is important to remember that the size of the neighborhood (and whether we are talking about locality or globality) depends on the size of the dataset.

4.2. Datasets

The main properties of each baseline dataset are: the number of samples M , the dimensionality N , and the number of classes K . Here, we consider datasets with (1) a large number of samples and relatively low dimensionality, (2) a smaller number of samples but a larger number of features, (3) highly imbalanced data (RCV-Reuters), and (4) skewed data (Small NORB). The most common datasets used in different experiments are described in Table 2. A complete breakdown of the datasets used in this work can be found in the Appendix A.

Table 4.2: The list of some baseline datasets.

Dataset	N	M	K	Short description
MNIST	784	70 000	10	Well balanced set of grayscale images of handwritten digits.
Fashion-MNIST	784	70 000	10	More difficult MNIST version. Instead of handwritten digits it consists of apparel images.
Small NORB	2048	48 600	5	It contains stereo image pairs of 50 uniform-colored toys under 18 azimuths, 9 elevations, and 6 lighting conditions.
RCV-Reuters	30	804 409	8	Corpus of press articles preprocessed to 30D by PCA.

Preprocessing

Using PCA, we can introduce explained variance, which measures the proportion to which a mathematical model accounts for the variability (dispersion) of a given data set. Often variability is quantified as variance; then the more specific term explained variance can be used. In simpler words, the explained variance tells us how much information can be attributed to each of the principal components. Using this, we can draw a graph showing the relation between the total explained variance and the number of principal components used. We can observe in Figure 4.4 that we can decrease the number of dimensions to ~ 100 with PCA, still having 90% of information. We use this method when we do not necessarily need large (multidimensional) datasets. When we mention that the dimensionality of MNIST is 100-D, this means that it has been preprocessed using this methodology.

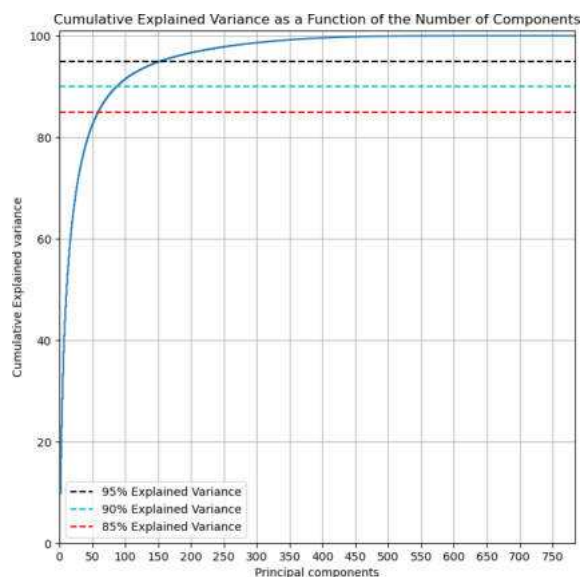


Figure 4.4: Cumulative explained variance by number of principal components for MNIST dataset.

In addition, we prepared a set of artificially generated "simple" data sets whose internal data structure gives insight into the quality of embeddings created by different methods. They are described in detail in the Appendix A. The Mammoth data set was

preprocessed, and labels were added based on the Z-axis value (colors were assigned by the height of the skeleton).

4.3. Baseline methods

Although there are many algorithms for visualizing multivariate data and the topic has been intensively researched for years, there are currently two methods that are used as a starting point for ongoing research: UMAP[90] and t-SNE[86]. Most of the concepts that are currently being tested and developed originate from them. Furthermore, LargeVis[131] is an interesting approach to visualization of the NN -graph, which is what IVHD is doing as well. For these reasons, these are the first three baseline methods (already described in Chapter 2) used for comparison with IVHD. In addition, recent research has contributed to the PaCMAP and TriMAP methods. The first optimizes a low-dimensional embedding using three types of point pairs: neighbor pairs, mid-near pairs, and furthest pairs. The second method uses triplet constraints to create a low-dimensional embedding of a set of points.

TriMap

DR method, which focuses on preserving the global structure of data in an embedding. The similarities between points appear to be insufficient to capture the global structure. Instead, TriMap uses a higher-order structure to construct the embedding using triples:

$$(i, j, k) \Leftrightarrow \text{point } i \text{ is closer to point } j \text{ than point } k. \quad (4.9)$$

The key idea of TriMap is derived from semi-supervised metric learning[3]: Given an initial low-dimensional representation for the data points, the triplet information from the high-dimensional representation of the points is used to improve the quality of the embedding. Similarly, the TriMap is initialized with a low-dimensional PCA embedding, and this embedding is then modified using a set of carefully selected triplets from the high-dimensional representation. TriMap chooses a subset $\mathcal{T} = \{(i, j, k)\}$ of triplets and assigns a weight $\omega_{ijk} \geq 0$ to each triplet: a higher value of ω_{ijk} implies that the pair (i, k) is located much farther away than the pair (i, j) . TriMap defines the loss of the triplet (i, j, k) as follows:

$$C_{ijk} = \omega_{ijk} \frac{s(\mathbf{y}_i, \mathbf{y}_k)}{s(\mathbf{y}_i, \mathbf{y}_j) + s(\mathbf{y}_i, \mathbf{y}_k)}, \quad s(\mathbf{y}_i, \mathbf{y}_j) = (1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}, \quad (4.10)$$

where $s(y_i, y_j)$ is a similarity function between y_i and y_j . The choice of s is motivated by the good performance of the Student's t distribution for similarities in low dimensions in the t-SNE method. Note that the loss of the triplet (i, j, k) approaches zero as $\|y_i - y_j\|$ decreases and $\|y_i - y_k\|$ increases. To reflect the relative similarities in high dimension, the weight of the triplet (i, j, k) is defined as:

$$\hat{\omega}_{ijk} = d_{ik}^2 - d_{ij}^2 \geq 0, \quad (4.11)$$

in which d_{ij} is any distance measure between x_i and x_j of high dimension. As the default distance, TriMap considers the squared Euclidean distance with scaling introduced in [152]: $d_{ij}^2 = \frac{\|x_i - x_j\|^2}{\sigma_{ij}}$, where $\sigma_{ij} = \sigma_i \sigma_j$ and σ_i are set to the average Euclidean distance between x_i and the set of nearest neighbors of x_i from neighbors of the 4- to 6-th grade. This choice of σ_{ij} adaptively adjusts the scaling according to the density of the data. We change the final weights by subtracting the minimum weight value calculated on all triplets and applying a tempered log transformation, $\omega_{ijk} = \log_t(1 + \hat{\omega}_{ijk} - \omega_{min})$, where $\omega_{min} = \min_{(i',j',k') \in \mathcal{T}} \hat{\omega}_{i'j'k'}$. Then, the tempered logarithm[100] is used:

$$\log_t(u) = \frac{1}{1-t}(u^{1-t} - 1), \quad t \neq 1. \quad (4.12)$$

Transformation \log_t smooths the weights and prevents triplets with large weights from dominating total loss. Note that the limit case $t \rightarrow 1$ recovers the standard log. The default value of t is 0.5.

Algorithm 13: TriMAP scheme.

Input: data matrix \mathbf{X} .

procedure TriMAP:

1. Find a set of triplets $\mathcal{T} = \{(i, j, k)\}$.
 2. Initialize \mathbf{Y} using PCA.
 3. Minimize the C_{TriMAP} by applying full batch gradient descent with momentum using the delta-bar-delta method.
 4. Return \mathbf{Y} .
-

To construct the embedding, TriMAP considers a small subset of all possible triplets (i, j, k) for which the closest point j belongs to the set of nearest neighbors of the point i and the farther point k is among the points that are more distant from i than j , chosen uniformly at random. For each point, it considers its nearest neighbors $m=10$ and the sample $m'=5$ triplets per nearest neighbor. This yields $m \times m' = 50$ nearest-neighbor triplets per point. Furthermore, we also add random triplets $r=5$ (i, j, k) per each point i where j and k are sampled uniformly at random and their order can change based on their proximity to i . Thus, the overall complexity of the optimization step is linear in number of points n . The computational complexity is dominated by the nearest-neighbor search. The final loss is defined as the sum of the losses of the triplets sampled in \mathcal{T} :

$$C_{TriMap} = \sum_{(i,j,k) \in \mathcal{T}} C_{ijk}. \quad (4.13)$$

The loss is minimized by using the full-batch gradient descent with momentum by using the delta-bar-delta method. The whole procedure is designed in a way that samples the informative triplets from the high-dimensional representation of a set of points and assigns weights to these triplets to reflect the relative similarities of these points. Although TriMap can also be used for the triplet embedding task, it focuses mainly on the DR.

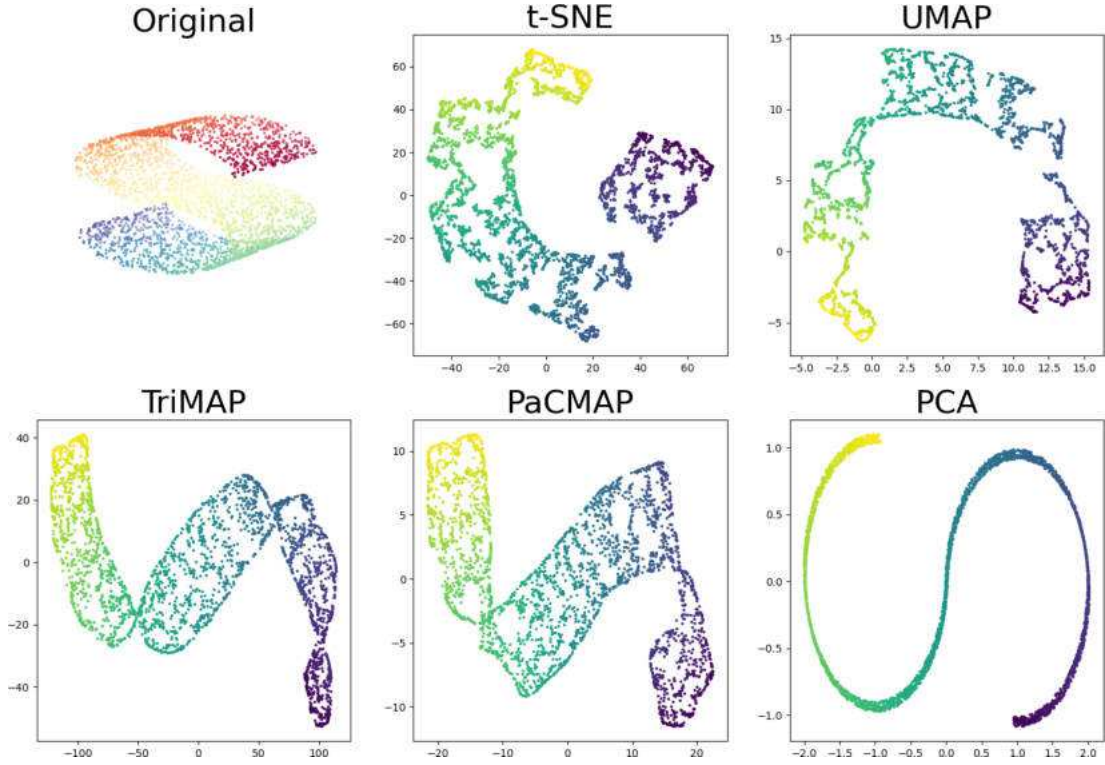


Figure 4.5: 2-D visualizations of the S-curve dataset. As shown, both TriMAP and PaCMAP successfully unveils the underlying structure in the original dataset.

PaCMAP

The goal of Pairwise Controlled Manifold Approximation Projection (PaCMAP) is also to preserve both local and global structure (as in TriMAP). It does this by distinguishing between three types of edges in the graph: neighbor pairs (NB), mid-near pairs (MN), and further pairs (FP). The first group consists of the n_{NB} nearest neighbors of each observation in the high-dimensional space. Similarly to TriMAP, the following scaled distance metric is used:

$$d_{ij}^2 = \frac{\|x_i - x_j\|^2}{\sigma_{ij}}, \text{ where } \sigma_{ij} = \sigma_i \sigma_j \quad (4.14)$$

where σ_i is the average distance between i and its Euclidean nearest fourth to sixth neighbors. These are used to construct neighbor pairs $(i, j_t), t = 1, 2, \dots, n_{NB}$. The second group consists of $N \cdot n_{MN}$ mid-near pairs selected by randomly sampling six additional observations from each observation and using the second smallest of them for the mid-near pair. Finally, the third group consists of a random selection of n_{FP} furthest points from each observation. For convenience, the number of pairs of mid-near and other points is determined by the parameter ratio MN and the ratio FP that specify the ratio of these quantities to the number of nearest neighbors, that is, $n_{MN} = MN_{ratio} \cdot n_{NB}$ and $n_{FP} = FP_{ratio} \cdot n_{NB}$. Since the number of neighbors n_{NB} is typically an order of magnitude smaller than the total number of observations, random sampling effectively chooses non-nearest neighbors as mid-near and further pairs. The weights

associated with neighbor, near-neighbor and other pairs in iteration t are set to the following default values:

- for $t \in [\tau_1, \tau_2)$: $w_{NB} = 2$, $w_{MN}(t) = 1000 \cdot (1 - \frac{t-1}{\tau_2-1}) + 3 \cdot \frac{t-1}{\tau_2-1}$, $w_{FP} = 1$;
- for $t \in [\tau_2, \tau_3)$: $w_{NB} = 3$, $w_{MN} = 3$, $w_{FP} = 1$;
- for $t \in [\tau_3, n_{iterations}]$: $w_{NB} = 1$, $w_{MN} = 0$, $w_{FP} = 1$;

PaCMAP also uses three distinct cost functions for each type of pair:

$$C_{NB} = \frac{\hat{d}_{ij}}{10 + \hat{d}_{ij}}, \quad C_{MN} = \frac{\hat{d}_{ik}}{10000 + \hat{d}_{ik}}, \quad C_{FP} = \frac{1}{1 + \hat{d}_{il}}, \quad (4.15)$$

where $\hat{d}_{ab} = \|y_a - y_b\|^2 + 1$. The final cost function, where the weights are dynamically updated is defined as follows:

$$C_{PaCMAP} = w_{NB} \cdot \sum_{i,j \text{ are neighbors}} \frac{\hat{d}_{ij}}{10 + \hat{d}_{ij}} + w_{FP} \sum_{i,l \text{ are furthest neighbors}} \frac{1}{1 + \hat{d}_{il}} + w_{MN} \cdot \sum_{i,k \text{ are mid-near neighbors}} \frac{\hat{d}_{ik}}{10000 + \hat{d}_{ik}} \quad (4.16)$$

Algorithm 14: Pairwise Controlled Manifold Approximation Projection scheme.

Input: data matrix \mathbf{X} .

procedure PaCMAP:

1. Find a set of near, mid-near and furthest pairs.
 2. Initialize \mathbf{Y} using PCA.
 3. Begin three optimization phases that satisfies:
 $\tau_1 = 1 \leq \tau_2 \leq \tau_3 \leq n_{iterations}$.
 Default values: $\tau_1 = 1$, $\tau_2 = 101$, $\tau_3 = 201$.
 4. Set initial weights.
 5. Run AdamOptimizer for $n_{iterations}$ to minimize the cost function C_{PaCMAP} , while simultaneously adjusting the weights.
 6. Return \mathbf{Y} .
-

Dynamic Optimization. The optimization process consists of three phases that aim to avoid the local optimum. In the first phase, the goal is to improve the initial distribution of embedded points to one that preserves both global and local structure to some extent but mainly global structure. This is achieved by heavily weighting mid-near pairs. During the first phase, PaCMAP gradually reduces the weights on the pairs mid-near, allowing the algorithm to gradually shift from the global structure to the local structure. In the second phase, the goal is to improve the local structure while preserving the global structure captured in the first phase by assigning a small (but not zero) weight to mid-near pairs. Together, the first two phases attempt to avoid local optima using a process that reveals similarities to simulated annealing and the *early exaggeration* technique used by t-SNE. However, early exaggeration places more emphasis on

neighbors rather than midpoints, while PaCMAP focuses on midpoint pairs first and neighbors later. Finally, in the third phase, the focus is on improving the local structure by reducing the weight of the center pairs of points to zero and the weight of neighbors to a smaller value, emphasizing the role of the repulsive force that helps separate possible clusters and make their boundary clearer.

4.4. Experiments

In this section, we present visualizations of small artificially generated datasets that provide an easy way to evaluate the specific properties of the baseline DR algorithms. Additionally, we present hardware on which calculations were performed for specific cases.

4.4.1. Hardware

All CPU implementations of the baseline embedding were executed in two environments, depending on the scale of the dataset processed. Mid-scale datasets were visualized on Macbook Pro 2,3 GHz 8-Core Intel Core i9, 16 GB 2667 MHz DDR4. Large-scale datasets were processed in HPC Prometheus environment: Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz, 64 GB RAM. The codes were compiled using Apple Clang version 13.1.6 and GCC-10.1.

All GPU/CUDA implementations of the baseline embedding methods were executed on the separated remote server with: CPU Intel Xeon E5-2620 v3, GPU Nvidia GeForce GTX 1070 (1920 CUDA cores, 8GB GDDR5), 252 GB RAM, OS: Ubuntu 18.04.3, architecture x86, 64. Codes were compiled using GCC-7.4 and CUDA Toolkit 10.0.

4.4.2. Synthetic datasets

To evaluate the DR properties of the baseline methods, we begin with experiments on the synthetic datasets described in Table A.3. In this way, we can make a preliminary assessment of which methods have more advantages in terms of locality and globality.

In Figure 4.6, we conducted an experiment that presented how a two-dimensional mesh with evenly spaced points was embedded by all baseline methods while using different metrics to calculate distances. Both the PaCMAP and TriMAP methods do not provide the possibility of using precalculated distances or the Chebyshev norm in the calculation process. It can be seen that both t-SNE and IVHD achieve the best embedding, which preserves the global structure and shape of the mesh without introducing any distortion. UMAP introduce distortions and that curl the mesh during the embedding process. In addition, we can see that all methods cause compression of points, that are located in the corners of the mesh. The overall shape is best preserved by IVHD. The mesh edges in the embedding were arranged as straight lines, and the data points are mostly evenly spaced and also arranged as straight lines, just like in the original mesh.

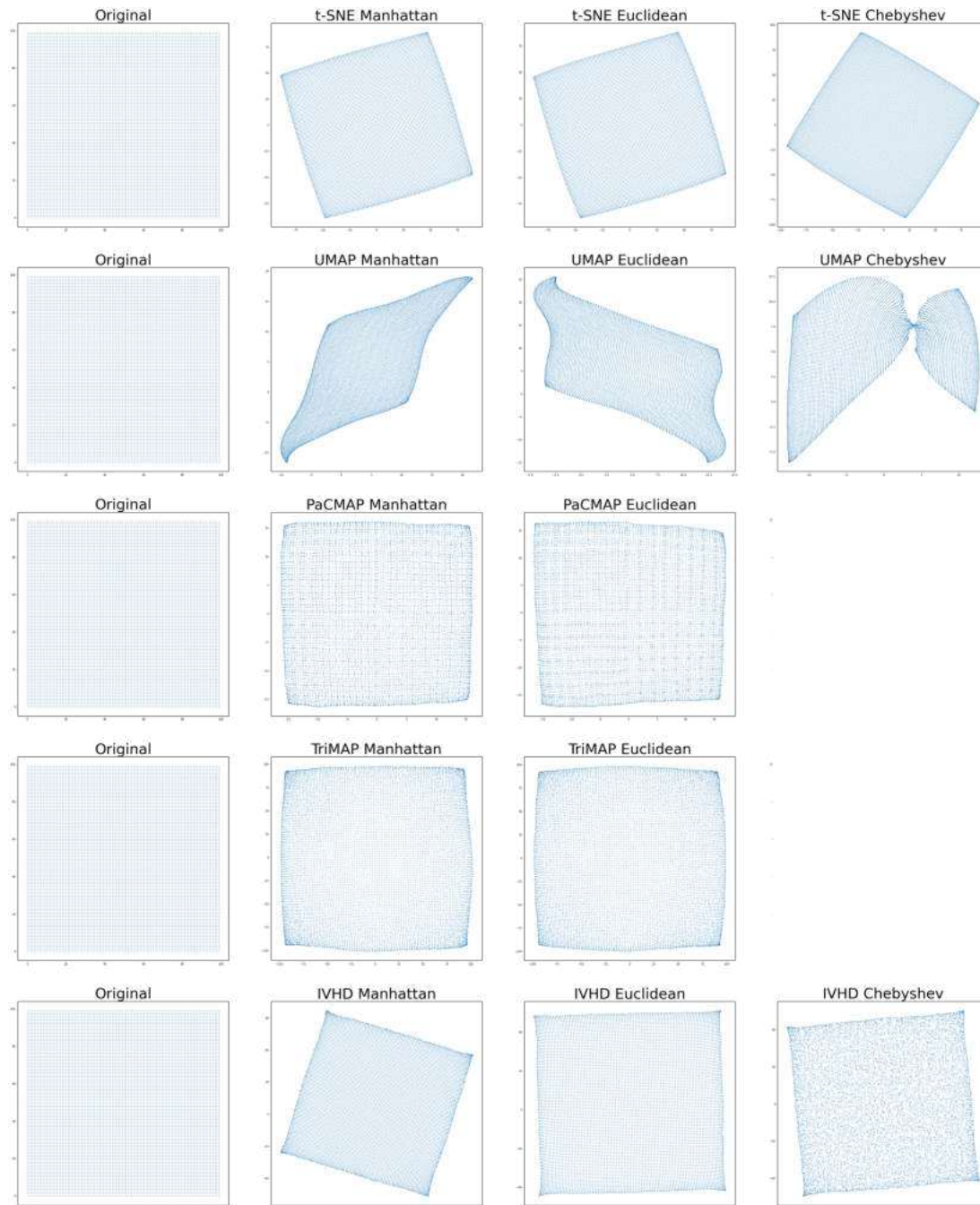


Figure 4.6: Methods comparison for embedding 2-dimensional grid using various metrics for calculating distance.

In Figure 4.7, we demonstrated how t-SNE, UMAP and IVHD embed 2-dimensional grid into 3-dimensional space, when using periodic boundary condition (PBC). Theoretically, in this process one should obtain a torus-like structure. In PBC, the geometry of the unit cell satisfies perfect two-dimensional tiling, and when an object passes through one side of the unit cell, it re-appears on the opposite side with the same velocity. As we can see, IVHD is the only method, that generates torus-like structure among given set of methods. TriMAP, LargeVis and PaCMAP were not considered in this experiment because their implementations do not allow passing precalculated

graph as input to the algorithm.

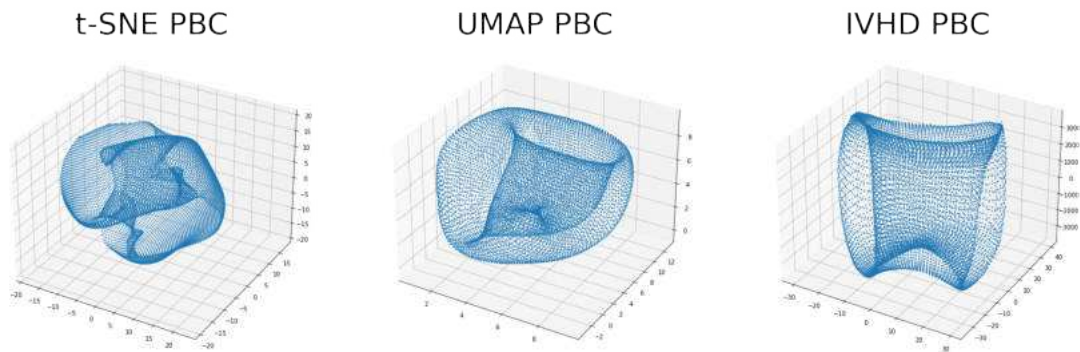


Figure 4.7: Methods comparison for embedding 2-dimensional grid into 3-dimensional space using periodic boundary condition.

In Figure 4.8 we show that t-SNE, TriMAP, PaCMAP, and IVHD manage to obtain the global structure of the data in which we can see the ball and the two spheres that surround it. UMAP, on the other hand, is completely unable to separate classes.

