UMAP vs PaCMAP

A comparison of **dimensionality reduction** algorithms' abilities to **preserve local and global data structures**

Linnéa Gustafsson, linneag2@kth.se
DD2470 Advanced Topics in Visualization and Computer Graphics
KTH Royal Institute of Technology

Before we begin, a quick recap:

Dimensionality reduction (DR)

- High-dimensional space → low-dimensional space (in visualization, 2D/3D)

*"The goal of dimension reduction for data visualization is to take **high dimensional data** and **project it down to 2 or 3 dimensions so that humans can understand its structure**."* [1]

[1] *Dimension Reduction for Data Visualization*, Duke Course Notes, Cynthia Rudin

Now, let’s begin the presentation:

I. Research question

The question

How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

The question

How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

The question is based on “the PaCMAP paper”:
Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization [1]

[1] Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *The Journal of Machine Learning Research, 22*(1), 9129-9201.

The question

How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

Preservation of local and global structures

- There is **no strict definition** of local or global structure preservation
- One way to think about it:
  - **Local structure-preservation** methods
    - **preserve neighborhoods**, so that neighbors in the high-dimensional space are still neighbors in low-dimensional space
  - **Global structure-preservation** methods
    - **preserve overall relative placement of large clusters**

Dimension Reduction for Data Visualization, Duke Course Notes, Cynthia Rudin

Preservation claims of algorithms

- UMAP
  - Aims to preserve **local structure**
  - Preserves **more of the global structure *than other algorithms (e.g., t-SNE)*** [1]
- PaCMAP
  - Aims to preserve **both local and global structure** [2]

[1] McInnes, L, Healy, J, Melville, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018
[2] Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *The Journal of Machine Learning Research, 22*(1), 9129-9201.

The question

How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

The question

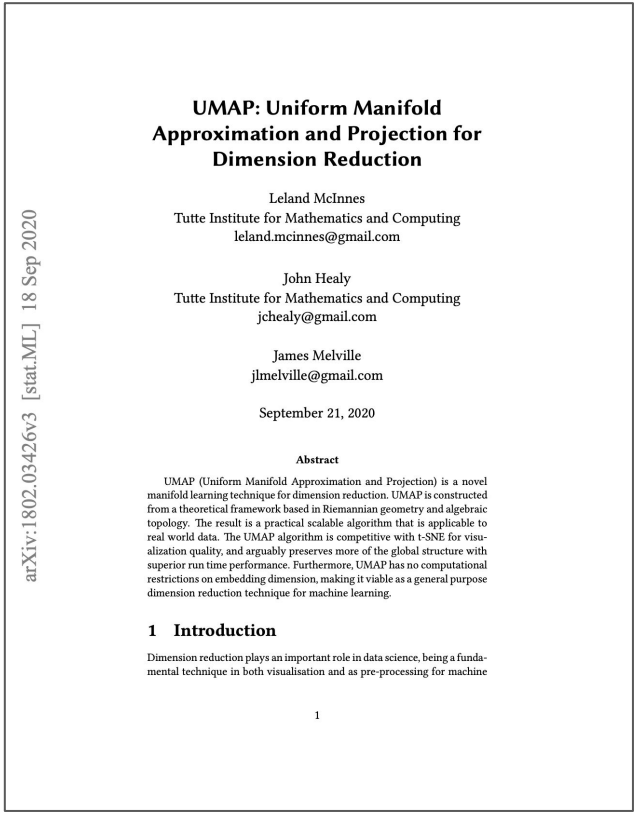
How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

UMAP

- UMAP: Uniform Manifold Approximation and Projection



UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Leland McInnes
Tutte Institute for Mathematics and Computing
leland.mcinnes@gmail.com

John Healy
Tutte Institute for Mathematics and Computing
jchealy@gmail.com

James Melville
jlmelville@gmail.com

September 21, 2020

Abstract

UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that is applicable to real world data. The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance. Furthermore, UMAP has no computational restrictions on embedding dimension, making it viable as a general purpose dimension reduction technique for machine learning.