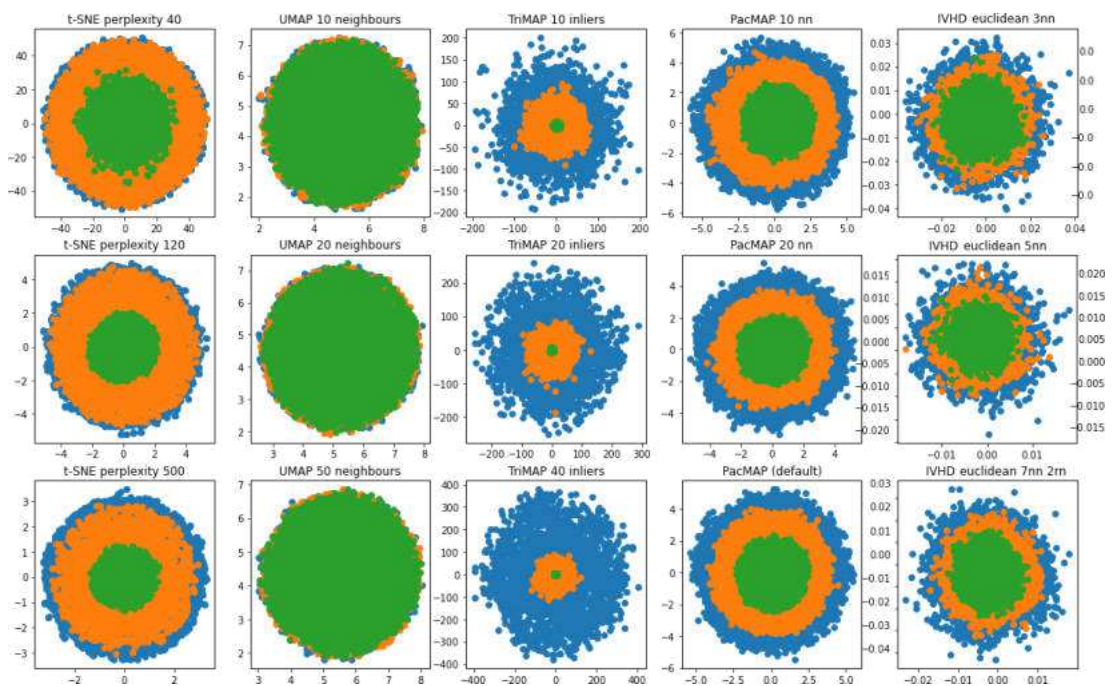


Figure 4.8: Methods comparison for 15k-sample ball inside of 2 spheres. Both ball and spheres were generated in 30-D space.

It is worth mentioning that different parameterizations were tested for each dataset. Here, the best results obtained for each method are presented (both in terms of visual aspect and in terms of metrics obtained).

The mammoth data set (Figure 4.10) is a 3D point cloud of a mammoth skeleton. Its main research value is that, with this data set, the preservation of local and global properties can be perfectly verified. Ideally, we would like the location of subsequent

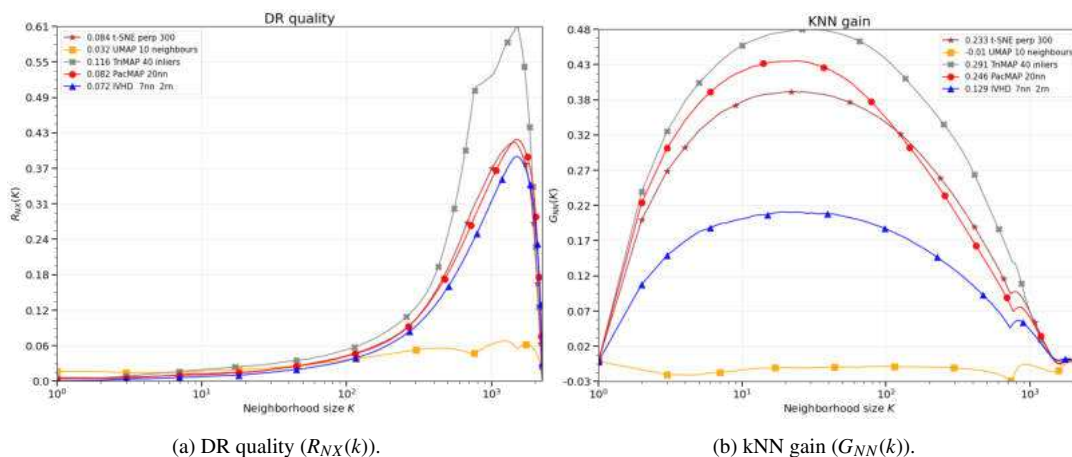


Figure 4.9: KNN gain and DR quality obtained for comparison of different DR methods for "ball in 2 spheres" dataset.

mammoth body parts to be logically correct (e.g., mammoth tusks should be attached to the head). Visualization of original mammoth data set is presented in Appendix A.

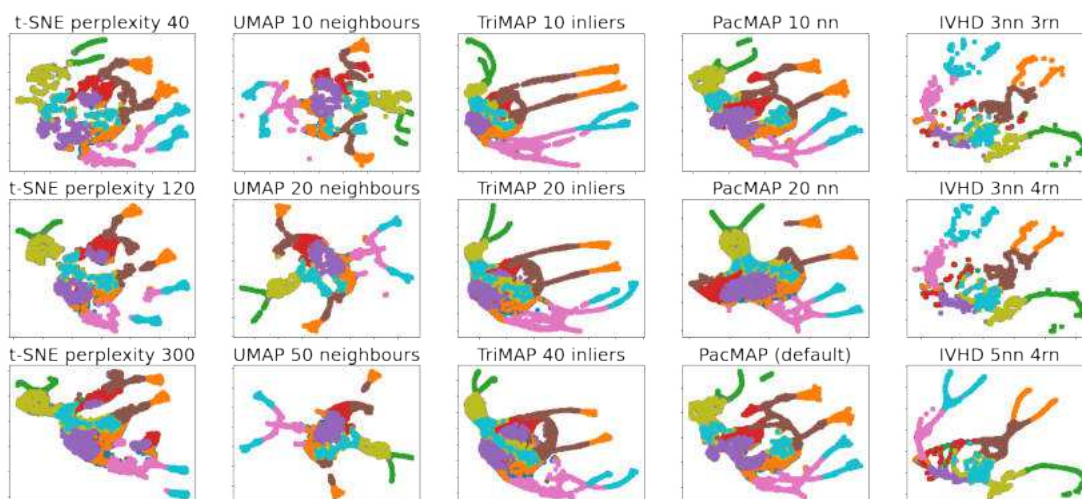


Figure 4.10: Methods comparison for Mammoth dataset.

PaCMAP, TriMap and IVHD methods achieve a good balance of local and global properties in mammoth dataset embedding. The next mammoth body parts are properly embedded in the low-dimensional 2-D space. For t-SNE and UMAP, it is evident that the overall structure of the embedding is not as clear and obvious, where classes are incorrectly distributed in space.

Similarly to Figure 4.8, in Figure 4.12, we can see that t-SNE and PaCMAP manage to maintain the global and local structure of the data. What is interesting is that IVHD is the only method that was capable of separate classes by inserting a "space" between the ball and the sphere. Classes in other methods are highly congested. Furthermore, we see that TriMap was unable to split the classes as clearly as in 4.8 (although both datasets are fairly similar). UMAP was again unable to separate classes at all.

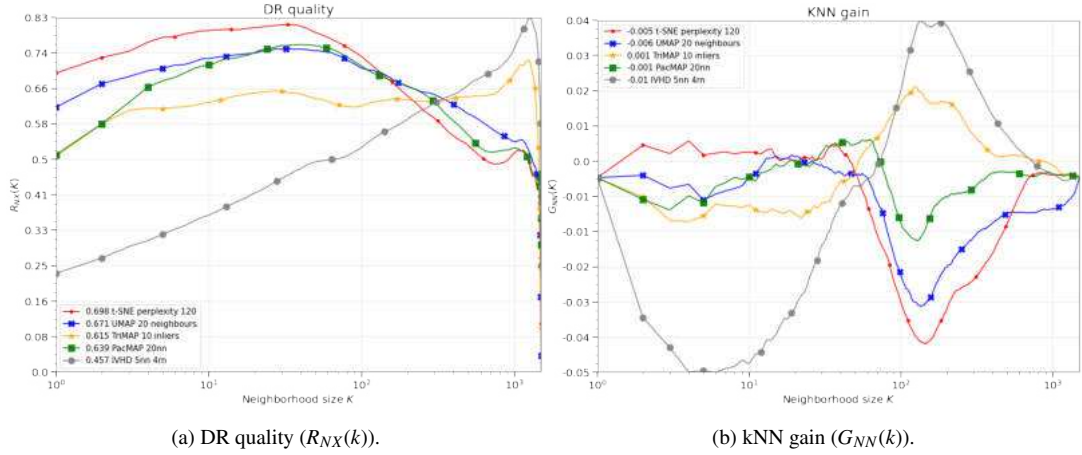


Figure 4.11: KNN gain and DR quality obtained for comparison of different DR methods for "Mammoth" dataset.

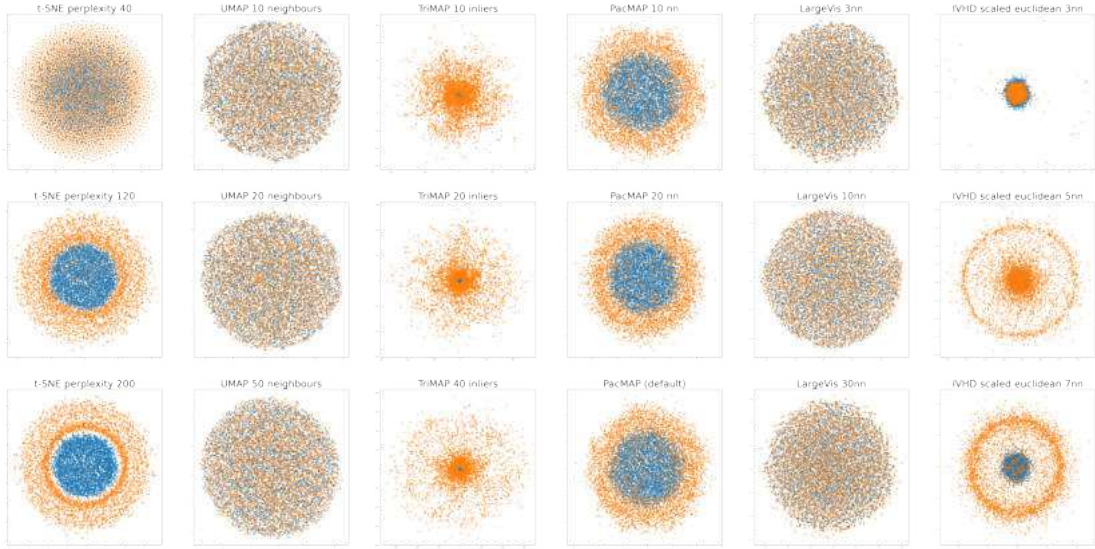


Figure 4.12: Methods comparison for ball inside a sphere dataset.

The KNN gain and the DR quality achieved for this data set (Figure 4.13) support the conclusions described above. By far, the worst performer was UMAP. In terms of DR quality, IVHD and TriMap are creating visualizations of the highest quality.

Using Figs. 4.12 and 4.8 helps us visualize *crowding phenomenon* observed during dimension reduction by different data embedding methods. Using the "ball inside the sphere" dataset as our framework, we assume that a high-dimensional ball contains two classes of points. Class 1 is inside the ball $B_2^n(1)$ and class 2 is inside a spherical shell outside class 1. When $\{X_i^N\}_{i=1} \subseteq \mathbb{R}^2$ is two-dimensional, a direct visualization (Table A.3 in Appendix A) shows the correct relation between the two classes. When $\{X_i^N\}_{i=1} \subseteq \mathbb{R}^5$ is five-dimensional and assigned to \mathbb{R}^2 using the DE method, its geometrical structure is distorted. Then, we see a severe crowding phenomenon: the second class is pushed towards the center. This is due to the difference in norm concentration between high and low dimensions: a ball in higher dimensions has a volume that

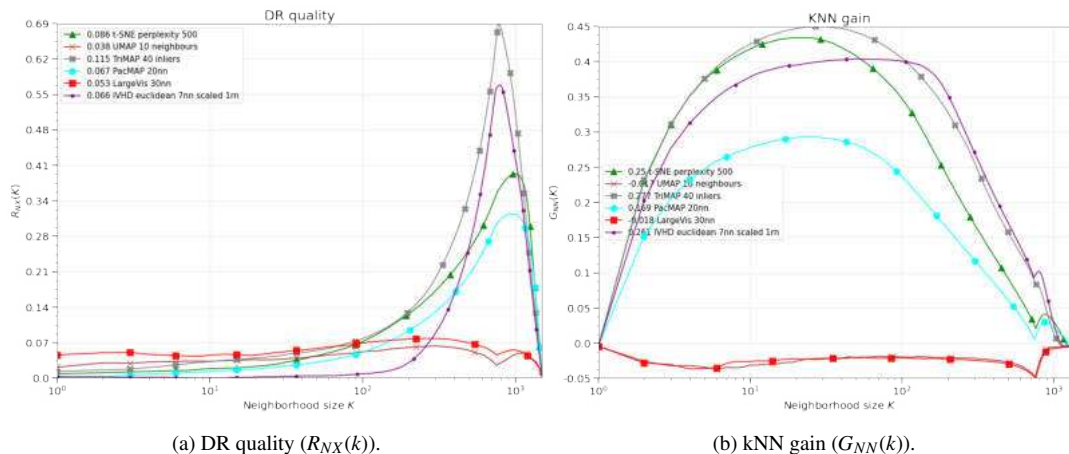


Figure 4.13: KNN gain and DR quality obtained for comparison of different DR methods for ball inside 1 sphere dataset.

Table 4.3: Precision values cf_{nn} and trustworthiness T_{nn} for the FMNIST and small-NORB data sets (and various embedding methods).

	FMNIST					smallNORB				
	cf_{15}	cf_{50}	cf_{100}	T_{15}	T_{50}	cf_{15}	cf_{50}	cf_{100}	T_{15}	T_{50}
t-SNE	0.956	0.950	0.947	0.987	0.976	0.951	0.845	0.754	0.973	0.946
UMAP	0.951	0.951	0.949	0.978	0.975	0.911	0.845	0.768	0.966	0.938
TriMap	0.944	0.944	0.943	0.967	0.977	0.914	0.863	0.813	0.967	0.948
PaCMAP	0.956	0.955	0.955	0.974	0.973	0.951	0.918	0.852	0.968	0.939
LargeVis	0.944	0.942	0.941	0.961	0.952	0.922	0.867	0.795	0.965	0.913
IVHD	0.894	0.893	0.892	0.938	0.933	0.904	0.847	0.764	0.922	0.906

grows faster with radius $Vol(B_2^n(r)) \sim r^n$ [35], where $B_2^n(r)$ is a ball in \mathbb{R}^n with radius r . As a result, high-dimensional points are more likely to be distributed near the surface. When mapped to low dimensions, since there is less room near the surface, the points are pushed towards the center.

4.4.3. Mid-scale baseline datasets

The second part of evaluating the DR properties of the baseline methods is to perform experiments on mid-scale ($N < 150000$) datasets. Analyses of the datasets that are not covered in the following section are found in the Appendix B. It is worth mentioning that the most efficient method in terms of computational time is IVHD. It is the fastest and requires the least amount of memory. The timings obtained for embedding all datasets using baseline methods are presented in the Appendix B.

Poza Based on the visualizations and diagrams presented in Figs. 4.14, 4.15, one can make the following observations considering visualized datasets:

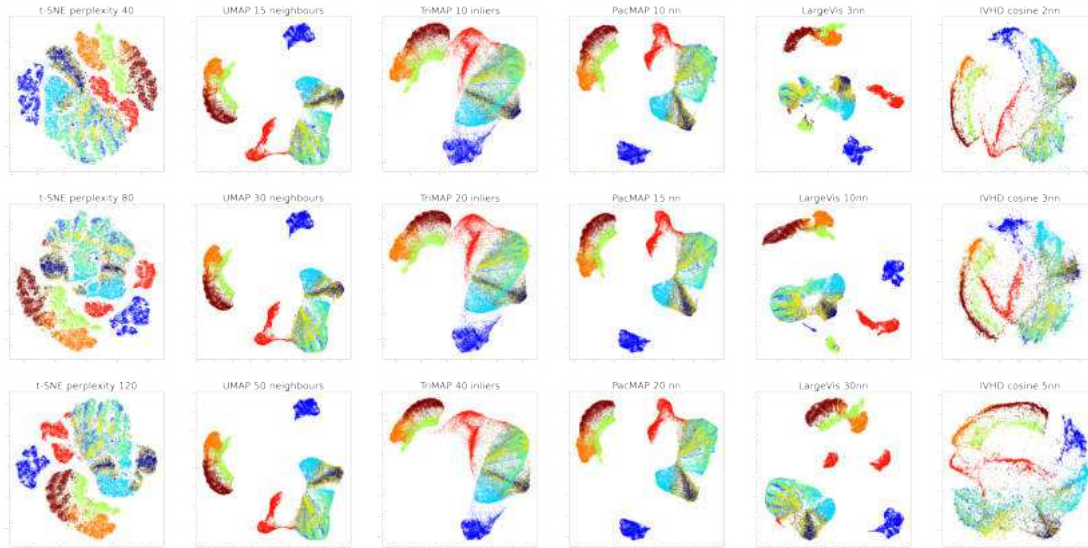
1. IVHD applied to Fashion-MNIST in Fig. 4.14 forms separate clusters of different, mostly elongated shapes. Moreover, the generated mapping is fuzzy. On the other hand, the *-MAP* methods are much better at creating rounded and clearly sepa-

rated clusters. Additionally, in t-SNE some classes are mixed and fragmented. In terms of DR quality, TriMap, UMAP, and PaCMAP are achieving the best results. It is also reflected by using the metrics cf and T , which have a very stable value regardless of the number of neighbors considered. IVHD starts to exceed t-SNE and LargeVis, when a very large neighborhood is considered ($k > 1000$).

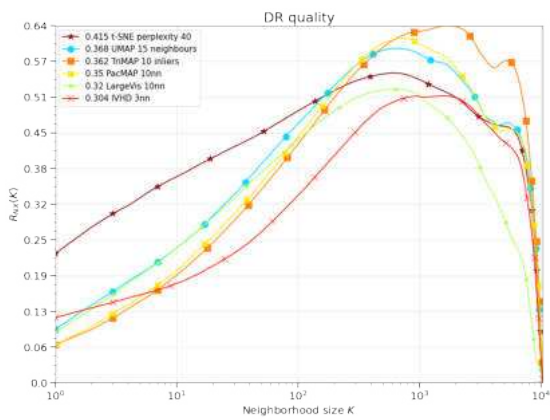
2. All methods were more or less capable of preserving the distance ratio between both (high- and low-dimensional) spaces (Fig. 4.14). Trimap seems to have the most condensed points on the diagonal of the Shepard diagram, which indicates the best preservation of the distance ratio.
3. For the smallNORB dataset, the TriMap and IVHD methods achieve the best results (Fig. 4.15) in terms of both the DR quality and the sustained distance ratio shown in the Shepard diagram (Fig. 4.15). In terms of the cf metric, PaCMAP achieves the highest score for all neighborhood sizes. Again, it generates separate clusters of different classes, which are mostly elongated in shape. All the embeddings generate similar in terms of global distribution of classes visualization that contains: 1) separated three classes in LD space (brown, green, orange), 2) separation of other two classes (red, blue) into common sub-space, and 3) mixing of the other four classes.
4. All embedding methods split the smallNORB dataset (Fig. 4.15) internally, which means that single classes are divided into separate groups. This indicates that all methods managed to find discrepancies between pictures that contain the same class of objects, but with different lighting or rotation.

As shown above, the quality and fidelity of the embeddings strongly depend on the structure of a specific dataset, user expectations, and their data visualization requirements. In general, the local structure of the *source* data is better reconstructed by the baseline algorithms. This is not a surprise because the IVHD algorithm does not use the correct ordering of k NN neighbors and the original distances between a sample y_i and its k NN neighbors.

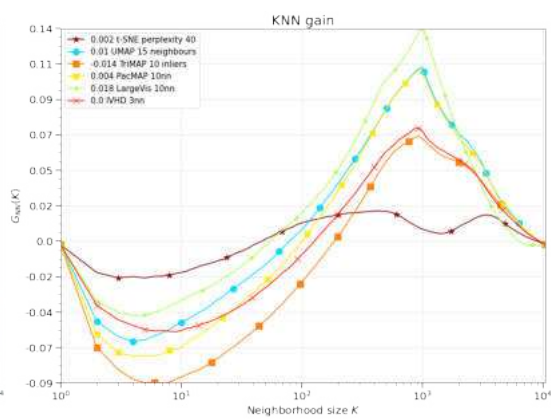
The results obtained for the large-scale baseline datasets ($M > 10^6$) were created using CUDA implementations (*IVHD-CUDA*) and are described in Chapter 5. This was due to much faster computation times and the fact that the reference methods did not generate visualizations for such large sets in a relatively reasonable time (<12h). Furthermore, the t-SNE, TriMAP, and PaCMAP CPU implementations generated errors (probably due to insufficient memory resources). It is worth mentioning that UMAP is the only method that contains benchmarks for large-scale datasets (although the resources needed for the calculations are much higher than for CUDA algorithms).



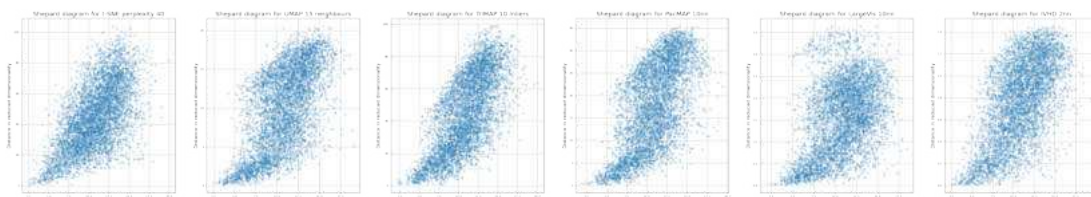
(a) Methods comparison for FMNIST dataset.



(b) DR quality ($R_{NX}(k)$).

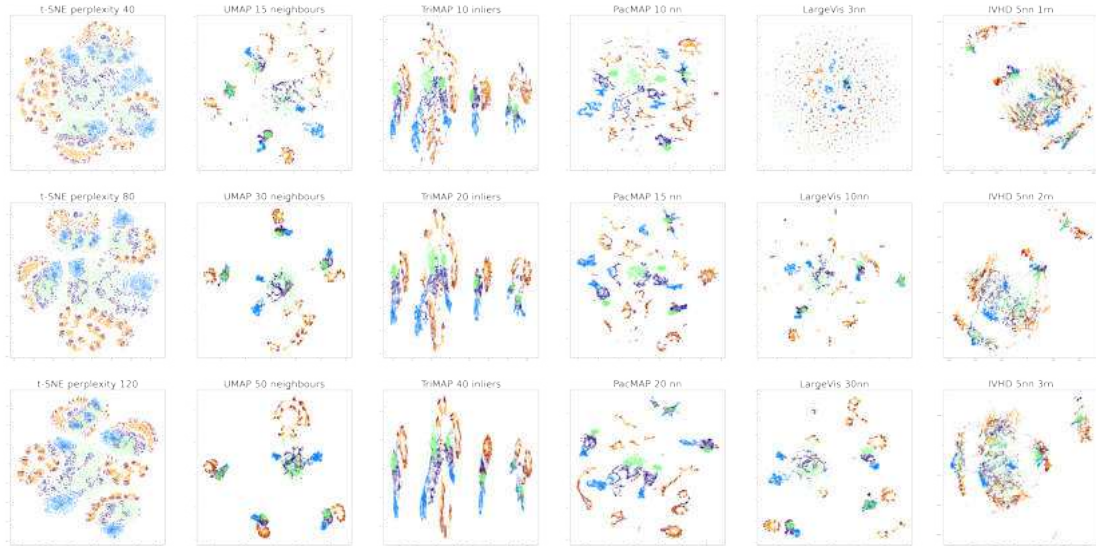


(c) kNN gain ($G_{NN}(k)$).

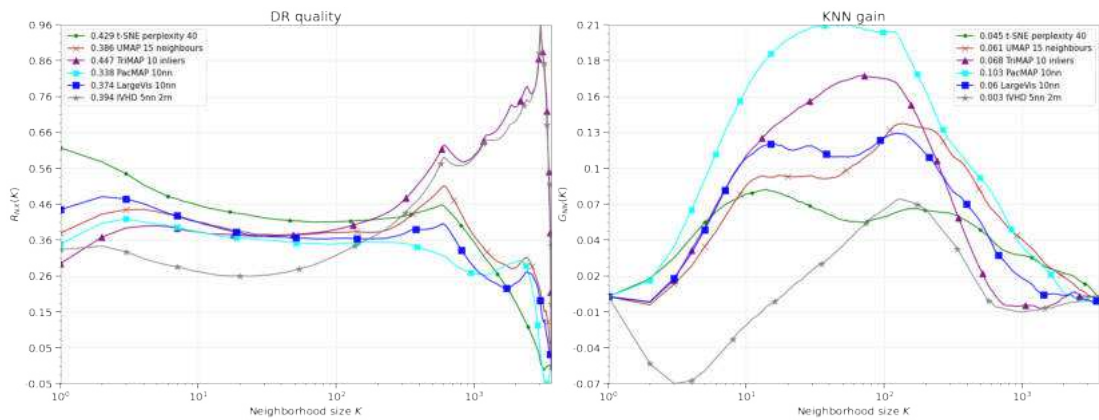


(d) Shepard Diagram ($G_{NN}(k)$).

Figure 4.14: Visualizations, kNN gain, DR quality and Shepard Diagrams obtained for comparison of different DR methods for FMNIST dataset.

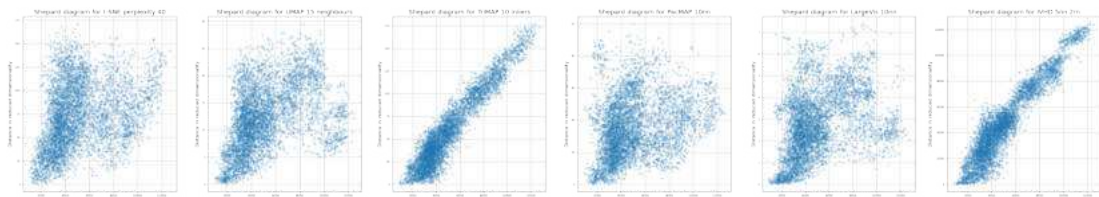


(a) Methods comparison for smallNORB dataset.



(b) DR quality ($R_{NX}(k)$).

(c) kNN gain ($G_{NN}(k)$).



(d) Shepard Diagram ($G_{NN}(k)$).

Figure 4.15: Visualizations, kNN gain, DR quality and Shepard Diagrams obtained for comparison of different DR methods for smallNORB dataset.

CHAPTER 5

GPU-EMBEDDING OF KNN-GRAPH REPRESENTING LARGE AND HIGH-DIMENSIONAL DATA

Despite the fact that there are many algorithms for the visualization of high-dimensional data and that this topic has been extensively studied for years, there are only a few implementations of modern data embedding algorithms in the GPU/CUDA environment [27, 42, 105]. This was the root cause of the creation of IVHD-CUDA [93], which was compared with publicly available CUDA implementations, which generated the best embeddings of large datasets: BH-SNE-CUDA [27] and Anchor-t-SNE [42] (At-SNE). All algorithms consist of two stages: (1) generate of a weighed kNN-graph; (2) run a proper embedding procedure which is based on: definition of a loss function and its minimization.

5.1. *k*NN graph generation

*k*NN-graph approximates a n D non-Cartesian manifold immersed in \mathbb{R}^N sampled by the feature vectors y_i . Here, we consider the *k*NN-graph construction procedure shipped by the FAISS library [64]. Its authors claim that it is currently the fastest available *k*NN search algorithm implemented on GPU. The FAISS *k*NN-search procedure merges a very efficient and well-parallelized exact *k*NN algorithm and indexing structures that allow an approximate search. To achieve a high-efficiency search with good accuracy, we employ the IVFADC [19] indexing structure. It uses two levels of quantization combined with vector encoding for compressing high-dimensional vectors. The main idea behind the use of indexes is to divide the input space and divide all input samples into a number of clusters represented by their centroids. To handle a query, the algorithm compares it with the centroid centers, picks the centroid that is the most similar to the query vector, and performs an exact search within the set of samples belonging to this centroid. In addition to the efficient *k*NN algorithm, its CUDA implementation is extremely well optimized [64]. We use the same FAISS *k*NN-graph generation procedure in both the baseline and IVHD algorithms.

5.2. BH-SNE-CUDA, At-SNE, IVHD-CUDA

BH-SNE-CUDA

BH-SNE (Barnes-Hut t-SNE [85]) is an approximation of the t-SNE method that can reduce the computational complexity of DE from $O(M^2)$ to $O(M \log M)$ using the Barnes-Hut approximation (described in Section 2.4.1) but at the cost of increasing algorithmic complexity and, consequently, decreasing parallelization efficiency [85]. Its GPU version, the BH-SNE-CUDA algorithm, does not introduce any changes to the BH-SNE [85] algorithm and just matches the instructions and data flows to the GPU architecture [27].

AtSNE-CUDA

The AtSNE-CUDA algorithm (Anchor-t-SNE [42]) was created to deal with issues related to t-SNE, such as sensitivity to initial conditions and inadequate reconstruction of the global data structure. The authors of the AtSNE-CUDA method claim that their method is 50% faster than BH-SNE-CUDA and has lower memory requirements. As shown in [42], it still generates good quality embeddings, although some k NN quality scores are better for the classic t-SNE or LargeVis algorithms [131].

Loss function: Similar to all t-SNE clones, AtSNE minimizes the regularized KL divergence. Information about the local structure of the data can be obtained from the set of approximated k NN neighbors of each y_i . Meanwhile, to reconstruct the global structure of the data, the x_i points receive 'pulling forces' from the so-called *anchor points* generated from the original data. In this way, *anchor points* are responsible for maintaining the mutual positions and shape of classes in X . Consequently, the hierarchical embedding [42] optimizes both the positions of the *anchor points* and regular data samples.

Let A represent the set of *anchor points* in the high-dimensional space and B its low-dimensional embedding. The probabilities $P(Y)$ and $Q(X)$ denote the high- and low-dimensional distributions of *anchor points*, respectively. To preserve both global and local information, the following loss function is minimized:

$$E(.) = \sum_i KL(P(Y)||Q(X)) + \sum_i KL(P(A)||Q(B)) + \sum_i \|b_i - \frac{\sum_{y_k \in C_{b_i}} y_k}{|C_{b_i}|}\|, b_i \in C_{b_i} \quad (5.1)$$

where C_{b_i} denotes the set of points whose cluster center of the K-means is b_i . The last term is a regularization term explained in detail in [42].

Optimization: The AtSNE-CUDA algorithm preserves global information by minimizing the term $KL(P(A)||Q(B))$ in Eq.5.1, while the first term of the loss function is responsible for preserving the local structure of the data. The reason is that the gradient of the loss function is more complicated than that defined by Eq. 3.9, AtSNE-CUDA uses the Hierarchical Optimization Algorithm [42]. This algorithm consists of two optimization layers: global and local. Optimization proceeds in a top-down manner. The main idea is to optimize the two-layer layout alternately: 1) fix the layout of the ordi-

nary points and optimize the layout of the anchor points; 2) fix the anchor point layout and optimize the layout of ordinary points.

IVHD-CUDA

In IVHD-CUDA, the force-directed optimization scheme is used, similar in spirit to the synchronous *momentum* and *Nesterov* methods, described in detail in [33]. Let *connection* mean two "particles" x_i and x_j (the nearest or random neighbors) in a 2D embedded space X joined by an edge. The main IVHD-CUDA implementation loop consists of the following steps: 1) calculating "forces" for each *connection* where the "force" is proportional to the gradient dependent on the two "particles'" positions and their velocities; 2) summing up the forces and 3) updating positions and velocities for all particles simultaneously. The simplified pseudocode is presented below as Algorithm 15.

Algorithm 15: Simplified IVHD-CUDA scheme.

Input: k NN-graph precalculated with FAISS library.

initialization;

while *running* **do**

 allocate subsets of force components to threads;

for *thread in threads* **do**

 | calculate force components for each connection;

end

 join threads;

 allocate subsets of samples to threads;

for *thread in threads* **do**

 | sum force/gradient components;

 | calculate new velocities;

 | update particle positions;

end

end

Because the steps are executed sequentially, this allows one to avoid the race condition between CUDA threads. Each connection receives two different memory addresses where the calculated values can be stored, one *positive* and one *negative*. If a particle i attracts particle j (*positive* value), then at the same time a particle j attracts particle i (*negative* value). In the first step, each thread calculates a subset of the force / gradient components between the particles. During the same time, they execute the same instructions, and thus no branching is required. However, this approach has a minor bottleneck. To save the result for each *connection*, a given thread must execute instructions from *write* to completely different memory addresses that represent two samples. Meanwhile, minimizing the number of *read/write* operations is always beneficial, as they are very time-consuming.

In the second step, each thread is assigned to a set of particles to process. As precautions against processing particles that interact with different numbers of particles

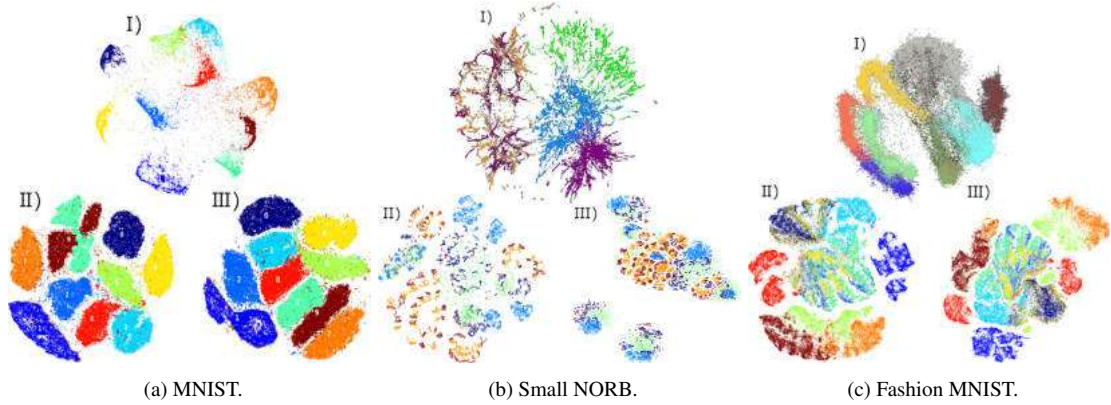


Figure 5.1: Visualization of datasets using: I) IVHD-CUDA, II) BH-SNE-CUDA, III) AtSNE.

(which might cause branch divergence), we sort the particles before the first iteration by considering the number of force components to calculate. After sorting, it is guaranteed that all threads in a warp will process samples with the same number of *connections*. The percentage share of each component of the DE procedure (without k NN graph generation), e.g., for 70.000 feature vectors with approximately three neighbors each ($nn=2$ and $rn=1$), is as follows: (1) calculation of the force component (78.97%), (2) update of positions (20.6%), (3) initialization (0.43%). In the IVHD-CUDA implementation, the global register and constant memory are used. While *connections*, the force components and positions are stored in global memory, the algorithm parameters are stored in constant memory. A register memory is used for auxiliary and temporary calculations.

Data embedding comparison

In this section, we briefly compare IVHD-CUDA with the most efficient and robust publicly available GPU-implemented data embedding codes: BH-SNE-CUDA and AtSNE-CUDA. The most important parameters of the methods are collected in Table B.1. For the BH-SNE-CUDA and AtSNE-CUDA methods, we use default parameters (appx. 15 parameters) proposed in [42, 85] and submitted to the GitHub repository with the respective GPU codes [21, 27]. IVHD parameters are also selected by default, for example, $rn = 1$ while nn is automatically fitted as a minimal value producing the largest connected component, which approximates (with a given high accuracy) the k NN-graph. The value of $c \in [0.01, 0.1]$ from Eq. 3.16 is the only parameter that can be interactively adjusted.

All datasets used in this section are described in Appendix A. There you can find the exact dimensionality of the dataset and a brief description of what is in the dataset.

Based on the visualizations presented in Figs. 5.1, 5.2, one can make the following observations:

1. In Fig. 5.1a, IVHD clearly reflects the global structure of MNIST classes and their separation. However, due to the "crowding effect", the classes become very

dense in the centers and sparse between them. On the other hand, the results of graph visualization presented in [32], demonstrate that the "crowding effect" can be controlled by tuning IVHD-CUDA parameters. Consequently, even the fine-grained neighborhood can be preserved with astonishing high precision (see [34]).

2. For the smallNORB dataset (see Fig.5.1b), IVHD-CUDA was able to visualize three large clusters clearly separable with a fine-grained data structure. For BH-SNE-CUDA and AtSNE-CUDA the quality of embeddings is much worse and more fragmented. Changes in the parameter values of the baseline algorithms (*perplexity*) do not improve the quality of the visualization.

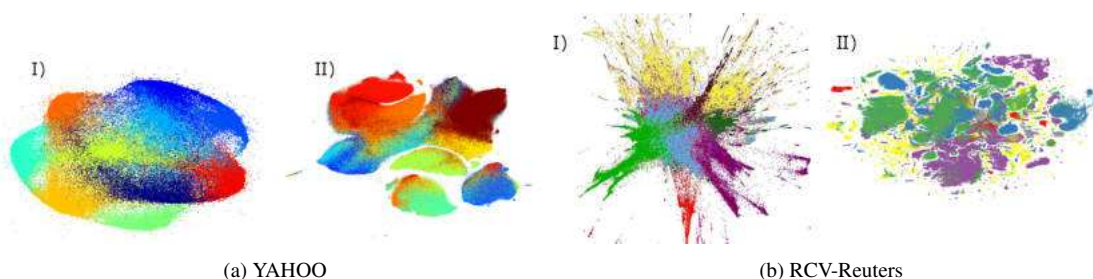


Figure 5.2: Visualization of datasets using: I) IVHD-CUDA, II) AtSNE-CUDA.

3. IVHD applied to Fashion-MNIST (see Fig.5.1c) creates separate clusters of various shapes, mainly elongated. Moreover, the generated mapping is fuzzy. It is clearly reflected in the decreasing values of cf_{nn} , particularly for $nn = 100$. However, both AtSNE-CUDA and BH-SNE-CUDA are much better at creating oblate and clearly separated clusters. As shown in Table 5.1, the values cf_{nn} are more stable than those for the IVHD method. Nevertheless, unlike the IVHD result, some classes reproduced by the baseline methods are mixed and fragmented.
4. Similar conclusions can be drawn by visualizing the RCV and YAHOO 2D embeddings (see Figs.5.2a, 5.2b, respectively). IVHD-CUDA generates more fuzzy output than AtSNE-CUDA, but the samples from the same class are closer together than those generated by the baseline method. Meanwhile, AtSNE-CUDA is capable of separate clearly visible but fragmented clusters. BH-SNE-CUDA was too slow to visualize these large datasets in a reasonable time budget.

As shown in Table 5.1, IVHD-CUDA is the fastest method for the chosen baseline datasets. Unfortunately, we were not able to control the run-times of AtSNE-CUDA and BH-SNE-CUDA implementations; thus comparisons for various time budgets are not possible. As shown above, the quality and fidelity of the embeddings strongly depend on the structure of a specific dataset, user expectations, and their data visualization requirements. In general, the local structure of the *source* data is better reconstructed by the two SNE baseline algorithms. This is not a surprise because the IVHD algorithm does not use the correct ordering of k NN neighbors and original distances between a sample y_i and its k NN neighbors. Moreover, the value of k (nn in

Table 5.1: The precision values cf_m for various data sets and embedding methods. The timings show the overall embedding time ($time$), k NN-graph generation time ($time_{gg}$), net embedding time ($time_{emb}$). The results are averages over 5-10 simulations.

Dataset	Algorithm	$time$ [s]	$time_{gg}$ [s]	$time_{emb}$ [s]	cf_2	cf_{10}	cf_{100}
MNIST	BH-SNE-CUDA	32.588		26.775	0.94	0.938	0.933
	AtSNE	15.980	5.813	10.167	0.944	0.943	0.938
	IVHD-CUDA	7.326		1.261	0.946	0.936	0.924
FMNIST	BH-SNE-CUDA	32.913		26.179	0.757	0.755	0.738
	AtSNE	17.453	6.734	10.719	0.76	0.757	0.737
	IVHD-CUDA	8.177		1.443	0.767	0.726	0.670
Small NORB	BH-SNE-CUDA	38.673		23.161	0.944	0.919	0.745
	AtSNE	20.521	15.517	5.009	0.97	0.94	0.73
	IVHD-CUDA	16.151		0.634	0.936	0.921	0.828
RCV-Reuters	BH-SNE-CUDA	-		-	-	-	-
	AtSNE	220.39	45.302	175.088	0.82	0.82	0.818
	IVHD-CUDA	60.72		15.418	0.835	0.828	0.803
YAHOO	BH-SNE-CUDA	-		-	-	-	-
	AtSNE	628.63	52.930	575.7	0.686	0.686	0.686
	IVHD-CUDA	70.12		18.930	0.668	0.662	0.653
Amazon 20M	BH-SNE-CUDA	-		-	-	-	-
	AtSNE	-	11060	-	-	-	-
	IVHD-CUDA	12218		1158	0.524	0.522	0.518

this dissertation) is extremely small compared to AtSNE-CUDA and BH-SNE. However, despite this drastic approximation, IVHD properly preserves the class structure and its relative locations. For fine-grained structures of classes (such as in the Small NORB dataset), IVHD outperforms its competitors in both efficiency and, slightly, in the cf accuracy. Moreover, unlike AtSNE-CUDA and BH-SNE-CUDA, IVHD can visualize separated and not fragmented classes. The same can be observed for the highly imbalanced RCV dataset. For MNIST and Fashion MNIST datasets, the dominance of t-SNE-based methods in reproducing the local "microscopic" data structure is visible, mainly due to the strong "crowding effect" seen for IVHD embeddings. However, the "macroscopic" view of the classes is more convincing for IVHD visualization, although due to the high *perplexity* parameter ($perplexity = 50$), the baseline algorithms are also tuned for better coarse-grained visualization.

The main advantage of IVHD is that after storing the k NN-graph in the disc cache and neglecting the computational time required for its generation, the embedding of IVHD can be more than one order of magnitude faster than the baseline methods. This allows for a very detailed interactive exploration of multiscale data structure by employing broad spectrum of parameter values and various versions of the stress function without the need for the k NN-graph recalculation.

As shown in Fig. 5.3, IVHD was the only method that was able to generate visualizations for the Amazon20M dataset (and its subsets) in a reasonable period of

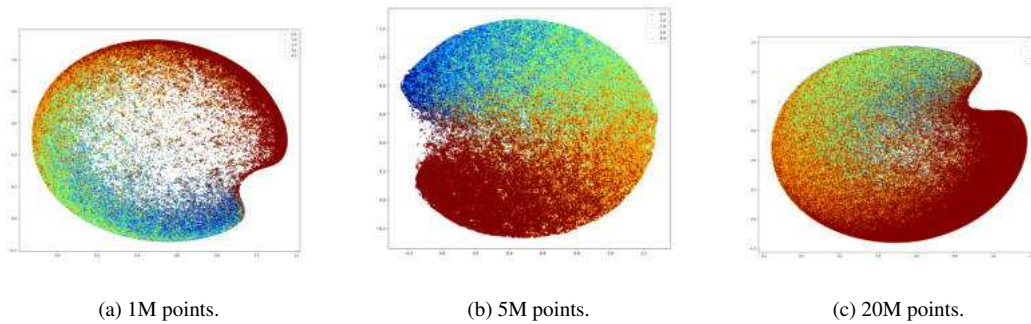


Figure 5.3: Visualization of Amazon dataset using IVHD-CUDA.

time. One can see a very clear separation of the 5 classes that appear in this dataset, which contains book reviews from the Amazon platform. Other methods (both those implemented in the CPU and GPU environments) ended with errors or were unable to generate visualizations within 12h, so that the calculation process was interrupted and canceled. This shows that IVHD is the only method that can actually deal with very large datasets containing real-world data in a very fast time using minimal resources, confirming that it is the most efficient and memory-saving method for visualization of high-dimensional data.

CHAPTER 6

META-PLATFORM FOR SUPERVISED VISUALIZATION

Supervised learning is a subcategory of machine learning and artificial intelligence. Its goal is to estimate statistical relationships using labeled training data. It has enabled a wide variety of basic and applied discoveries, from the discovery of biomarkers in omics data [142] to the recognition of objects in images [72]. Dimensionality reduction itself is often used as a preliminary step in data mining, both as a pre-processing for classification or regression and for visualization. Most dimensionality reduction techniques to date are unsupervised and do not consider class labels (e.g., PCA [60], MDS [39], t-SNE [86] and its variants [114, 12, 42], Isomap [134], LargeVis [131]). These methods require large amounts of data and are often sensitive to noise, which can hide important data features. Various attempts at supervised dimensionality reduction methods that take into account auxiliary annotations (e.g., class labels) have been successfully implemented to increase classification accuracy or improve data visualization. Many of these supervised techniques consider labels as a loss function in the form of similarity or dissimilarity matrices, thus creating an undue emphasis on separation between class clusters, which does not realistically reflect local and global relationships in the data. Additionally, these approaches are often sensitive to parameter tuning, which can be difficult to configure without a clear quantitative notion of visual superiority. As for UMAP [90, 119], it can be used for standard unsupervised dimension reduction, but the algorithm itself offers significant flexibility, allowing it to be extended to perform other tasks, including making use of categorical label information to perform supervised dimension reduction and even metric learning.

In this chapter, we describe a novel meta-platform for supervised visualization, which allows each visualization method to be considered as supervised.

Meta-platform for supervised high-dimensional visualization

The idea is based on the use of centroids obtained using any clustering method (such as, for example, k-means) (Fig. 6.2). We generate two types of centroids: 1) local (intra-class) and 2) global (for the entire dataset). Using the centroids thus obtained, parallel layers of neurons of both types are formed. Feed-forward of such a network

boils down to calculating the Euclidean distances of a given sample with respect to the centroids to which it is "closest" (you can also use here any classification algorithm, e.g. SVM, which will classify the sample into the appropriate class), and then the activation function is applied. What we achieve in this way is a trained classifier against which we will embed a given sample with the appropriate set of centroids. After this embedding, any method can be used for data visualization, and as a result, we obtain a supervised variation of this method.

Experiment with random dataset. A random 100-D dataset was generated and visualized with t-SNE and PCA. Before visualization, we embedded the data set using a different set of centroids obtained with the k-means method.

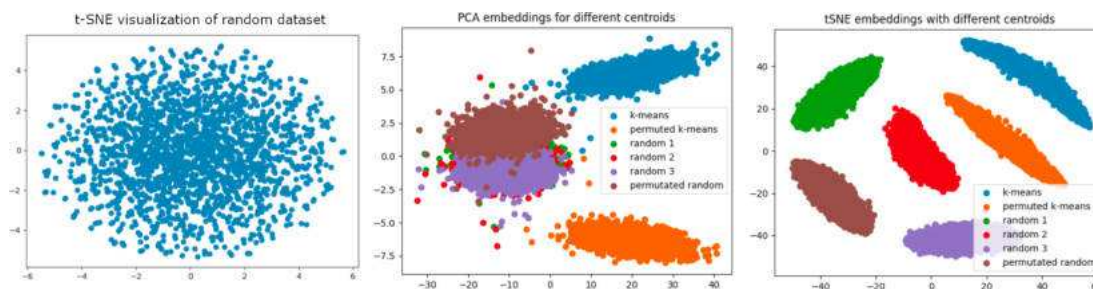


Figure 6.1: Embedding dataset with different set of centroids. Left: random 100-D dataset visualized with t-SNE. Middle: random dataset embedded with the centroids, visualized with PCA. Right: random dataset embedded with the centroids, visualized with t-SNE.