1 Introduction

Dimension reduction plays an important role in data science, being a fundamental technique in both visualisation and as pre-processing for machine

1

McInnes, L, Healy, J, Melville, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018

UMAP

- Algorithm: *k*-neighbour based graph learning algorithm
  - Step 1: Graph construction (data → graph)
    - Construct local weighted graph based on distance between points (closer ⇔ larger weight
    - Patch graphs together ⇒ Fuzzy topological representation of the data
  - Step 2: Graph layout (graph → lower dimension)
    - Cost function is minimized

McInnes, L, Healy, J, Melville, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018

The question

How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

The question

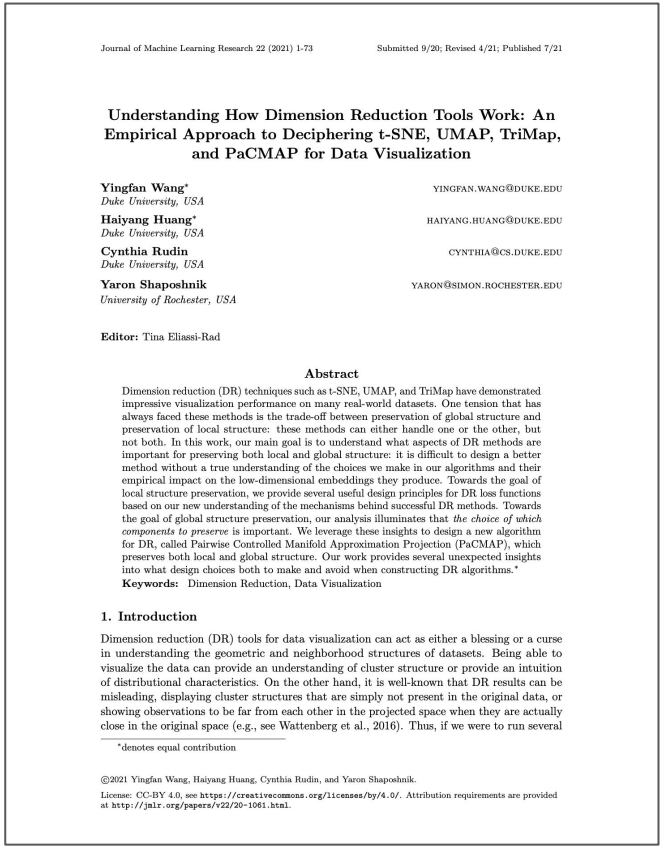
How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

PaCMAP

- PaCMAP: Pairwise Controlled Manifold Approximation Projection



Journal of Machine Learning Research 22 (2021) 1-73 Submitted 9/20; Revised 4/21; Published 7/21

Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization

Yingfan Wang* YINGFAN.WANG@DUKE.EDU
Duke University, USA
Haiyang Huang* HAIYANG.HUANG@DUKE.EDU
Duke University, USA
Cynthia Rudin CYNTHIA@CS.DUKE.EDU
Duke University, USA
Yaron Shaposhnik YARON@SIMON.ROCHESTER.EDU
University of Rochester, USA

Editor: Tina Eliassi-Rad

Abstract

Dimension reduction (DR) techniques such as t-SNE, UMAP, and TriMap have demonstrated impressive visualization performance on many real-world datasets. One tension that has always faced these methods is the trade-off between preservation of global structure and preservation of local structure: these methods can either handle one or the other, but not both. In this work, our main goal is to understand what aspects of DR methods are important for preserving both local and global structure: it is difficult to design a better method without a true understanding of the choices we make in our algorithms and their empirical impact on the low-dimensional embeddings they produce. Towards the goal of local structure preservation, we provide several useful design principles for DR loss functions based on our new understanding of the mechanisms behind successful DR methods. Towards the goal of global structure preservation, our analysis illuminates that *the choice of which components to preserve* is important. We leverage these insights to design a new algorithm for DR, called Pairwise Controlled Manifold Approximation Projection (PaCMAP), which preserves both local and global structure. Our work provides several unexpected insights into what design choices both to make and avoid when constructing DR algorithms.*