As shown in Fig. 6.1, embedding with different sets of centroids causes clusters to separate, when visualized with the PCA or t-SNE method. This is because a dataset embedded with different sets of centroids actually has different bases, and hence the clusters are separate in space. This introduces the need for global centroids (which will provide a common frame-of-reference in the embedding).

Experiment with artificially generated dataset. A 10-D dataset with overlapped classes. There are 5 classes with 500 points each. As a classification method, we use a "perfect classifier" (actual labels). We do not want the results to be distorted by the classifier.

As shown in Fig. 6.3, the dataset contains "X shapes" manifolds immersed in a 10-dimensional space. The overall structure of the dataset can be observed by plotting the first 3 dimensions (Fig. 6.3a). We specifically wanted certain classes to overlap to add realism to the dataset.

Based on Fig. 6.4, the following conclusions could be drawn:

1. For embedding based on distances from centroids belonging to a single recognized class (Fig. 6.4a), the classes are separated and retain their internal structure. However, the global structure has nothing to do with their original arrangement.
2. For embedding based on distances from centroids belonging to the entire dataset (Fig. 6.4 b), the effect is similar to that of pure t-SNE. Still, sizable areas overlap, global structure is visible, and differences are due to the way the k-means method finds clusters (less emphasis on dense areas).

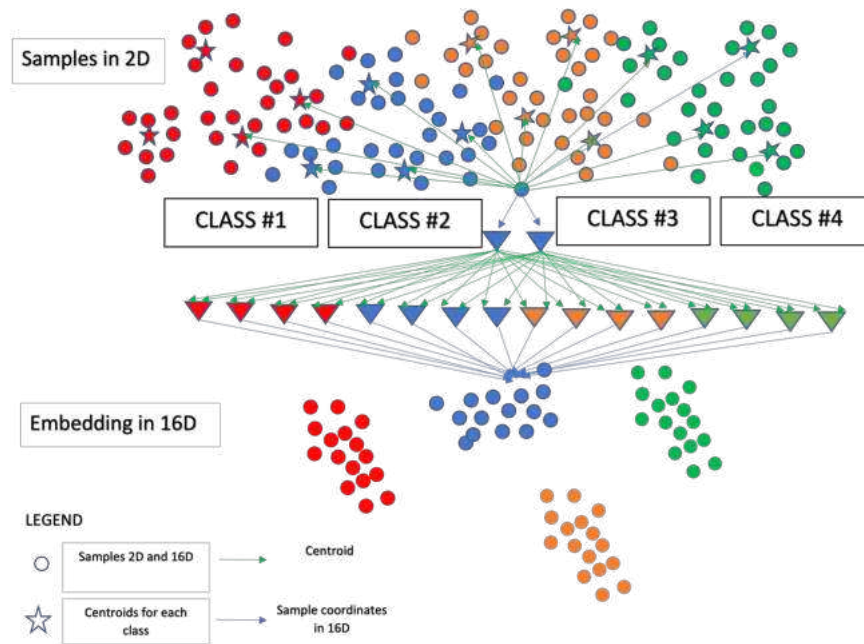


Figure 6.2: General idea of meta-platform for supervised visualization, which leverages centroids.

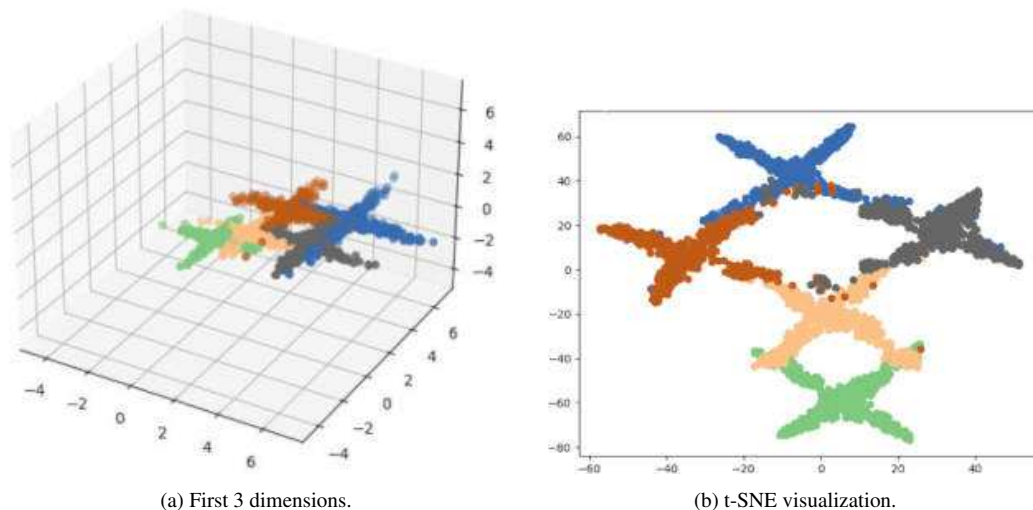


Figure 6.3: Artificially generated "X shapes" dataset (first 3 dimensions) and its t-SNE visualization.

3. For embedding based on hybrid distances (Fig. 6.4c), visualization preserves the separation of classes, intra-class structure, and also their mutual position in space (both the layout and the fact that yellow and green are closer together than the others).

Experiment with local and global centroids. Knowing how the DR quality and KNN gain metrics work 4.1, the following observations (which coincide with theory and intuition) can be made:

1. With an increase in the number of global centroids, the DR quality should shift to the right (better properties for a bigger neighborhood), and the KNN gain should

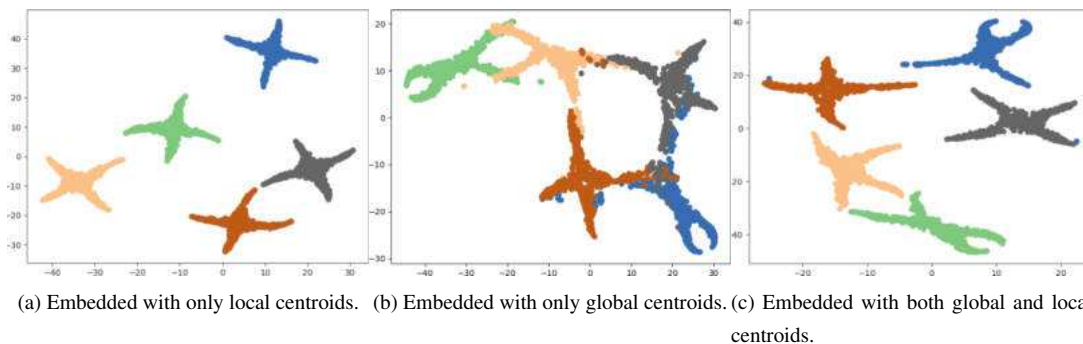


Figure 6.4: t-SNE visualization of artificially generated dataset "X shapes", which was preprocessed with different combination of centroids before the visualization.

be lower in a larger part of the "neighborhood size" spectrum. KNN gain gives us information on how the KNN classifier behaves when trained on the embedded dataset. It is intuitive that the classifiers' accuracy should decrease when data is embedded in a more global fashion. Both effects can be observed in the first row of Fig. 6.5.

2. However, with an increase in the number of local centroids, the DR quality should shift to the left (worse properties for a bigger neighborhood), yet the KNN gain should be higher in a larger part of the "neighborhood size" spectrum. There is less global information when more local centroids are used, so the separation of clusters is becoming more emphasized; thus, the KNN classifier used for metric calculation is obtaining data that are much more "tweaked" to its class separation aspect. Furthermore, both effects can also be observed (in the second row of Fig. 6.5).
3. When comparing Fig. 6.4 c and Fig. 6.5 a, it is also clear that the global geometry of the visualization improved. When analyzing the green class, it is next to the beige class (that is, between gray and brown), and the farthest class is blue. Exactly as in Fig. 6.3b, but with better class separation.

Experiment with mid-scale dataset. To verify how this idea behaves with realistic benchmark data, we applied it to the MNIST and FMNIST datasets. To show that it can be used for any unsupervised method, we visualized it with one that is implemented in a GPU environment (AtSNE [42]), one that is implemented in a standard CPU environment t-SNE and IVHD. To create supervised variants, we used a set-up with 100 global centroids and 100 local centroids. As shown in the following, in Figs. 6.6 and 6.8, in both cases, IVHD responds in the most extreme way, separating classes very effectively. This is a consequence of the fact that IVHD is a method derived from MDS, which in turn is a method that seriously takes into account the aspect of preserving the global structure of the entire data set.

Regarding performance, generating the set of $100gc$ and $100lc$ and then embedding the MNIST/FMNIST dataset using these centroids is a computational mark-up of about 1 minute (in the local setup described in Section 4.4). It is worth to mention that no

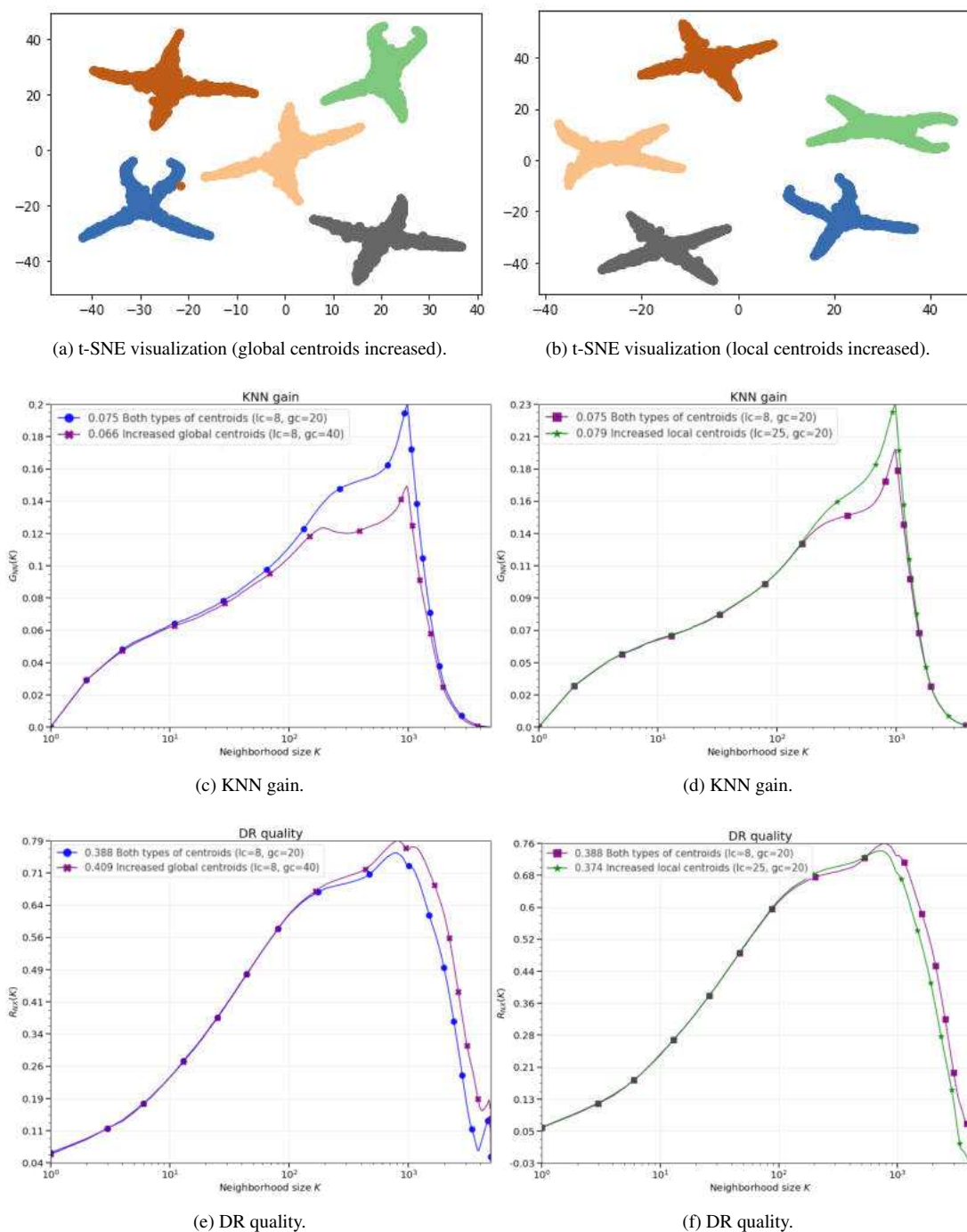


Figure 6.5: t-SNE visualization, DR quality and KNN gain of artificially generated dataset "X shapes". In the first column (on the left), we increased the number of global centroids (2 times) in the overall combination of centroids (Fig. 6.4 c). In the second column (on the right), we increased the number of local centroids (4 times) in overall centroid combination.

optimization was performed and basic Python methods were used for testing, so there is a huge room for improvement and speed-up (although it was not needed at this stage of the research).

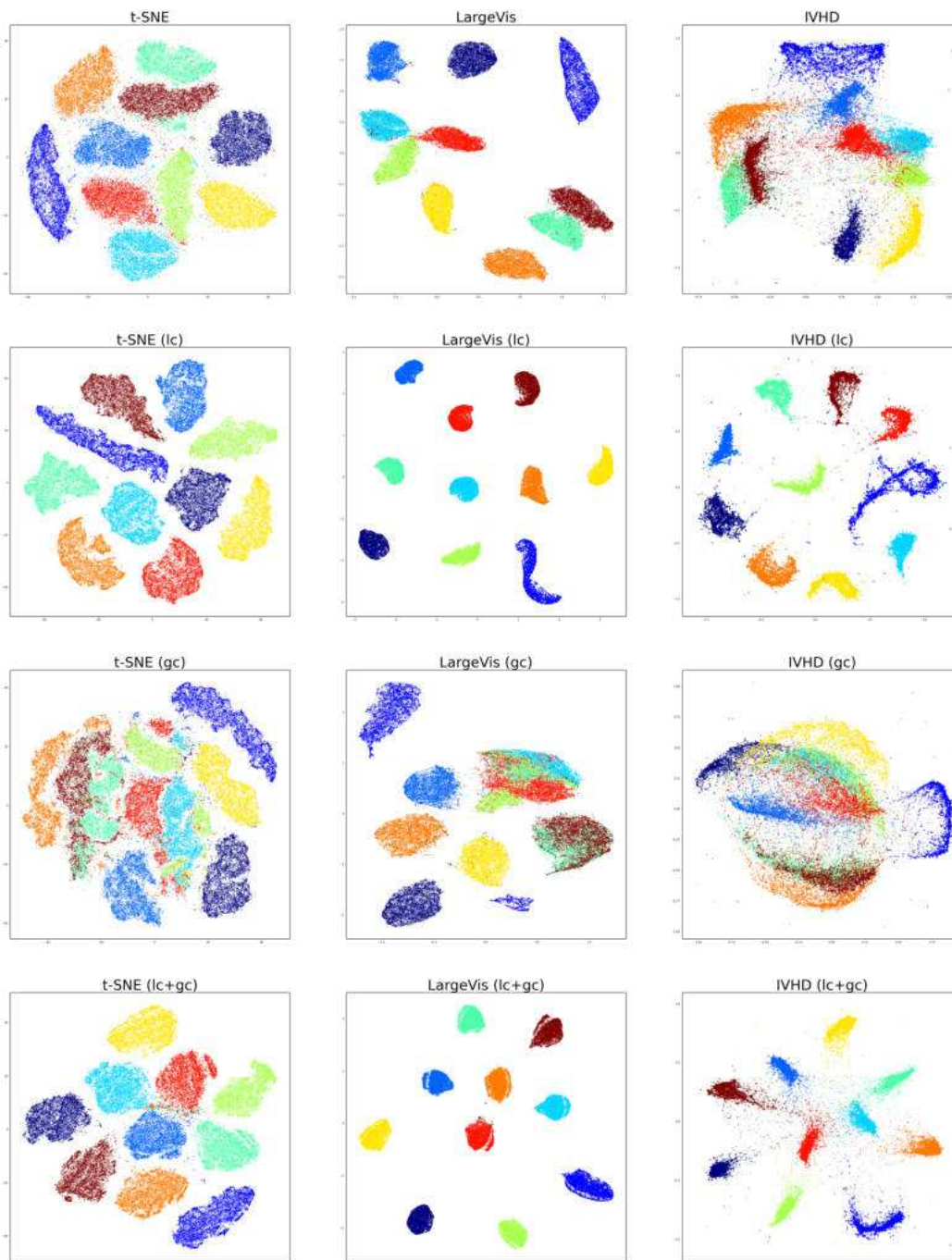


Figure 6.6: Method comparison on MNIST dataset for both unsupervised and supervised visualization variants. In next rows of the figure we observe: 1) default visualization of the unsupervised variant of the method, 2) supervised variant, that uses only local centroids, 3) supervised variant, that uses only global centroids and 4) 2) supervised variant, that uses both local and global centroids to perform embedding.

It should be noted that the supervised variant is not quite the type of method that is desirable for this type of data analysis. Much more preferable are methods that operate in unsupervised fashion and are efficient enough to analyze datasets in real time (which, as will be shown in the next chapter, can be useful for inspection and interpretation

of deep neural networks). The reason for this is that much of the data on which one operates in machine learning is unlabeled and its internal structure is unknown. Then the supervised methods become useless.

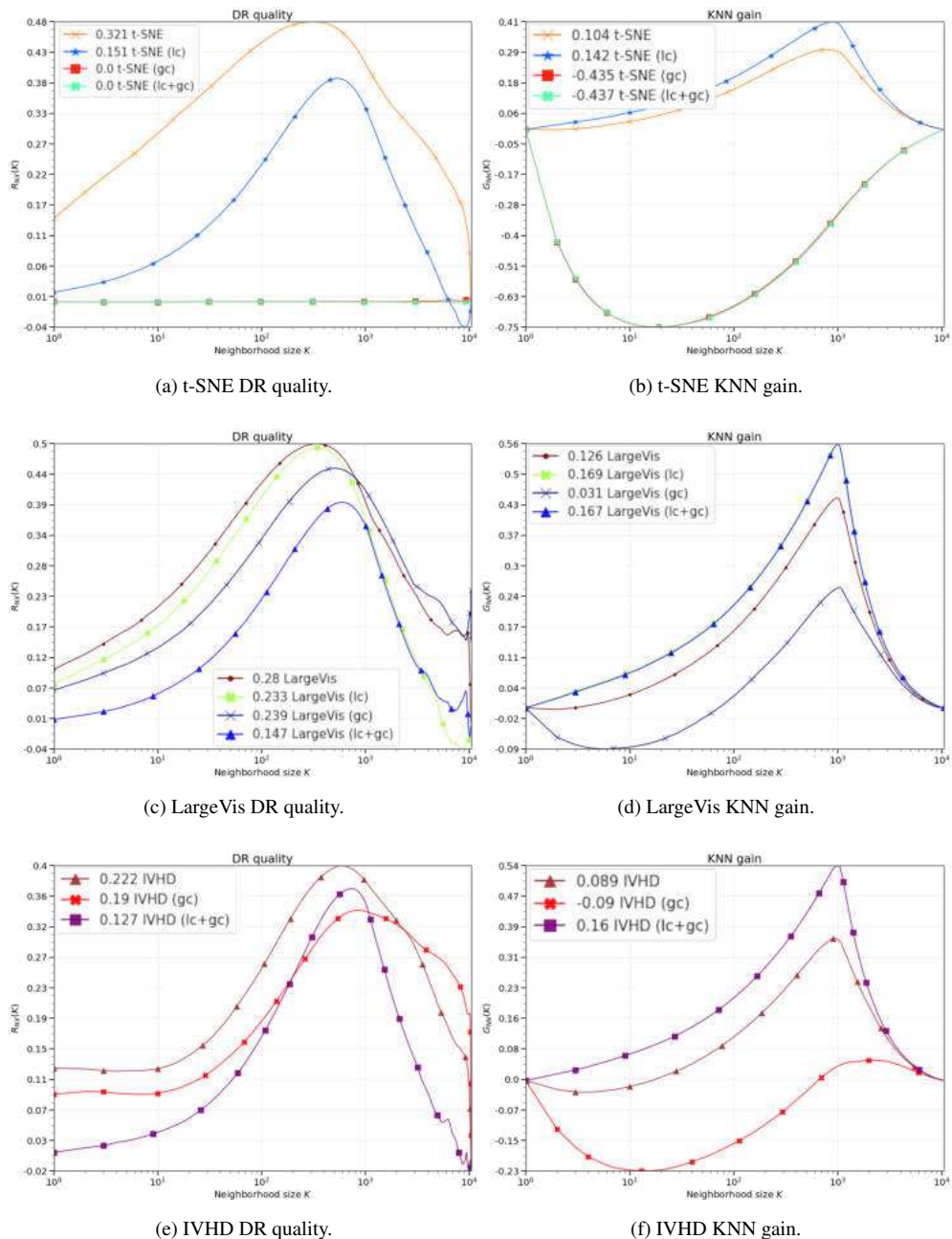


Figure 6.7: Metrics obtained for MNIST dataset for both unsupervised and supervised visualization variants for different methods.

For both sets of metrics obtained (Figs. 6.7 and 6.9), an increase in KNN gain and a slight decrease in DR quality are observed (analogically to Fig. 6.5). It shows that we can use this meta-platform for converting any algorithm from unsupervised to its supervised variant. To verify this thesis, we also visualized the centroid-embedded dataset with AtSNE implemented in the AtSNE CUDA environment (Appendix B).

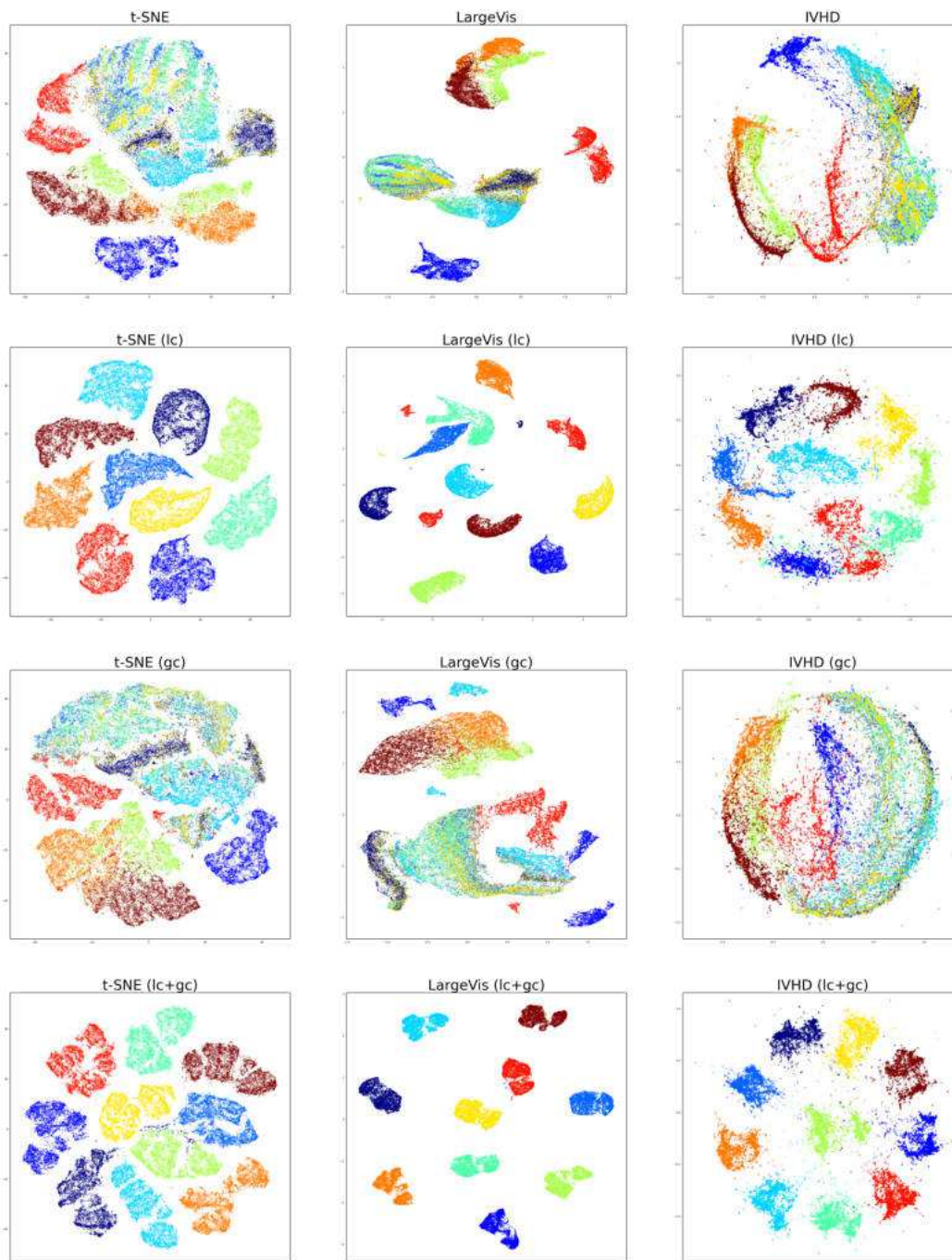


Figure 6.8: Method comparison on FMNIST dataset for both unsupervised and supervised visualization variants. In next rows of the figures we observe: 1) default visualization of the unsupervised variant of the method, 2) supervised variant, that uses only local centroids, 3) supervised variant, that uses only global centroids and 4) 2) supervised variant, that uses both local and global centroids to perform embedding.

The most important feature of the created meta-platform is that it is highly abstracted. We can use arbitrary classifier and unsupervised variant of the embedding method and obtain a supervised-like process of visualization in 2-D (or 3-D) space.

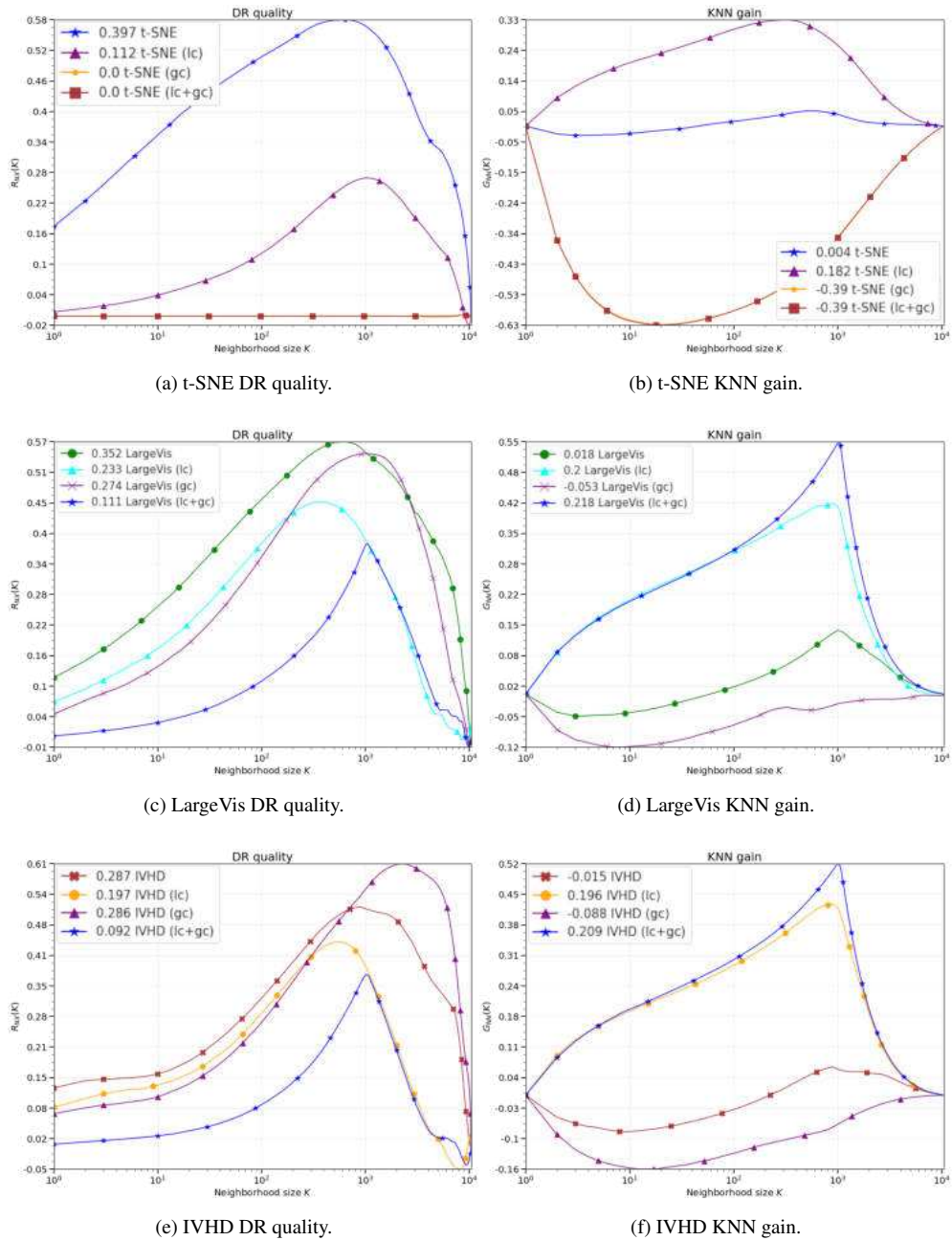


Figure 6.9: Metrics obtained for MNIST dataset for both unsupervised and supervised visualization variants for different methods.

CHAPTER 7

APPLICATION OF HDD VISUALIZATIONS IN ANNS INSPECTION AND INTERPRETATION

In the field of machine learning, advances in computational power and techniques for building and training artificial neural networks (ANNs) have allowed these models to achieve state-of-the-art performance in many pattern recognition applications [123]. However, effective ANN training is usually time-consuming and requires considerable knowledge [7]. Additionally, such networks are typically used as black boxes and it is widely believed that they learn high-dimensional representations of the original observations. In this chapter, we demonstrate the potential of DR techniques to provide insightful visual feedback considering ANNs. Although we focus on multilayer perceptrons and conventional neural networks, this approach is extensible to other types of network (e.g., LSTM or Elman networks [14]). In particular, proposed visualizations address the following two tasks [110]: 1) exploring the relationships between alternative representations of observations *learned* by ANNs and 2) exploring the relationships between artificial neurons. The projection-based visualization approach proposed in (1) can be found in the machine learning literature [28, 95, 129, 96, 55]. Using approach (2), which is related to techniques developed for feature-space exploration, we use projections to represent similarities between artificial neurons. It is a novel approach to visualizing the relationships between artificial neurons and classes, which was proposed in [110].

7.1. Experimental protocol

Visualization approach that we use is based on hidden-layer activations extracted from a network trained for a given dataset and can be divided into two parts: creating projections of these activations (1) and depicting the relationships between the neurons that originate these activations (2). It is largely inspired by the work presented in [110].

Datasets We include three well-known image classification benchmarks: MNIST [74], CIFAR-10 [71], and SVHN [101]. All datasets are described in detail in the Appendix A.

Neural networks architectures of two types were considered:

1. Multilayer perceptron (MLP): 3072 (784, for MNIST) input neurons, followed by four rectified linear hidden layers of 1000 neurons each. The output layer is softmax with 10 neurons. Dropout [129] is applied from the first hidden layer, increasing from 0.2 to 0.5 in steps of 0.1 per layer.
2. Convolutional neural network (CNN): $32 \times 32 \times 3$ input image ($28 \times 28 \times 1$, for MNIST), followed by a convolutional layer with $32 \ 3 \times 3 \times 3$ (or $3 \times 3 \times 1$) filters, a convolutional layer with $32 \ 3 \times 3 \times 32$ filters, a 2×2 max-pooling layer (dropout 0.25), a convolutional layer with $64 \ 3 \times 3 \times 32$ filters, a convolutional layer with $64 \ 3 \times 3 \times 64$ filters, a 2×2 max-pooling layer (dropout 0.25), a fully connected layer with 4096 (or 3136) neurons (dropout 0.5), a fully connected layer with 512 neurons and a softmax output layer with 10 neurons. All fully convolutional and connected layers (except the output) are rectified linear.

Although larger models (in number of layers and parameters) are used for certain difficult classification tasks, the architectures sketched above are fully realistic, typical for image classification tasks, and sufficiently complex to warrant exploration.

Training is performed by mini-batch stochastic gradient descent based on momentum. For MLPs, the batch size is 16, the learning rate is 0.01, the momentum coefficient is 0.9, and the learning decay is 10^{-9} . For CNNs, the batch size is 32, the learning rate is 0.01, the momentum coefficient is 0.9, and the learning decay is 10^{-6} . The initial weights of a neuron in layer l are sampled from a uniform distribution in $[-s, s]$, where $s = [6/(N^{(l-1)} + N^{(l+1)})]^{1/2}$, and the biases start at 0. We manually chose these hyperparameters, together with the architectures mentioned above, based on cross-validation using the predefined validation sets and experiments performed in [110].

Activations for a given layer, the subject of our analysis, are extracted for a random subset of 2000 observations from the test sets, strictly to facilitate visual presentation. This subset is always the same for a given dataset. In two cases, we also extract activations from a random subset of a training set. For CNNs, we only extract activations from fully connected layers.

Projections are created using a fast (approximate) implementation of t-distributed Stochastic Neighbor Embedding (t-SNE) and IVHD. We chose these techniques on the basis of their widespread popularity (t-SNE) and computational speed (IVHD). Both have the proven ability to preserve neighborhoods and clusters in projections (Chapter 4).

To carry out the research demonstrated in this section, we used Python, Keras, Theano [4], NumPy, scikit-learn [107], and our methods for feature space exploration. For visualization, we used the Python interface provided with the IVHD method, which made it possible to use all components of this study in one environment.

7.1.1. Exploring the relationships between alternative representations of observations learned by ANN

Inter-layer evolution: Let $\mathbb{A}\{1\}, \dots, \mathbb{A}\{T\}$ be a sequence of sets of (high-dimensional) activations, where each activation $a\{t\} \in \mathbb{A}\{t\}$ originates from the same observation as a single activation $a\{t+1\} \in \mathbb{A}\{t+1\}$. One way to visualize the evolution in such sets of activations is by reducing the dimensionality, applied in such a way that changes in the resulting projections reflect changes in the corresponding high-dimensional data [65]. This can be done by creating a projection $\mathbb{A}_p\{t\}$ for each activation set \mathbb{A}_t . When doing this, it is essential to eliminate the variability between projections that does not reflect changes in the high-dimensional data.

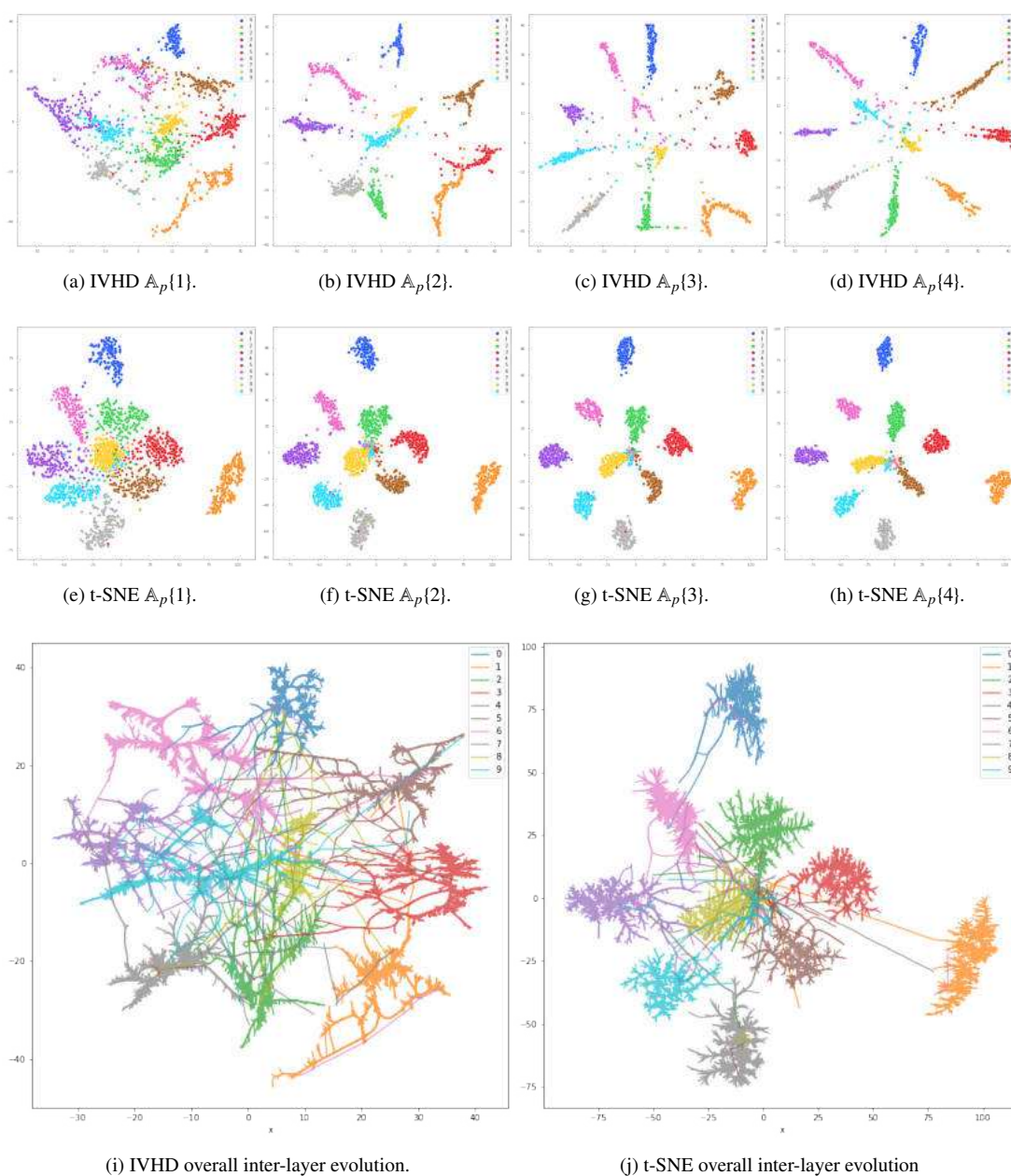


Figure 7.1: Inter-layer evolution of MLP after training using MNIST dataset.

It is important to remember that dimensionality reduction techniques, including t-SNE and IVHD, often produce large changes in global visual cluster placement for small data changes, which registration cannot eliminate [43]. For iterative techniques, an intuitive solution is to initialize the positioning in $\mathbb{A}_p\{t+1\}$ with the previously calculated $\mathbb{A}_p\{t\}$. In work [110], authors verified that this is a poor alternative, as it significantly biases the projection sequence to show evolution due to initialization in a (presumably) better state with respect to the optimization goal. Instead, a simple strategy is employed: calculate a projection (randomly initialized) \mathbb{A}_p of $\mathbb{A}\{1\}$ using t-SNE and use \mathbb{A}_p to initialize each $\mathbb{A}\{t\}$.

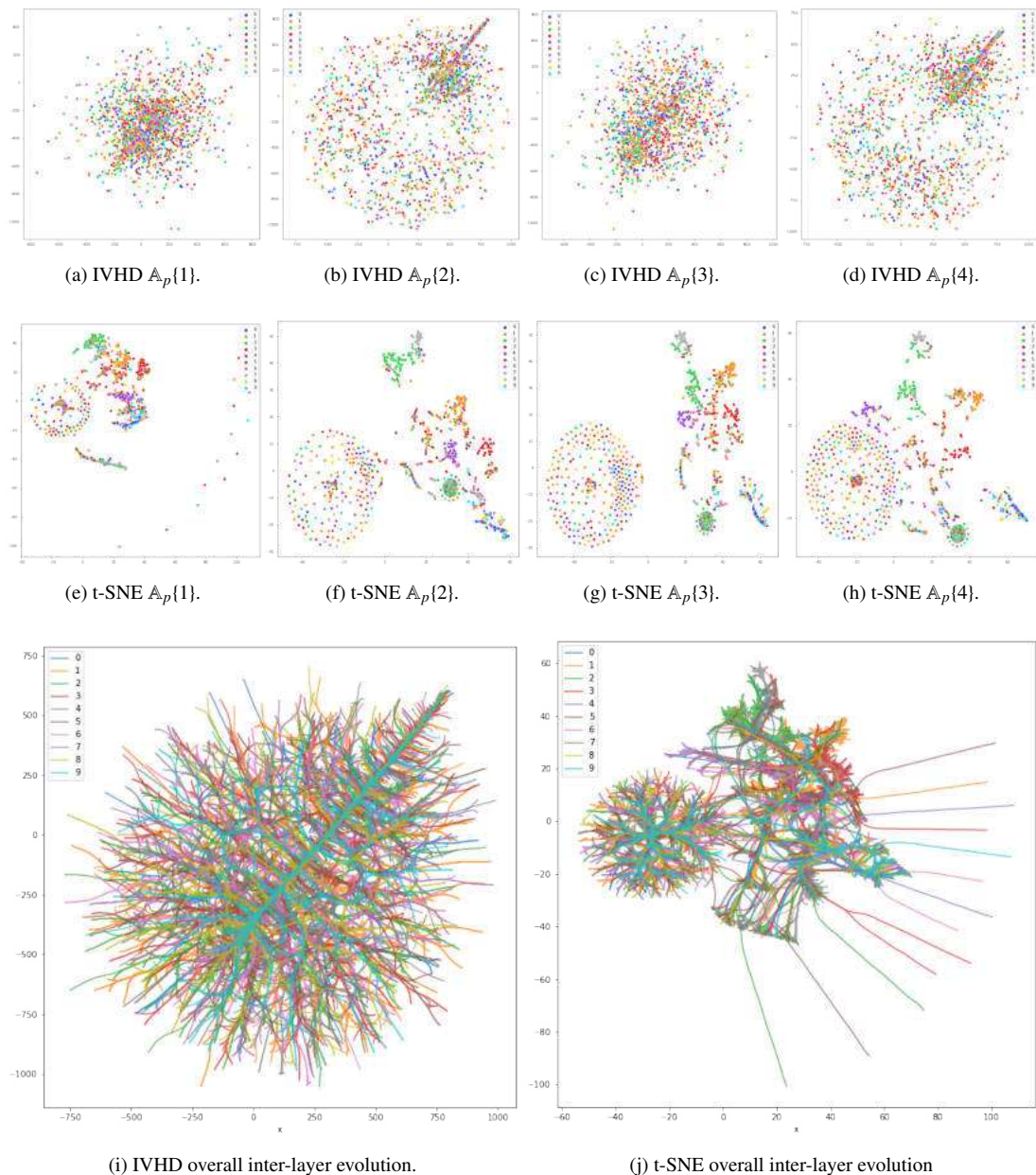


Figure 7.2: Inter-layer evolution of MLP after training using SVHN dataset.

Figures 7.1, 7.2, 7.3 present the evolution between layers for the test subsets (after training). The merged image summarizes a sequence of four projections, one per hid-

den layer, shown as thumb nails. The colors of the paths encode the classes, we can also observe how the activation data "flows" through the four layers of the network, by comparing next-layer projections (and overall inter-layer evolution). For MNIST, the beam shapes show that the visual clusters are quite stable in all layers. Thus, the network has reached a fairly good separation between classes in the early layers. The gradients also show that some visual clusters become more compact in later layers and that some clusters move away from others (e.g., the brightness pattern in the purple cluster).

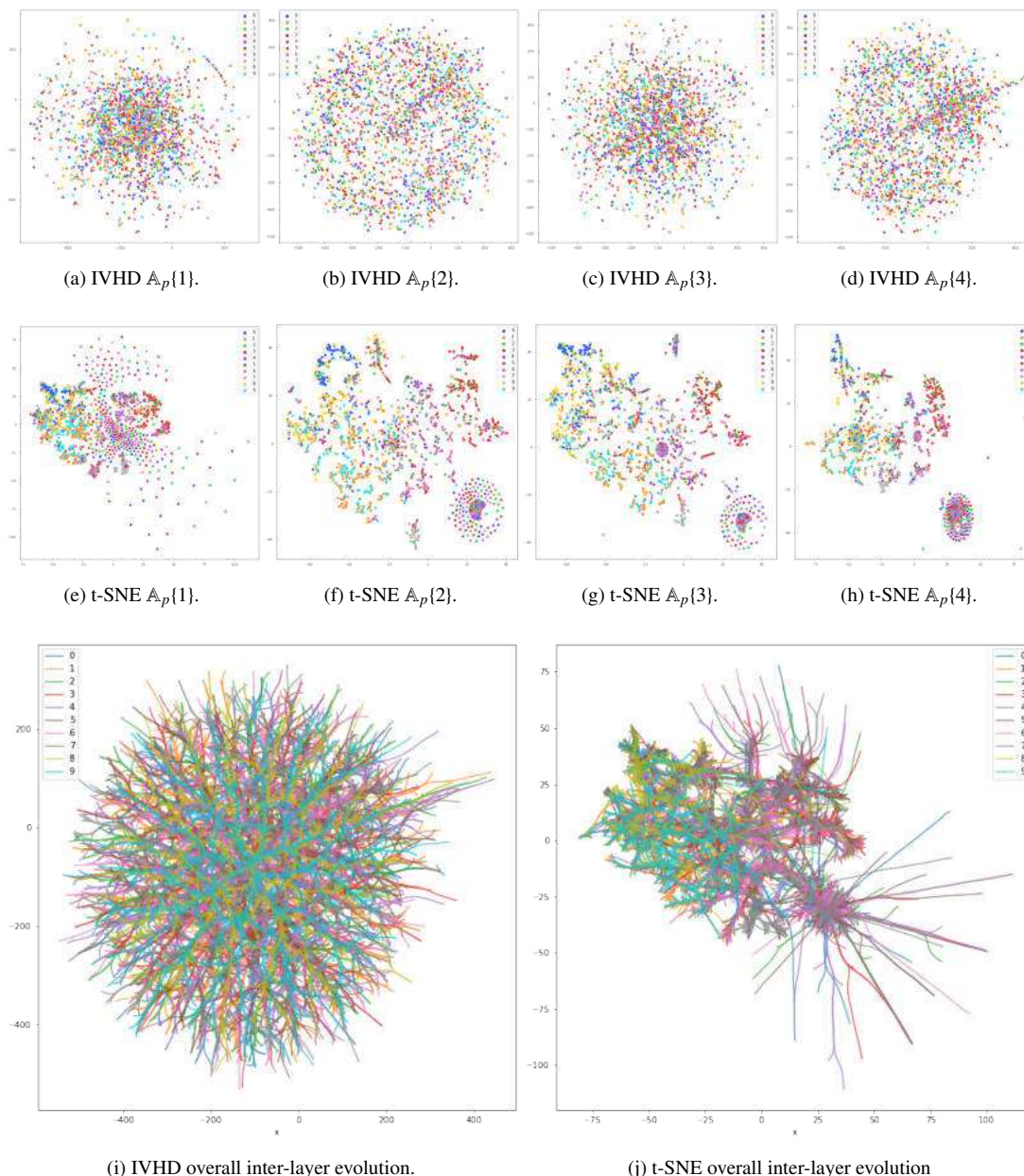


Figure 7.3: Inter-layer evolution of MLP after training using Cifar-10 dataset.

Inter-epoch evolution: The same idea as described above can be employed to visualize inter-epoch evolution. However, as shown in [110], the results in this case are significantly more difficult to interpret. This is due to a combination of large changes

in the very first epoch, high intra-visual-cluster variance between epochs, and a much larger number of frames (typically hundreds) to be summarized. For this reason, different strategy is used to visualize inter-epoch evolution. Consider again the sequence $\mathbb{A}\{1, \dots, \mathbb{A}\{T\}$ of the activation sets. For inter-epoch evolution, $\mathbb{A}\{t\} \subset \mathbb{R}^k$ for a fixed k , for all t . Therefore, we can create a projection for the set $\cup_k \mathbb{A}\{t\}$, which contains activations for all epochs. As we compute a single projection, there is no spurious inter-frame variation. Figure 7.4 shows the evolution between epochs for the last CNN hidden layer activations using this strategy, from epochs 0 to 100, in steps of 20 (12K points in total). It is interesting to note how the dimensionality reduction technique placed the points corresponding to earlier epochs in the center of the projection, considering that it does not explicitly receive this information.

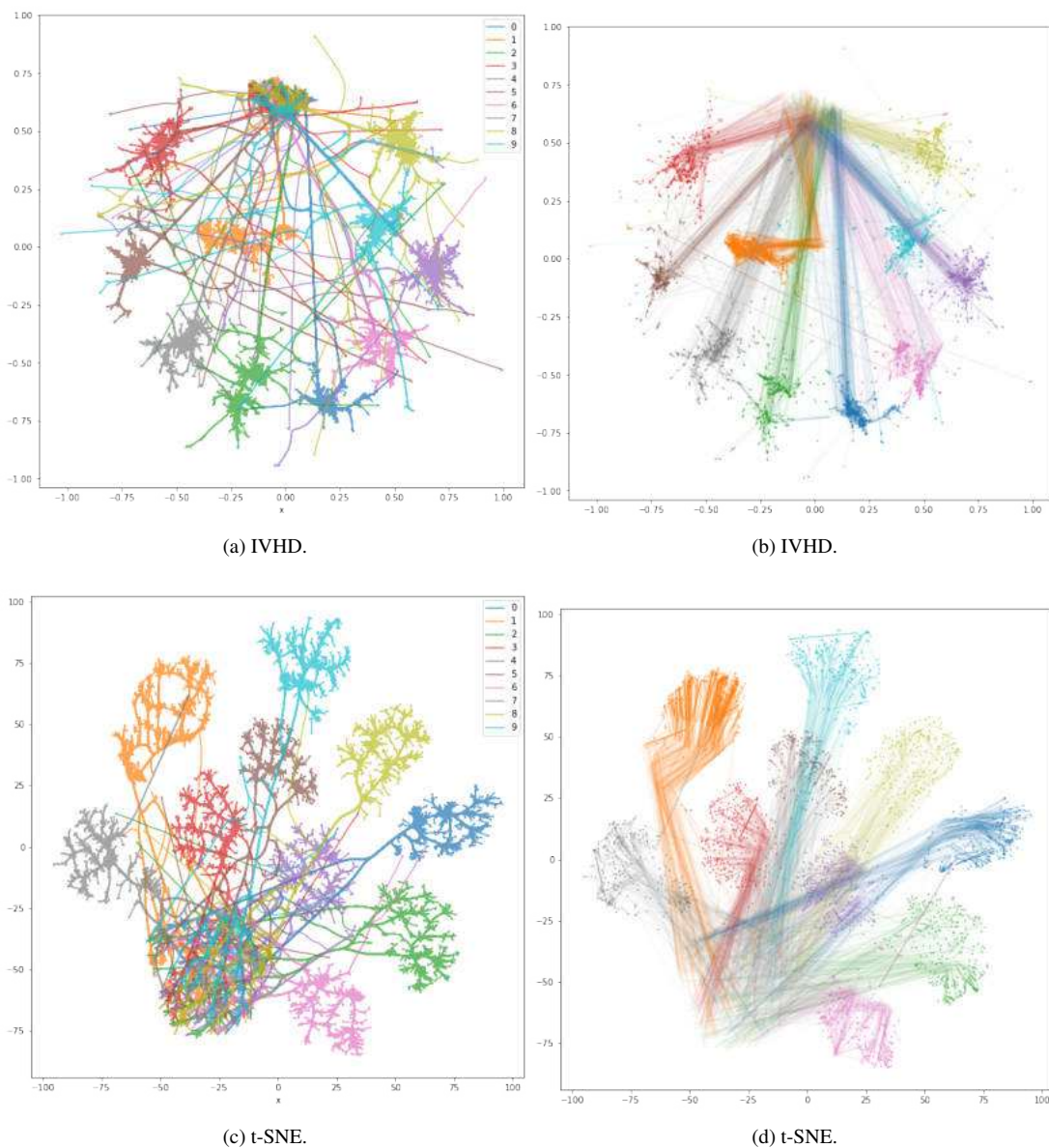


Figure 7.4: Inter-epoch evolution, last CNN hidden layer. On the right side: trace plot, where trail parts show later epochs and the points are showing the final position obtained in the last epoch.

7.1.2. Exploring the relationships between artificial neurons

As in [110], we introduce a projection of a neuron, where a point represents a neuron. Points are placed in 2D on the basis of the similarity between neurons. The dissimilarity $d_{i,j}$ between neurons i and j is defined as $d_{i,j} = 1 - |r_{i,j}|$, where $r_{i,j}$ is the empirical correlation coefficient (Pearson) between neurons i and j in a dataset composed of layer-1 activations (recall that each element of an activation vector is a neuron output). This metric captures both positive and negative linear correlations between pairs of neurons. From the pairwise dissimilarity matrix, a projection is computed using multidimensional scaling (MDS [39]). Although t-SNE is particularly concerned with preserving neighborhood relationships [86], MDS attempts to preserve global pairwise dissimilarities as much as possible, which is more appropriate in this scenario.

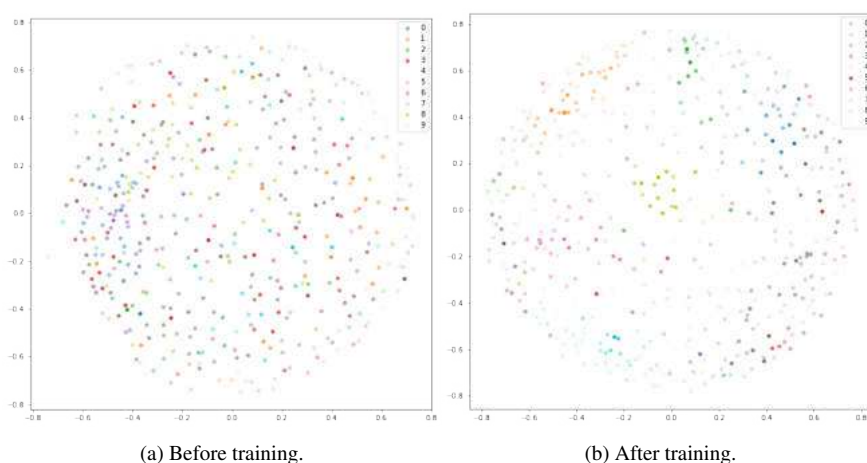


Figure 7.5: Activation and neuron projections of last CNN hidden layer activations before (on the left) and after (on the right) training. Neuron projection colors show the neurons’ power to discriminate class 0 vs rest.

Figure 7.6 shows the discriminative neuron map for the SVHN test subset, the last activations of the hidden layer, after training. The presence of compact visual clusters shows how the entire set of neurons can be (almost) partitioned into groups with related discriminative functions (specializations), even though the neuron projection is created without any class information. Neuronal activation and projections can be combined to elucidate the role of particular neurons.

In Figure 7.7 the discriminative neuron map for the MNIST test subset is presented (from the last activations of the hidden layer, after training). We can clearly see which neurons are associated to which class. For MNIST dataset, the results presented are the most intuitive and informative.

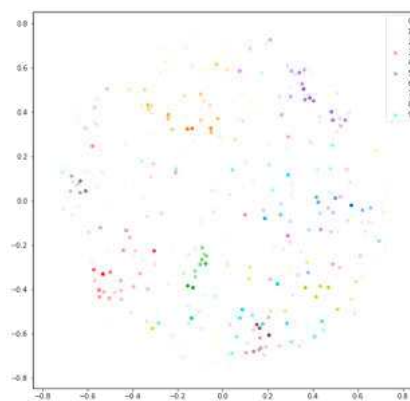


Figure 7.6: Discriminative neuron map of last CNN hidden layer activations after training, SVHN test subset.

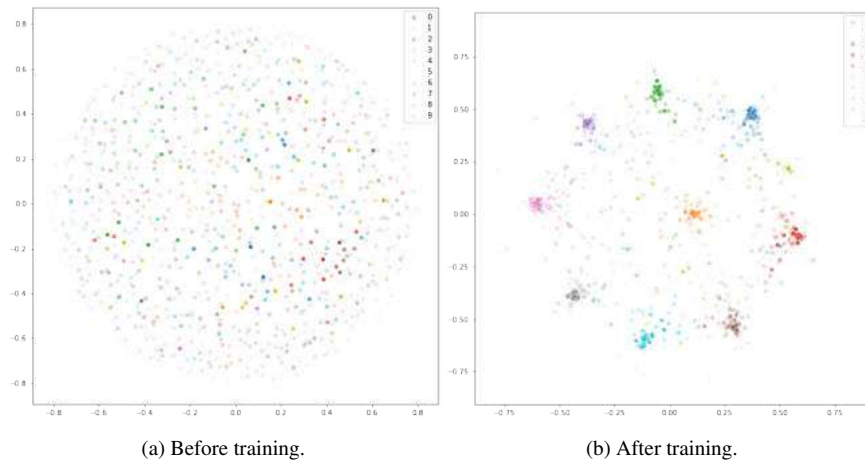


Figure 7.7: Activation and neuron projections of last MLP hidden layer activations before (on the left) and after (on the right) training. Neuron projection colors show the neurons' power to discriminate class 0 vs rest.

In this chapter, we show how dimensionality reduction can be used to visualize the relationships between learned representations (1) and between neurons (2) in artificial neural networks. Concerning the first task, the visualizations support the identification of confusion zones, outliers, and clusters in the internal representations computed by such networks. Separately, it was shown how to visually track inter-layer and inter-epoch evolution of learned representations. Concerning the second task, it was presented how to inspect the relationships between neurons and classes (specialization) and similarity between neurons.

CHAPTER 8

CONCLUSIONS

In this work, we explored areas of how data embedding methods can be analyzed using a common framework and showed that *dimensionality reduction (DR)* can be perceived as visualization of unweighted k NN-graphs, constructed for the source \mathbf{X} data. The framework made it possible to verify how knowledge extracted from different DE algorithms can help to develop the most efficient and memory-saving method for interactive visualization of high-dimensional data. In our study, we focus on different types of data, starting with synthetic datasets that can provide insight into very specific properties of a method and ending with large datasets to verify the effectiveness and precision of each method.

Firstly, based on recent results posted in a paper currently under review by the Promoter, we claim that t-SNE and UMAP are in fact computationally consistent with each other. Up to minor changes, one could have presented the theoretical foundation of UMAP and implemented it with the tSNE algorithm or vice versa. Furthermore, t-SNE can be simplified step by step to the most simplistic IVHD method.

Second, we show that IVHD with binary distances is the most time&memory optimal and is capable of analyzing truly large datasets and graphs in a reasonable amount of time [32], which is not achievable by other baseline methods. For this purpose, a very extensive analysis of methods of visualizing high-dimensional data was performed. We evaluated five well-recognized in the research community data embedding algorithms: t-SNE, LargeVis, UMAP, TriMAP and PaCMAP. The results obtained suggest that while all of these methods generate visualizations at a high level (in terms of metrics such as: DR quality, k NN gain, Shepard diagram), baseline methods are significantly inferior to IVHD in terms of efficiency. Through the use of GPU-CUDA environment, we made it possible to visualize datasets containing millions of data points in a reasonable amount of time. We also proposed several improvements to the IVHD method that did not impose a computational burden, but improved its visualization efficiency. Specifically:

- use the $L1$ norm in the final steps of algorithm calculation,

- utilize the reverse neighbor mechanism in the final steps of algorithm calculation (along with the norm $L1$),
- add auto-adaptation of time step during simulation,
- we implemented IVHD in the *viskit* visualization library [92], which contains different DR and optimization methods, which creates the possibility to mix different optimizers "on fly",

For datasets that are not that large, IVHD generates visualizations whose global values do not deviate from the average achieved by other methods, and in some cases the global properties are the best (smallNORB, mammoth dataset, synthetically generated ball in a sphere). However, IVHD has two shortcomings. First, it squeezes classes (crowding) (but it also preserves the neighborhood well [34]). It can be loosened at the end of the simulation by changing the parameters or using another method locally. Second, it generates "dust" between classes. These points are the result of reaching equilibrium of the forces that affect them. One way to improve this was to add a mechanism that leaves only the interaction with reverse neighbors in the last steps of the algorithm.

In the third part of this thesis, we introduced a meta-platform that allows any DR method to be used in a supervised fashion. It is highly abstracted, allowing for an arbitrary classifier and an unsupervised variant of any embedding method to obtain a supervised-like process of visualization in 2-D (or 3-D) space. The supervised variant is not exactly the type of method that is desirable in data visualization research. Much more preferable are methods operating in an unsupervised fashion, which are powerful enough to analyze datasets in real time. This is because much of the data on which machine learning operates is unlabeled and its internal structure unknown. Of course, there is also the problem of the availability of labeled data because sets of such data can still be difficult to access.

Finally, we use our highly efficient IVHD method to see how it can be applied to inspect and interpret the training process in ANN. We recognize that building transparent machine learning systems is a convergent approach to both extracting new domain knowledge and performing model validation. As machine learning is increasingly applied to real-world decision making, the need for a transparent machine learning process will continue to grow. Using IVHD, we show that relationships between learned representations and between neurons can be studied, and efficient and time&memory optimal high-dimensional data visualization methods can be very useful in this interpretation process. Furthermore, the entire DNN visualization framework can be further improved by efficiently implementing it on modern hardware, for example, GPGPUs (*IVHD-CUDA*).

In future work, we plan to further investigate the visualization and interpretation of DNNs. In particular, we are eager to develop and explore a unified procedure for real-time analysis of the training process, which could greatly contribute to the possibility of optimizing deep network architectures. Furthermore, although dimensionality reduction is among the most scalable methods for high-dimensional data visualization, it still has some problems. For this reason, we plan to improve our CPU implementation

of IVHD method, which could be further accelerated by using low-level AVX instructions or more advanced multithreading operations. We also plan to further investigate supervised visualization methods and our meta-platform. Perhaps supervised training variants of DE methods would be possible alongside DNN, which could somehow elevate the real-time visualization of the training process. We hope that these approaches will further improve the performance of large high-dimensional data visualization and DNN applied to sparse, high-dimensional, unstructured data.

APPENDIX A

DATASETS

In this appendix, we list and briefly describe the datasets used in this work. These datasets are summarized in Table A.1. The experiments in Chapter 5 and Appendix B were conducted based on popular image and English-language text datasets, namely -MNIST, SmallNORB, 20 Newsgroups, and RCV1. Some basic pre-processing was applied to the datasets (4.2). In the case of RCV, only the 10 largest components (classes) were taken into account during visualization to remove the huge granularity. GPU methods were also tested using a YAHOO dataset, which contains questions and answers from the YAHOO service. In addition, a set of small synthetic datasets was generated to allow simple benchmarking of the embedding methods.

Table A.1: The list of baseline datasets.

Dataset	N	M	K	Short description
MNIST	784	70 000	10	Well balanced set of grayscale images of handwritten digits.
Fashion-MNIST	784	70 000	10	More difficult MNIST version. Instead of handwritten digits it consists of apparel images.
Extended-MNIST Letters	784	103 600	26	Merges a balanced set of the uppercase and lowercase letters into a single set.
Small NORB	2048	48 600	5	It contains stereo image pairs of 50 uniform-colored toys under 18 azimuths, 9 elevations, and 6 lighting conditions.
20-NG	5000	18846	20	Collection of approximately 20 000 newsgroup documents corresponding to a different topic.
RCV-Reuters	30	804 409	8	Corpus of press articles preprocessed to 30D by PCA.
SVHN	1024	63200	10	32x32 images of real-world numbers obtained from house numbers in Google Street View images.
CIFAR-10	1024	60000	10	32x32 images of real-world objects (e.g. airplanes, birds, dogs).
YAHOO	100	1.4 million	10	Questions and answers from YAHOO. The answers service preprocessed with FastText [111].
Amazon20M	100	20 million	5	Book reviews from Amazon. The reviews were preprocessed with FastText [111].

A.1. Datasets

A.1 MNIST

The MNIST dataset [74] is a widely used benchmark for machine learning algorithms. It consists of images of handwritten digits (Figure A.1). In the original dataset, the images are represented by 256 grayscale levels, but in this work we used rescaled pixel intensities to the $[0, 1]$ interval.

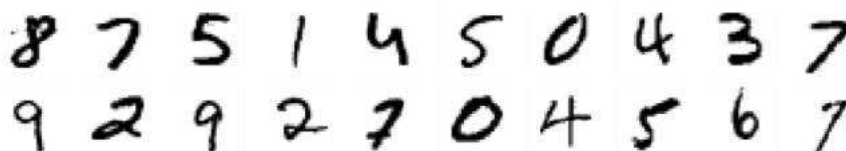


Figure A.1: Examples of MNIST dataset.

A.2 Fashion-MNIST

Fashion-MNIST [147] is a dataset of Zalando's product images, consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28×28 grayscale image associated with a label from 10 classes. Zalando intends Fashion-MNIST to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure as the training and testing splits.



Figure A.2: Examples of FMNIST dataset.

A.3 Extended-MNIST Letters

The EMNIST dataset [20] is a set of handwritten character digits converted to a 28×28 pixel image format and a data set structure that corresponds directly to the MNIST dataset. The Letters data set merges a balanced set of upper- and lowercase letters into a single 26-class set.

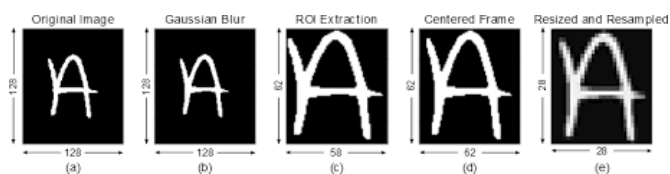


Figure A.3: Examples of EMNIST dataset.

A.4 smallNORB

The smallNORB dataset [75] is a dataset used for the recognition of 3D objects of its shape. It contains images of 50 toys belonging to five generic categories: four-legged animals, human figures, airplanes, trucks, and cars. The objects were imaged by two cameras under six lighting conditions, nine elevations (30 to 70 degrees every 5 degrees) and 18 azimuths (0 to 340 every 20 degrees).

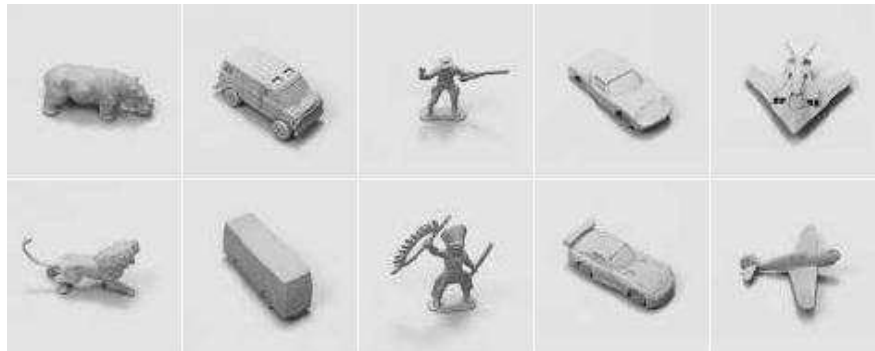


Figure A.4: Examples of smallNORB images.

A.5 20 News Groups

The 20 Newsgroups dataset [73] is a collection of approximately 20 000 newsgroup documents. The dataset is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*), while others are highly unrelated (e.g. *misc.forsale* and *soc.religion.christian*).

Table A.2: List of the 20 newsgroups, partitioned (more or less) according to topic.

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

A.6 RCV-Reuters

In 2000, Reuters Ltd. made available a large collection of Reuters News stories for use in the research and development of natural language processing, information retrieval, and machine learning systems. This corpus, known as the "Reuters Corpus, Volume 1" or RCV1 [80], is significantly larger than the older and well-known Reuters-

21578 collection, which is widely used in the text classification community. It is an archive of more than 800,000 manually categorized newswire stories made available by Reuters, Ltd. for research purposes.

A.7 Street View House Numbers (SVHN)

SVHN [101] is a real-world image dataset for developing machine learning and object recognition algorithms with a minimal data pre-processing and formatting requirement. It can be seen as similar in flavor to MNIST (e.g., the images are of small cropped digits), but incorporates an order of magnitude more labeled data (over 600,000 digit images) and comes from a significantly harder, unsolved, real-world problem (recognizing digits and numbers in natural scene images). It is obtained from house numbers in Google Street View images.



Figure A.5: Examples of SVHN images.

A.8 CIFAR-10

CIFAR-10 [71] consists of 60000 32x32 colored images in 10 classes, with 6000 images per class. The dataset is divided into five training batches and one test batch, each with 10000 images. The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. The classes are mutually exclusive, which means that there is no overlap between different classes (for example, automobiles and trucks).



Figure A.6: Examples of CIFAR-10 images.

There is also a variant of CIFAR-10 called CIFAR-100. It is like CIFAR-10, except that it has 100 classes containing 600 images each, which provides much greater diversity.

A.9 Amazon3M and Amazon10M

Amazon3M is sampled from all Amazon Reviews (233.1 million), which contains 3 million 100-D vectors. Amazon10M was sampled from Amazon20M. The description of whole Amazon dataset is placed below in the Amazon20M section.

A.10 Amazon20M

This dataset contains product reviews and metadata from Amazon, including 233.1 million reviews spanning May 1996 - October 2018. Reviews include ratings, text, helpfulness votes, descriptions, category information, price, brand, image features, and links. Amazon20M contains reviews from the Books section.

```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
  He is having a wonderful time playing these old hymns.
  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

Snippet A.1: Sample of the Amazon dataset record (*reviewText* is later preprocessed by FastText [111].)

A.11 Synthetic datasets

We prepared a set of artificially generated "simple" data sets whose internal data structure gives insight into the quality of embeddings created by different methods.

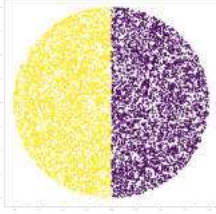
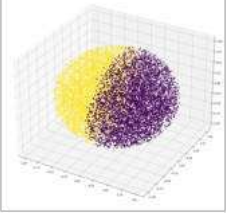
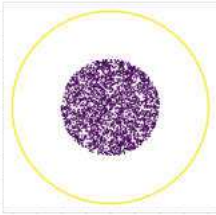
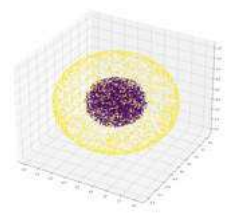
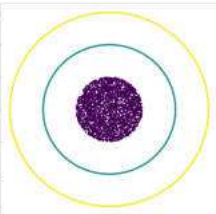
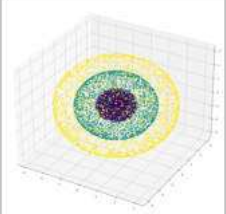
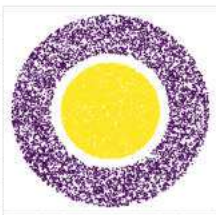
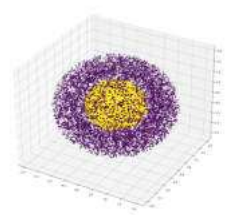
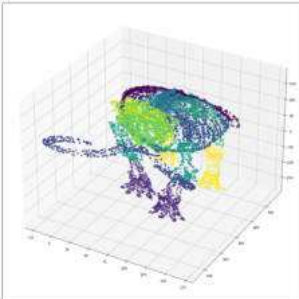
Description	2-D and 3-D example view	
An N-dimensional ball divided in half.		
An N-dimensional ball inside a sphere.		
An N-dimensional ball inside two spheres.		
An N-dimensional ball inside an N-dimensional empty ball.		
Mammoth dataset.		

Table A.3: Artificially generated ("synthetic") N-D baseline datasets and its 3-D (and 2-D) example views.

APPENDIX B

VISUALIZATIONS, TIMINGS AND METHODS PARAMETRIZATION

In this appendix, we include additional visualizations (for datasets not presented in the main body of dissertation), timings achieved for both CPU and GPU implementation of different methods and thorough parametrization, that was used for methods.

Table B.1: Glossary of the parameters of the DE algorithms used in the experiments. All IVHD parameters are displayed. We also include general parameters used for UMAP, t-SNE, TriMAP, and PaCMAP. All algorithms except IVHD require matching a considerably higher number of parameters (about 10) [27, 21]), so due to brevity only the most important ones are given, that is, *perplexity*, *nn*, *ni* (number of inlier points for triplet constraints).

Method	Dataset	Perplexity	ni	nn	rn	c	Metric
IVHD / IVHD-CUDA	MNIST	-	-	2	1	0.01	Euclidean
IVHD / IVHD-CUDA	Fashion-MNIST	-	-	2	1	0.01	Cosine
IVHD / IVHD-CUDA	Small Norb	-	-	5	1	0.01	Cosine
IVHD / IVHD-CUDA	RCV-Reuters	-	-	3	1	0.1	Cosine
IVHD / IVHD-CUDA	YAHOO	-	-	2	1	0.1	Cosine
t-SNE	All	40, 80, 120	-	-	-	-	Euclidean
UMAP	All	-	-	15,30,50	-	-	Euclidean
TriMap	All	-	12, 20, 40	-	-	-	Euclidean
PaCMAP	All	-	-	10, 15, 20	-	-	Euclidean
BH-SNE-CUDA	All	50	-	32	-	-	Euclidean
AtSNE-CUDA	All	50	-	100	-	-	Euclidean

All metrics were calculated and plotted to obtain the best configuration of a given method (from those presented above). The results were compared locally, and one that obtains the best metrics results was selected for the final drawing of the metrics.

An N-dimensional ball divided in half.

All methods were unable to separate the classes for a 30-dimensional sphere. The two classes completely overlap (Fig. B.1). Therefore, we performed cross-sectional visualizations for different dimensionalities of the ball (Fig. B.2). As you can see, separation is possible only for low-dimensionalities (3D-5D). Any higher dimensionality causes the classes to merge, which is the direct cause of *crowding effect*.

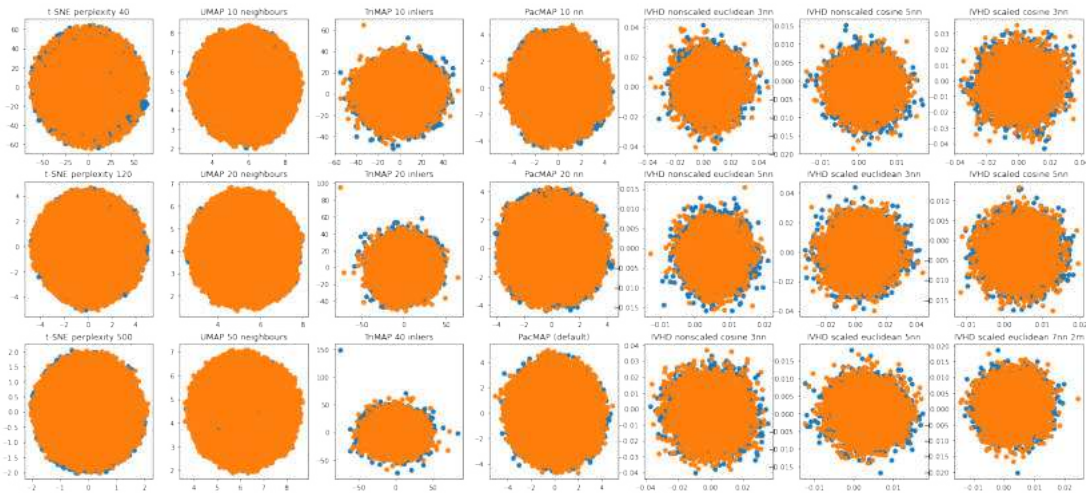


Figure B.1: Methods comparison for 10k-sample ball (30-D) splitted into 2 separate classes.

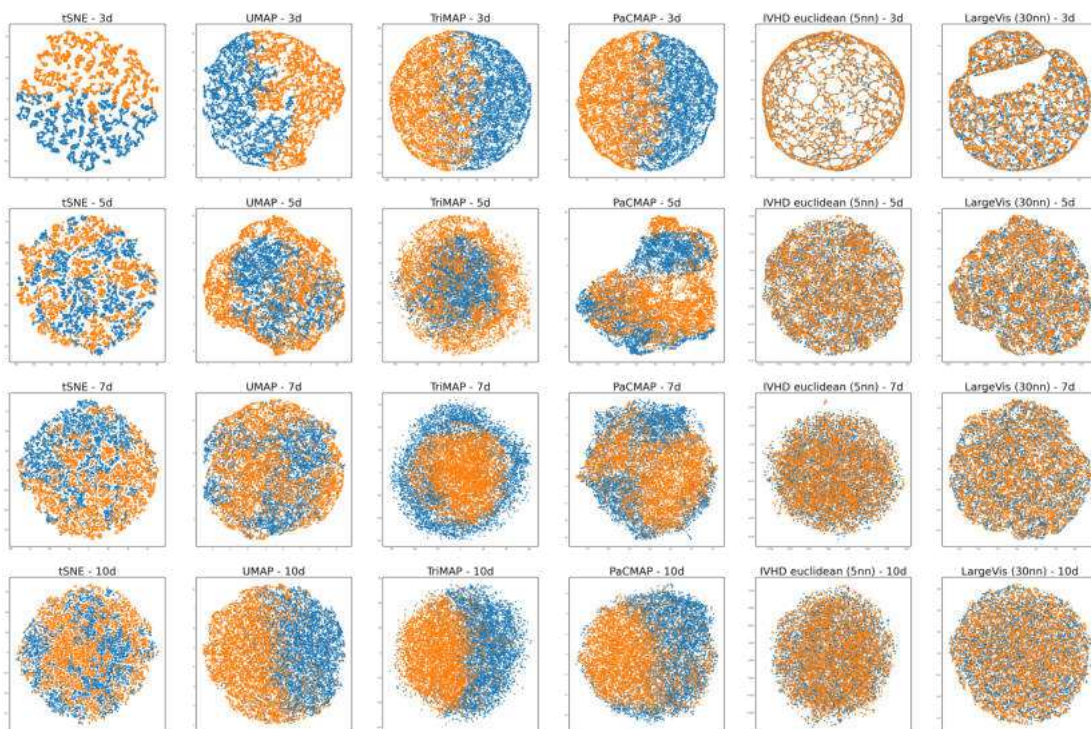


Figure B.2: Methods comparison for 10k-sample ball splitted into 2 separate classes. Various dimensionality.

Methods that have coped as such with the separation of low dimensions are UMAP, TriMAP, and PaCMAP (2nd, 3rd, 4th column from the left in Fig. B.2). IVHD, as the simplest and fastest method, is unfortunately unable to recognize the features responsible for the separations in such a homogeneous dataset. The same is true for both LargeVis and t-SNE (although t-SNE could separate two classes when embedded from 3-D). For all methods, we observe that they tend to distribute data fairly uniformly around the space, which may potentially contribute to the preservation of local structure and hinder the preservation of global structure. Visually, for the highest dimension (last row in Fig. B.2), the TriMap and UMAP methods are methods that create a clear separation of classes in the center of the ball.

In Figures B.3, B.4 we can observe the embedding of the fourth synthetic data set (an N-dimensional ball inside an N-dimensional ring). The best performance in terms of DR quality is achieved by the TriMAP method. For all the cases, there is a clear separation for two classes, the ball and the empty ball.

An N-dimensional ball inside of empty ball.

The best visualization of this dataset was undisputedly created using the TriMap method. There is not only a clear separation of the two classes, but the method was also able to create an inner ball that is bounded by a second hollow ball (a bold sphere). The rest of the methods had trouble rendering this shape. IVHD is the only method, in addition to TriMap, that was able to reflect the overall structure of the classes, but it is not as accurate as in the first case. UMAP and PaCMAP created the internal ball, but its surrounding was chaotic and completely did not reflect the real structure of this dataset.

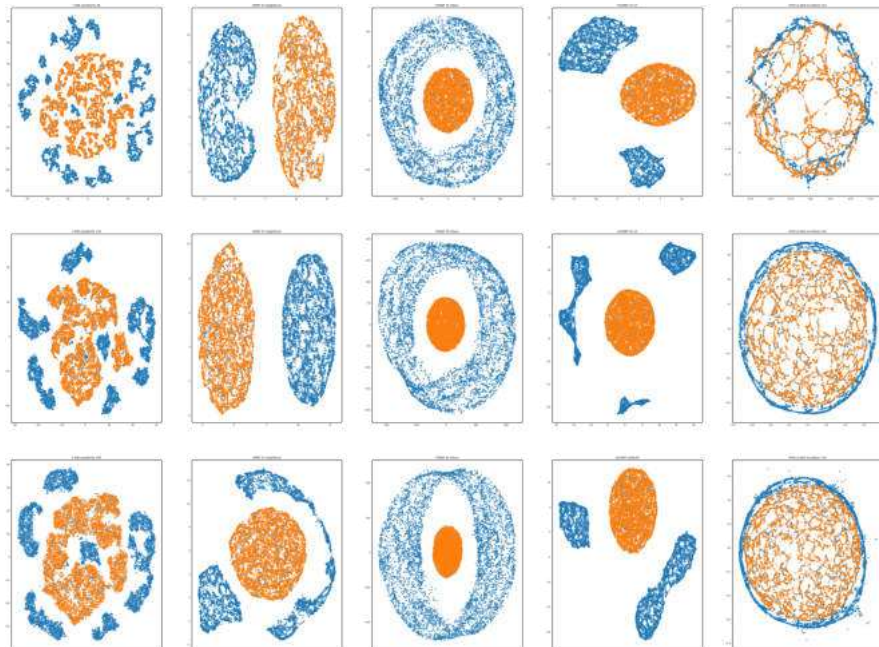


Figure B.3: Methods comparison for 10k-sample ball inside empty ball dataset. Both manifolds were generated in 30-D.

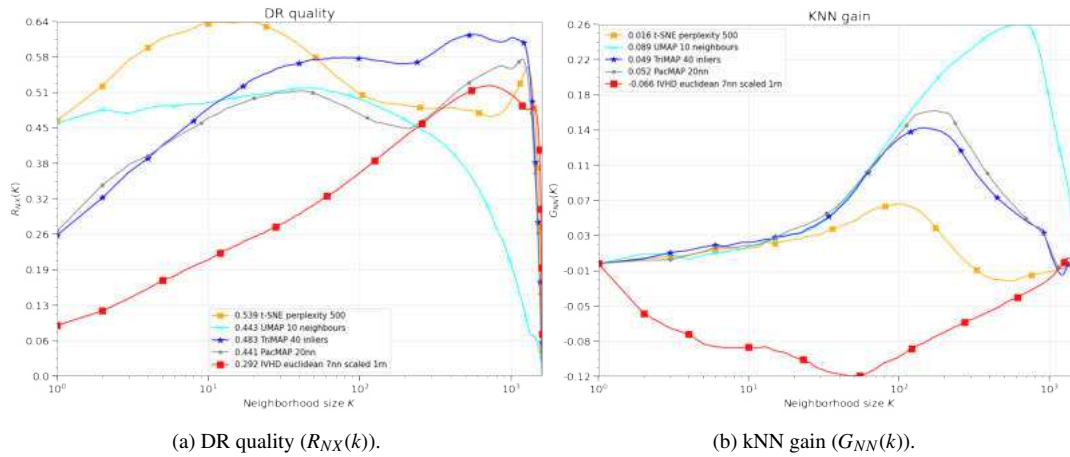


Figure B.4: KNN gain and DR quality obtained for comparison of different DR methods for ball inside empty ball dataset.

The metrics presented in Fig. B.4 confirm the conclusions described above, in which TriMap achieves the best results (DR quality plot), but UMAP completely fails to preserve quality. It generates very clear two clusters and for this reason it spikes in the kNN gain metric plot.

Mid-scale datasets

Based on the visualizations presented in Figs. B.7, B.5, B.5 and the diagrams presented in Figs. B.8, B.6 and B.10 one can make the following observations:

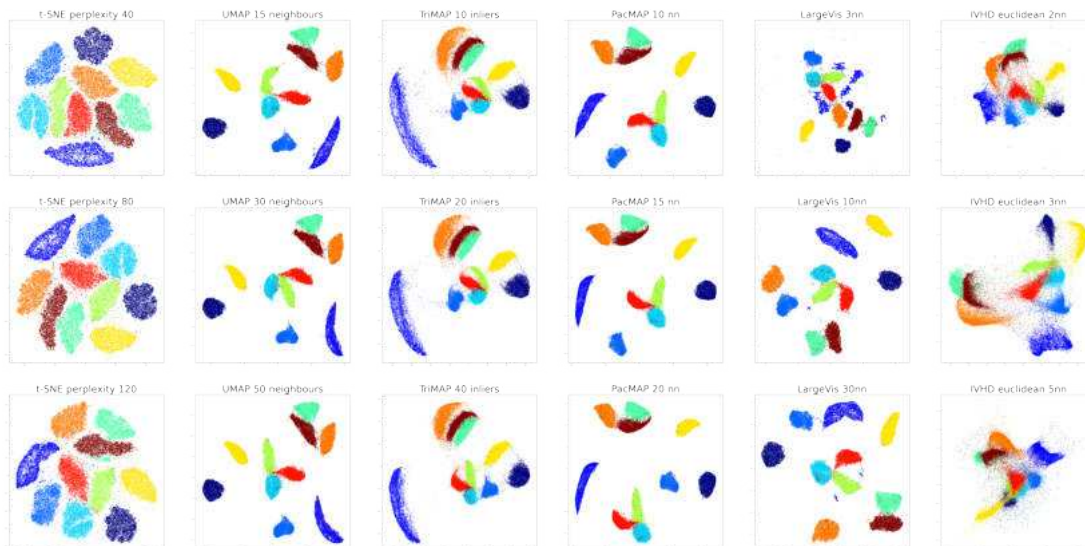


Figure B.5: Methods comparison for MNIST dataset.

1. IVHD applied to both MNIST and EMNIST forms separate clusters of different, mostly elongated shapes (similarly as with FMNIST in Fig. 4.14). Furthermore, the generated mapping is fuzzy. On the other hand, the LargeVis, PacMAP, t-SNE, and TriMap methods are much better at creating rounded and clearly separated

clusters. Additionally, in t-SNE, some classes are mixed and fragmented. In terms of DR quality, LargeVis, UMAP, and PaCMAP are achieving the best results.

- All methods were more or less capable of preserving the distance ratio between the (high and low dimensional) spaces for the EMNIST and MNIST dataset (Figs. B.8 and B.6). MAP methods seem to have the most condensed points on the diagonal of the Shepard diagram, which indicates the best preservation of the distance ratio. Although in both cases, IVHD obtains the most distorted Shepard diagram.

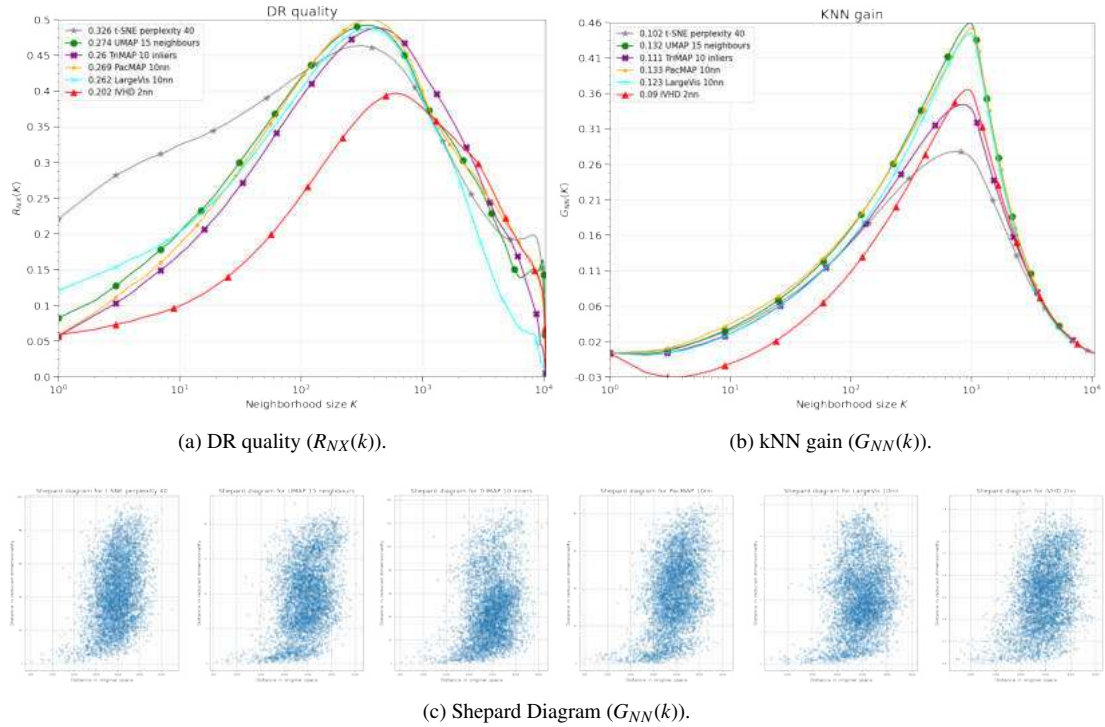


Figure B.6: KNN gain, DR quality and Shepard Diagrams obtained for comparison of different DR methods for MNIST dataset.

- The original IVHD implementation does not separate clusters in the EMNIST data set, as could be with proper adjustments. Therefore, this data set was chosen to be one of the benchmark data sets to show the improvements in IVHD presented in the section 3.4.
- For the 20 newsgroups dataset, we can see that there is a complete separation of classes in the case of the IVHD and LargeVis methods, resulting in very low DR quality and a huge KNN gain. For all methods, the distance ratio shown in the Shepard diagram (Fig. B.10) is not "linear". This means that there is no preservation of the distance ratio and that the embeddings do not preserve the original data structure. It is quite abject, especially in IVHD and LargeVis case, where we see this extreme cluster separation.
- In terms of DR quality, we can verify that PaCMAP has the highest quality in neighborhood size of 50–1000. For KNN gain, of course, LargeVis and IVHD



Figure B.7: Methods comparison for EMNIST-Letters dataset.

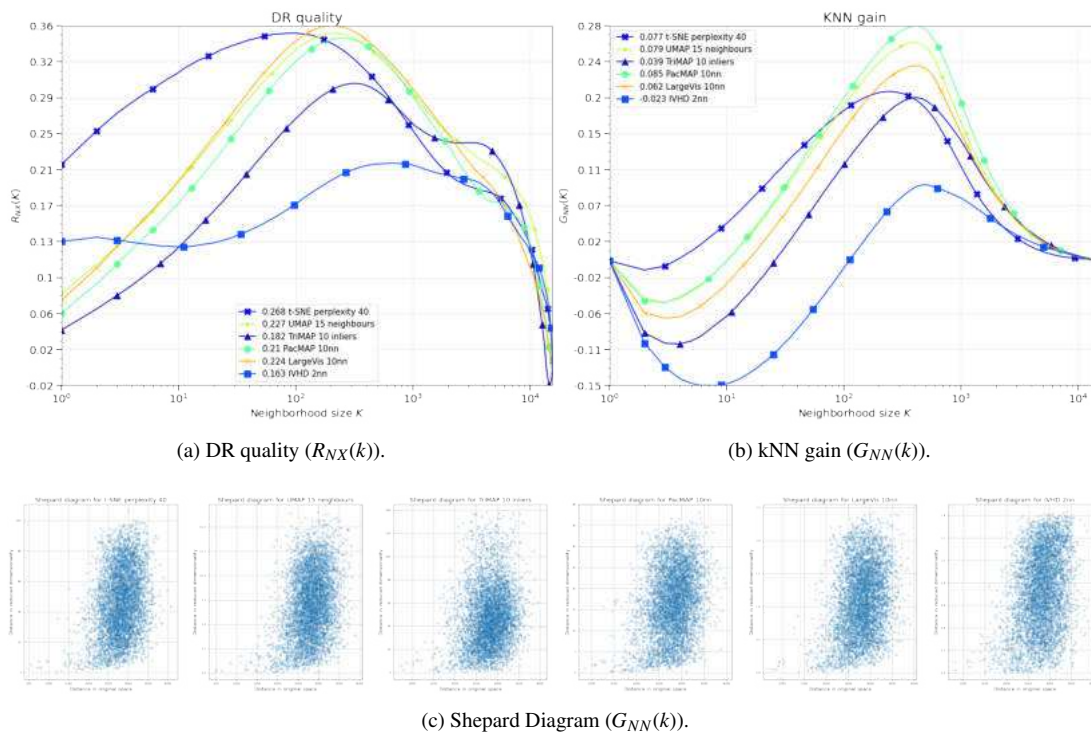


Figure B.8: KNN gain, DR quality and Shepard Diagrams obtained for comparison of different DR methods for EMNIST dataset.

obtain the highest score, both being the methods with the highest level of cluster separation. It is important to remember that it does not mean that embedding preserves the overall data structure (as mentioned, DR quality gives us that information).

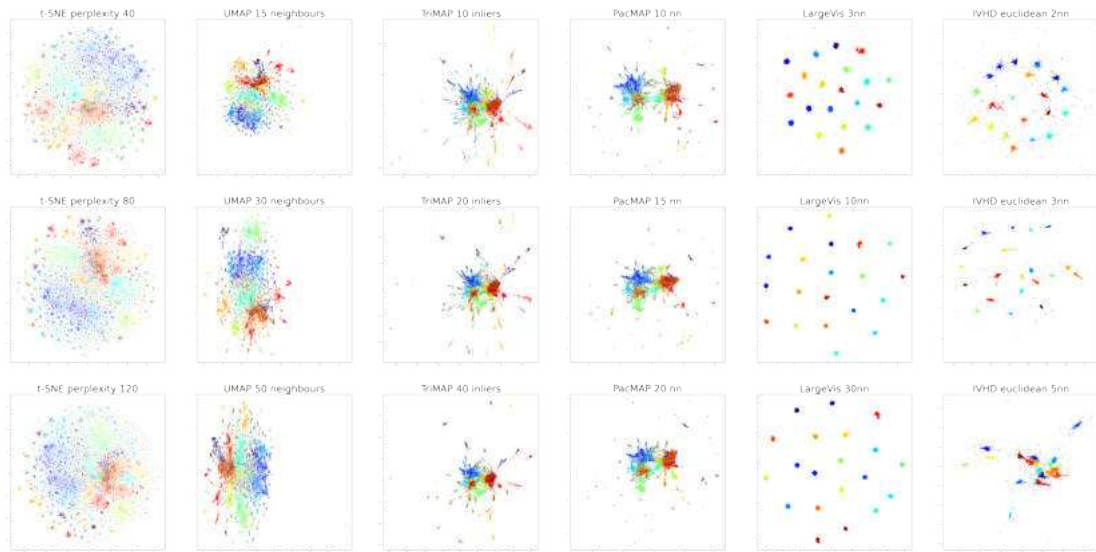
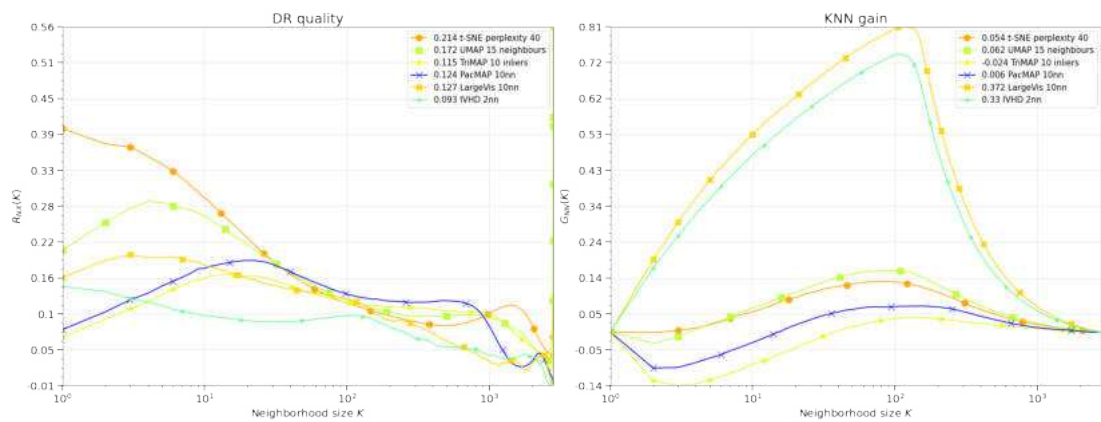
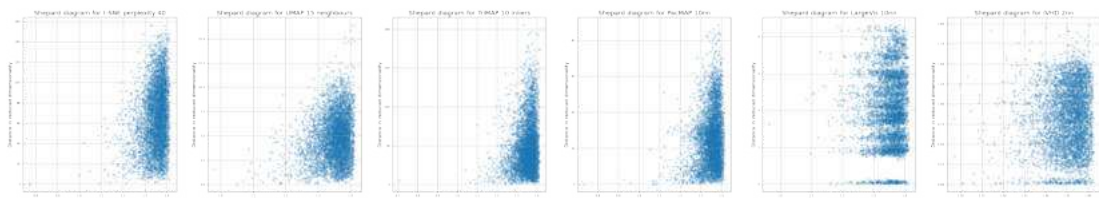


Figure B.9: Methods comparison for for 20-News Groups dataset.



(a) DR quality ($R_{NX}(k)$).

(b) kNN gain ($G_{NN}(k)$).



(c) Shepard Diagram ($G_{NN}(k)$).

Figure B.10: KNN gain, DR quality and Shepard Diagrams obtained for comparison of different DR methods for 20-News Groups dataset.

BIBLIOGRAPHY

- [1] Abe, M., Miyao, J., Kurita, T.: q-sne: Visualizing data using q-gaussian distributed stochastic neighbor embedding (2020), <https://arxiv.org/abs/2012.00999>
- [2] Amid, E., Warmuth, M.K.: TriMap: Large-scale Dimensionality Reduction Using Triplets. arXiv preprint arXiv:1910.00204 (2019)
- [3] Amid, E., Gionis, A., Ukkonen, A.: Semi-supervised kernel metric learning using relative comparisons (2016), <https://arxiv.org/abs/1612.00086>
- [4] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., Bengio, Y.: Theano: new features and speed improvements (2012), <https://arxiv.org/abs/1211.5590>
- [5] Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing System* **14** (04 2002)
- [6] Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15**, 1373–1396 (2003)
- [7] Bengio, Y.: Practical Recommendations for Gradient-Based Training of Deep Architectures, 437–478. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
- [8] Bentley, J.L.: Multidimensional binary search trees used for associative searching. *Commun. ACM* **18**(9), 509–517 (sep 1975)
- [9] Beygelzimer, A., Kakade, S., Langford, J.: Cover trees for nearest neighbor. In: *Proceedings of the 23rd International Conference on Machine Learning*. 97–104. ICML '06, Association for Computing Machinery, New York, NY, USA (2006)
- [10] de Bodt, C., Mulders, D., Lopez-Sanchez, D., Verleysen, M., Lee, J.A.: Class-aware t-SNE: cat-SNE. In: *ESANN*. 409–414 (2019)
- [11] Borg, I., Groenen, P.: *Modern Multidimensional Scaling: Theory and Applications*. Springer (2005)
- [12] Boytsov, A., Fouquet, F., Hartmann, T., Traon, Y.L.: Visualizing and exploring dynamic high-dimensional datasets with lion-tsne. *ArXiv* **abs/1708.04983** (2017)

- [13] Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences* **103**(5), 1168–1172 (2006)
- [14] Böhm, J.N., Berens, P., Kobak, D.: Attraction-repulsion spectrum in neighbor embeddings (2020)
- [15] Carreira-Perpiñán, M.A., Lu, Z.: The laplacian eigenmaps latent variable model. In: Meila, M., Shen, X. (eds.) *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*, 2, 59–66. PMLR, San Juan, Puerto Rico (21–24 Mar 2007)
- [16] Chao, G., Luo, Y., Ding, W.: Recent advances in supervised dimension reduction: A survey. *Machine Learning and Knowledge Extraction* **1**(1), 341–358 (2019)
- [17] Chen, L., Buja, A.: Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association* **104**(485), 209–219 (2009)
- [18] Cheng, Y., Wang, X., Xia, Y.: Supervised t-distributed stochastic neighbor embedding for data visualization and classification. *INFORMS J. on Computing* **33**(2), 566–585 (may 2021)
- [19] Cofaru, C.: Inverted file system with asymmetric distance computation for billion-scale approximate nearest neighbor search. <https://github.com/zgornel/IVFADC.jl/> (2019), [Online; accessed 20-November-2019]
- [20] Cohen, G., Afshar, S., Tapson, J., van Schaik, A.: Emnist: an extension of mnist to handwritten letters (2017), <https://arxiv.org/abs/1702.05373>
- [21] Cong Fu, Yonghui Zhang, D.C., Ren, X.: Anchor-t-sne for large-scale and high-dimension vector visualization. (2019), <https://github.com/ZJUlearning/AtSNE>, [Online; accessed 20-November-2019]
- [22] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: *Introduction to Algorithms*, Third Edition. The MIT Press, 3rd edn. (2009)
- [23] Crecchi, F., de Bodt, C., Verleysen, M., Lee, J.A., Bacciu, D.: Perplexity-free parametric t-sne. *arXiv preprint arXiv:2010.01359* (2020)
- [24] Damrich, S., Böhm, J.N., Hamprecht, F.A., Kobak, D.: Contrastive learning unifies *t*-sne and umap (2022)
- [25] Dasgupta, S., Freund, Y.: Random projection trees and low dimensional manifolds. In: *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*. 537–546. STOC '08, Association for Computing Machinery, New York, NY, USA (2008)

- [26] Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Twentieth Annual Symposium on Computational Geometry. 253–262. SCG '04, Association for Computing Machinery, New York, NY, USA (2004)
- [27] David M. Chan, Roshan Rao, F.H., Canny, J.F.: Gpu accelerated t-distributed stochastic neighbor embedding. *Journal of Parallel and Distributed Computing* **131**, 1–13 (2019), <https://github.com/CannyLab/tsne-cuda/>, [Online; accessed 20-November-2019]
- [28] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, 32, 647–655. PMLR, Beijing, China (22–24 Jun 2014)
- [29] Dong, W., Moses, C., Li, K.: Efficient k-nearest neighbor graph construction for generic similarity measures. In: Proceedings of the 20th International Conference on World Wide Web. 577–586. WWW '11, Association for Computing Machinery, New York, NY, USA (2011)
- [30] Dzwinel, W.: Virtual particles and search for global minimum. *Future Generation Computer Systems* **12**(5), 371–389 (1997)
- [31] Dzwinel, W., Błasiak, J.: Method of particles in visual clustering of multi-dimensional and large data sets. *Future Generation Computer Systems* **15**(3), 365–379 (1999)
- [32] Dzwinel, W., Wcisło, R., Czech, W.: ivga: A fast force-directed method for interactive visualization of complex networks. *Journal of Computational Science* **21**, 448–459 (2017)
- [33] Dzwinel, W., Wcislo, R., Matwin, S.: 2-d embedding of large and high-dimensional data with minimal memory and computational time requirements. arXiv preprint arXiv:1902.01108 (2019)
- [34] Dzwinel, W., Wcislo, R., Strzoda, M.: ivga: Visualization of the network of historical events. In: Proceedings of the 1st International Conference on Internet of Things and Machine Learning. ACM (2017)
- [35] Falconer, K.: *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley and Sons (2014)
- [36] Fefferman, C., Mitter, S., Narayanan, H.: *Testing the manifold hypothesis* (2013)
- [37] Floyd, R.W.: Algorithm 97: shortest path. *Communications of the ACM* **5**(6), 345 (1962)

- [38] France, S., Carroll, J.: Development of an agreement metric based upon the rand index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data. In: Machine Learning and Data Mining in Pattern Recognition, Lect. Notes Comput. Sc. 4571, 499–517 (07 2007)
- [39] France, S.L., Carroll, J.D.: Two-way multidimensional scaling: A review. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **41**(5), 644–661 (2011)
- [40] Friedman, J.H., Bentley, J.L., Finkel, R.A.: An algorithm for finding best matches in logarithmic expected time. ACM Trans. Math. Softw. **3**(3), 209–226 (sep 1977)
- [41] Fruchterman, T.M., Reingold, E.M.: Graph drawing by force-directed placement. Software: Practice and experience **21**(11), 1129–1164 (1991)
- [42] Fu, C., Zhang, Y., Cai, D., Ren, X.: Atsne: Efficient and robust visualization on gpu through hierarchical optimization. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 176–186 (2019)
- [43] García-Fernández, F.J., Verleysen, M., Lee, J.A., Díaz, I.: Stability Comparison of Dimensionality Reduction Techniques Attending to Data and Parameter Variations. In: Aupetit, M., van der Maaten, L. (eds.) EuroVis Workshop on Visual Analytics using Multidimensional Projections. The Eurographics Association (2013)
- [44] Ghojogh, B., Crowley, M.: Unsupervised and supervised principal component analysis: Tutorial (2019), <https://arxiv.org/abs/1906.03148>
- [45] Ghojogh, B., Ghodsi, A., Karray, F., Crowley, M.: Locally linear embedding and its variants: Tutorial and survey (2020), <https://arxiv.org/abs/2011.10925>
- [46] Ghojogh, B., Ghodsi, A., Karray, F., Crowley, M.: Multidimensional scaling, sammon mapping, and isomap: Tutorial and survey (2020), <https://arxiv.org/abs/2009.08136>
- [47] Ghojogh, B., Ghodsi, A., Karray, F., Crowley, M.: Uniform manifold approximation and projection (umap) and its variants: Tutorial and survey (2021)
- [48] Ghojogh, B., Karray, F., Crowley, M.: Eigenvalue and generalized eigenvalue problems: Tutorial (2019)
- [49] Ghojogh, B., Sikaroudi, M., Shafiei, S., Tizhoosh, H., Karray, F., Crowley, M.: Fisher discriminant triplet and contrastive losses for training siamese networks. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE (jul 2020)

- [50] Gibson, H., Faith, J., Vickers, P.: A survey of two-dimensional graph layout techniques for information visualisation. *Information Visualization* **12**, 324–357 (2013)
- [51] Gisbrecht, A., Schulz, A., Hammer, B.: Parametric nonlinear dimensionality reduction using kernel t-sne. *Neurocomputing* **147**, 71–82 (2015), advances in Self-Organizing Maps Subtitle of the special issue: Selected Papers from the Workshop on Self-Organizing Maps 2012 (WSOM 2012)
- [52] Goodfellow I., B.Y., A., C.: *Deep learning*. The MIT Press (2017)
- [53] Gower, J.C., Dijksterhuis, G.B.: *Procrustes Problems*. Oxford University Press (01 2004)
- [54] Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). 2, 1735–1742 (2006)
- [55] Hamel, P., Eck, D.: Learning features from music audio with deep belief networks. In: ISMIR. 339–344 (01 2010)
- [56] Hatcher, A.: *Algebraic topology*. Cambridge University Press, Cambridge (2002)
- [57] Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* **22**(1), 5–53 (jan 2004)
- [58] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006)
- [59] Hinton, G., Roweis, S.: Stochastic neighbor embedding. *Advances in neural information processing systems* **15**, 833–840 (2003)
- [60] Hotelling, H.: Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 498–520 (1933)
- [61] Ingram, S., Munzner, T., Olano, M.: Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics* **15**(2), 249–261 (2009)
- [62] Ioffe, S.: Probabilistic linear discriminant analysis. In: *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*. 531–542. ECCV'06, Springer-Verlag, Berlin, Heidelberg (2006)
- [63] Jiang, Q., Jia, M.: Supervised laplacian eigenmaps for machinery fault classification. In: 2009 WRI World Congress on Computer Science and Information Engineering. 7, 116–120 (2009)
- [64] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. arXiv preprint arXiv:1702.08734 (2017)

- [65] Jäckle, D., Fischer, F., Schreck, T., Keim, D.A.: Temporal mds plots for analysis of multivariate data. *IEEE Transactions on Visualization and Computer Graphics* **22**(1), 141–150 (2016)
- [66] Kaya, M., Bilge, H.S.: Deep metric learning: A survey. *Symmetry* **11**(9) (2019)
- [67] Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013)
- [68] Ko, H.K., Jo, J., Seo, J.: Progressive Uniform Manifold Approximation and Projection. In: Kerren, A., Garth, C., Marai, G.E. (eds.) *EuroVis 2020 - Short Papers*. The Eurographics Association (2020)
- [69] Kochenderfer, M.J., Wheeler, T.A.: Algorithms for optimization. The MIT Press (2019)
- [70] Kouropteva, O., Okun, O., Pietikäinen, M.: Incremental locally linear embedding. *Pattern Recognition* **38**(10), 1764–1767 (2005), <https://www.sciencedirect.com/science/article/pii/S0031320305001792>
- [71] Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009)
- [72] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. 1097–1105. NIPS'12, Curran Associates Inc., Red Hook, NY, USA (2012)
- [73] Lang, K.: Newsweeder: Learning to filter netnews. In: *Proceedings of the Twelfth International Conference on Machine Learning*. 331–339 (1995)
- [74] LeCun, Y.: Mnist database (1998), <http://yann.lecun.com/exdb/mnist/>
- [75] LeCun, Y.: The smallnorb database (2005), <https://cs.nyu.edu/~ylclab/data/norb-v1.0-small/>
- [76] Lee, J.A., Peluffo-Ordóñez, D.H., Verleysen, M.: Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing* **169**, 246–261 (2015), learning for Visual Semantic Understanding in Big Data ESANN 2014 Industrial Data Processing and Analysis
- [77] Lee, J.A., Renard, E., Bernard, G., Dupont, P., Verleysen, M.: Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing* **112**, 92–108 (2013), advances in artificial neural networks, machine learning, and computational intelligence
- [78] Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing* **72**(7), 1431–1443 (2009), advances in Machine Learning and Computational Intelligence

- [79] Lei, Y.K., Xu, Y., Zhang, S.W., Wang, S.L., Ding, Z.G.: Fast isomap based on minimum set coverage. In: Huang, D.S., Zhang, X., Reyes García, C.A., Zhang, L. (eds.) *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*. 173–179. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
- [80] Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research* **5**, 361–397 (2004)
- [81] Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., Kluger, Y.: Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nature methods* **16**, 243 – 245 (2019)
- [82] Liu, W.m., Chang, C.I.: Variants of principal components analysis. In: 2007 IEEE International Geoscience and Remote Sensing Symposium. 1083–1086 (2007)
- [83] Lombaert, H., Zikic, D., Criminisi, A., Ayache, N.: Laplacian forests: Semantic image segmentation by guided bagging. In: *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*. 17, 496–504 (09 2014)
- [84] van der Maaten, L.: Learning a parametric embedding by preserving local structure. In: *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*. 5, 384–391 (2009)
- [85] van der Maaten, L.: Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research* **15**, 3221–3245 (2014)
- [86] van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008)
- [87] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders (2015)
- [88] Marcílio-Jr, W.E., Eler, D.M., Paulovich, F.V., Martins, R.M.: Humap: Hierarchical uniform manifold approximation and projection (2021)
- [89] Marsland, S.: Swissroll database (1998), <https://people.cs.uchicago.edu/~dinoj/manifold/swissroll.html>
- [90] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018)
- [91] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality (2013)
- [92] Minch, B.: Viskit library. <https://gitlab.com/bminch/viskit>

- [93] Minch, B., Nowak, M.P., Wcislo, R., Dzwinel, W.: Gpu-embedding of knn-graph representing large and high-dimensional data. *Computational Science – ICCS 2020* **12138**, 322 – 336 (2020)
- [94] Mishra, S., Misra, A.: Structured and unstructured big data analytics. In: 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC). 740–746 (2017)
- [95] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. *Nature* **518**(7540), 529–533 (Feb 2015)
- [96] Mohamed, A.r., Hinton, G., Penn, G.: Understanding how deep belief networks perform acoustic modelling. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 4273–4276 (2012)
- [97] Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP (2009)
- [98] Narayan, A., Berger, B., Cho, H.: Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. *bioRxiv* (2020)
- [99] Narayan, A., Berger, B., Cho, H.: Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature biotechnology* **39**(6), 765—774 (June 2021), <https://europepmc.org/articles/PMC8195812>
- [100] Naudts, J.: Deformed exponentials and logarithms in generalized thermostatics. *Physica A: Statistical Mechanics and its Applications* **316**(1-4), 323–334 (dec 2002)
- [101] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011 (2011), http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf
- [102] Oja, H., Sirkiä, S., Eriksson, J.: Scatter matrices and independent component analysis. *AUSTRIAN JOURNAL OF STATISTICS* Volume **35**, 175–189 (01 2006)
- [103] Ouyang, D., Wen, D., Qin, L., Chang, L., Zhang, Y., Lin, X.: Progressive top-k nearest neighbors search in large road networks. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. 1781–1795. SIGMOD '20, Association for Computing Machinery, New York, NY, USA (2020)
- [104] Pawliczek, P., Dzwinel, W.: Interactive data mining by using multidimensional scaling. *Procedia Computer Science* **18**, 40–49 (2013)

- [105] Pawliczek, P., Dzwinel, W., Yuen, D.: Visual exploration of data by using multidimensional scaling on multicore cpu, gpu, and mpi cluster. *Concurrency and Computation Practice and Experience* **3**, 1–21 (2014)
- [106] Pawliczek, P., Dzwinel, W., Yuen, D.A.: Visual exploration of data with multithread mic computer architectures. In: *International Conference on Artificial Intelligence and Soft Computing*. 25–35. Springer (2015)
- [107] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay: Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12**(85), 2825–2830 (2011)
- [108] Pezzotti, N., Lelieveldt, B.P.F., van der Maaten, L., Höllt, T., Eisemann, E., Vilanova, A.: Approximated and user steerable tsne for progressive visual analytics (2015)
- [109] Qu, T., Cai, Z.: A fast isomap algorithm based on fibonacci heap. In: Tan, Y., Shi, Y., Buarque, F., Gelbukh, A., Das, S., Engelbrecht, A. (eds.) *Advances in Swarm and Computational Intelligence*. 225–231. Springer International Publishing, Cham (2015)
- [110] Rauber, P.E., Fadel, S.G., Falcão, A.X., Telea, A.C.: Visualizing the hidden activity of artificial neural networks. *IEEE Transactions on Visualization and Computer Graphics* **23**(1), 101–110 (2017)
- [111] Research, F.A.: Library for fast text representation and classification. <https://github.com/facebookresearch/fastText> (2018), [Online; accessed 20-November-2019]
- [112] Ribeiro, B., Vieira, A., Carvalho das Neves, J.: Supervised isomap with dissimilarity measures in embedding learning. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) *Progress in Pattern Recognition, Image Analysis and Applications*. 389–396. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
- [113] Ridder, D., Kouropteva, O., Okun, O., Pietikäinen, M., Duin, R.: Supervised locally linear embedding. In: *Proceedings of Artificial Neural Networks and Neural Information Processing*. 2714, 333–341 (01 2003)
- [114] Robinson, I., Pierce-Hoffman, E.: Tree-sne: Hierarchical clustering and visualization using t-sne (2020)
- [115] Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.d.F., Rodrigues, F.A.: Clustering algorithms: A comparative approach. *PLOS ONE* **14**(1), 1–34 (01 2019)
- [116] Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)

- [117] Roychowdhury, S., Ghosh, J.: Robust laplacian eigenmaps using global information. *Manifold Learning and its Applications* (12 2011)
- [118] Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2017)
- [119] Sainburg, T., McInnes, L., Gentner, T.Q.: Parametric umap embeddings for representation and semi-supervised learning (2020), <https://arxiv.org/abs/2009.12981>
- [120] Sainburg, T., McInnes, L., Gentner, T.Q.: Parametric umap: learning embeddings with deep neural networks for representation and semi-supervised learning. arXiv preprint arxiv.2009.12981 (2020)
- [121] Salton, G.: Mathematics and information retrieval. *Journal of Documentation* **35**(1), 1–29 (Jan 1979)
- [122] Sammon, J.W.: A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* **C-18**, 401–409 (1969)
- [123] Schmidhuber, J.: Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117 (2015)
- [124] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE (jun 2015)
- [125] Schulz, A., Hinder, F., Hammer, B.: DeepView: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction. In: *Proceedings of the 29th International Joint Conference on Artificial Intelligence* (jul 2020)
- [126] Shaw, B., Jebara, T.: Structure preserving embedding. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 937–944. ACM (2009)
- [127] Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: *Proceedings of the 1968 23rd ACM National Conference*. 517–524. ACM '68, Association for Computing Machinery, New York, NY, USA (1968)
- [128] Silpa-Anan, C., Hartley, R.: Optimised kd-trees for fast image descriptor matching. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8 (2008)
- [129] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**(56), 1929–1958 (2014), <http://jmlr.org/papers/v15/srivastava14a.html>
- [130] Takane, Y.: 11 applications of multidimensional scaling in psychometrics. In: Rao, C., Sinharay, S. (eds.) *Psychometrics, Handbook of Statistics*, 26, 359–400. Elsevier (2006)

- [131] Tang, J., Liu, J., Zhang, M., Mei, Q.: Visualizing large-scale and high-dimensional data. Proceedings of the 25th International Conference on World Wide Web (Apr 2016)
- [132] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web (may 2015)
- [133] Tao, Y., Papadias, D., Lian, X.: Reverse knn search in arbitrary dimensionality. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30. 744–755. VLDB '04, VLDB Endowment (2004)
- [134] Tenenbaum, J., Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)* **290**, 2319–23 (01 2001)
- [135] Tharwat, A., Gaber, T., Ibrahim, A., Hassanien, A.E.: Linear discriminant analysis: A detailed tutorial. *Ai Communications* **30**, 169–190, (05 2017)
- [136] Thorpe, M.F., Duxbury, P.M.: Rigidity theory and applications. Springer Science & Business Media (1999)
- [137] Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61(3)**, 611–622 (1999)
- [138] Van Der Maaten, L., Postma, E., Van den Herik, J.: Dimensionality reduction: a comparative review. *J Mach Learn Res* **10**, 66–71 (2009)
- [139] Venna, J., Kaski, S.: Neighborhood preservation in nonlinear projection methods: An experimental study. In: Artificial Neural Networks — ICANN 2001. 485–491. Springer Berlin Heidelberg, Berlin, Heidelberg (2001)
- [140] Venna, J., Peltonen, J., Nybo, K., Aidos, H., Kaski, S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research* **11(13)**, 451–490 (2010)
- [141] Vladymyrov, M., Carreira-Perpiñán, M.: Locally linear landmarks for large-scale manifold learning. In: Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 8190, 256–271 (09 2013)
- [142] Vogelstein, J., Park, Y., Ohshima, T., Kerr, R., Truman, J., Priebe, C., Zlatic, M.: Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. *Science (New York, N.Y.)* **344** (03 2014)
- [143] W., T.L.: An Introduction to Manifolds. 2nd ed. New York: Springer (2011)
- [144] Wang, R., Zhang, X.: Capacity preserving mapping for high-dimensional data visualization. arXiv preprint arXiv:1909.13322 (2019)

- [145] Witten, D.M., Tibshirani, R.: Supervised multidimensional scaling for visualization, classification, and bipartite ranking. *Computational Statistics and Data Analysis* **55**(1), 789–801 (2011)
- [146] W.S., T.: *Theory and methods of scaling*. Wiley (1958)
- [147] Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms (2017), <https://github.com/zalandoresearch/fashion-mnist>
- [148] Xu, X., Zhou, X.: d-sne: Domain adaptation using stochastic neighborhood embedding (2019)
- [149] Yang, L.: Data embedding techniques and applications. *ACM Proceedings of the 2nd international workshop on Computer vision meets databases* 29–33 (2005)
- [150] Yang, X., Lin, Y., Liu, R., He, Z., Wang, C., Dong, J.S., Mei, H.: Deepvisualinsight: Time-travelling visualization for spatio-temporal causality of deep classification training (2022)
- [151] Yianilos, P.N.: Data structures and algorithms for nearest neighbor search in general metric spaces. In: *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. 311–321. SODA '93, Society for Industrial and Applied Mathematics, USA (1993)
- [152] Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. 1601–1608. NIPS'04, MIT Press, Cambridge, MA, USA (2004)
- [153] Zhang, Y., Tiño, P., Leonardis, A., Tang, K.: A survey on neural network interpretability (2021)
- [154] Zhao, K., Lu, H., Mei, J.: Locality preserving hashing. *Proceedings of the AAAI Conference on Artificial Intelligence* **28**(1) (Jun 2014)