Keywords: Dimension Reduction, Data Visualization

1. Introduction

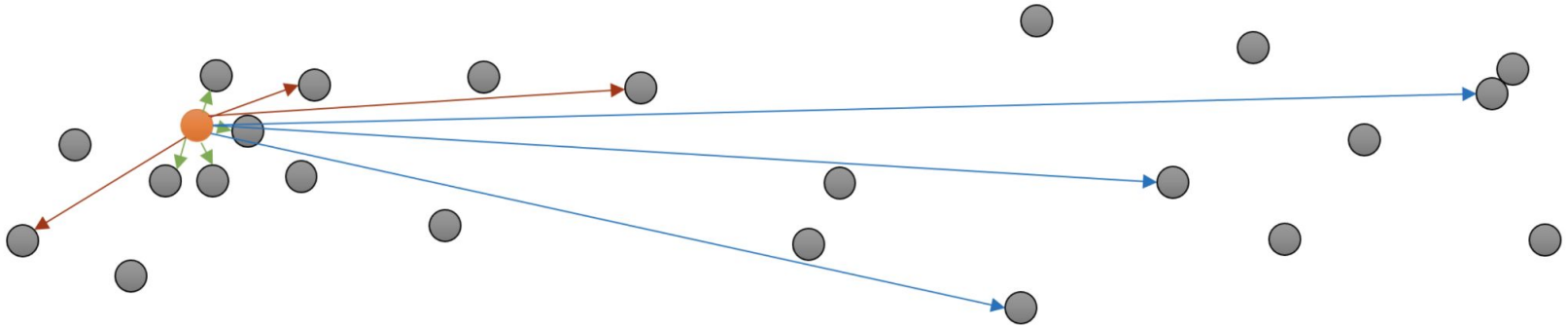
Dimension reduction (DR) tools for data visualization can act as either a blessing or a curse in understanding the geometric and neighborhood structures of datasets. Being able to visualize the data can provide an understanding of cluster structure or provide an intuition of distributional characteristics. On the other hand, it is well-known that DR results can be misleading, displaying cluster structures that are simply not present in the original data, or showing observations to be far from each other in the projected space when they are actually close in the original space (e.g., see Wattenberg et al., 2016). Thus, if we were to run several

*denotes equal contribution

Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *The Journal of Machine Learning Research, 22*(1), 9129-9201.

PaCMAP

- Algorithm
  - Step 1: Global structure optimization — *Preserve the **red** point pairs*
  - Step 2: Global and local optimization — *Preserve the **red** and **green** (and **blue**) point pairs*
  - Step 3: Refinement of local structure — *Preserve the **green** and **blue** point pairs*



Green: neighbors; **Red**: mid-near points; **Blue**: further points.

Dimension Reduction for Data Visualization, Duke Course Notes, Cynthia Rudin

The question

How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

The question

How does the performance of **UMAP** compare to **PaCMAP** in **preserving local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

Evaluation: Local structures

- KNN accuracy
  - Accuracy of KNN classifier applied to reduced dataset
- SVM accuracy
  - Accuracy of SVM classifier applied to reduced dataset

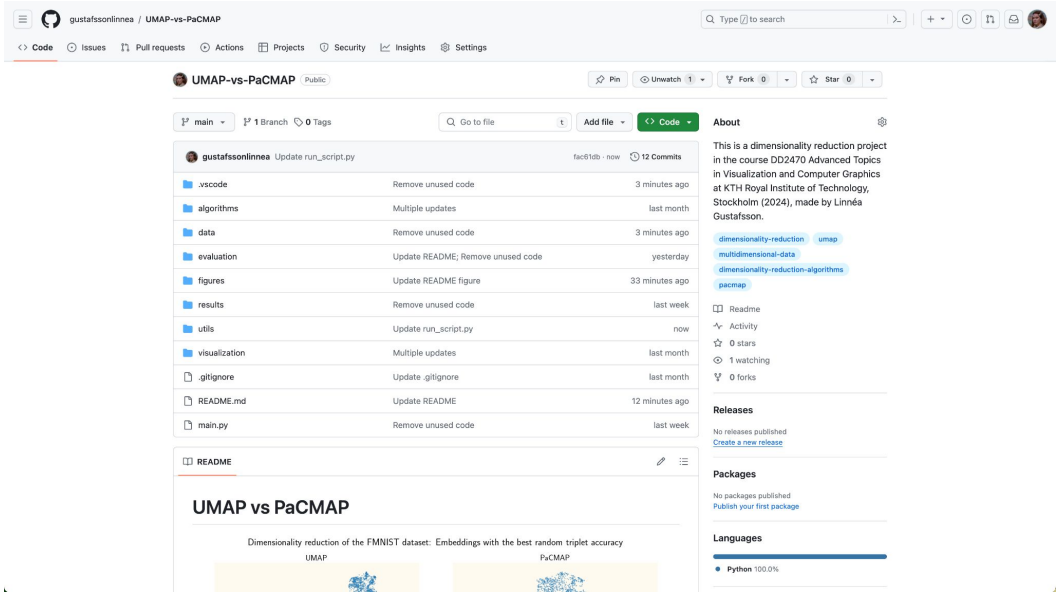
Evaluation: Global structures

- Random triplet accuracy
  - Percentage of triplets whose relative distance in the high- and low-dimensional spaces maintain their relative order
- Centroid triplet accuracy
  - Percentage of preserved centroid triplets

II. Implementation

Code

- Implemented in Python
- UMAP [1]
- PaCMAP [2]
- Data loading & two datasets
- Hyperparameter tuning
- Visualization
- Evaluation [3]
- Etc.



See details in GitHub repository (QR code in end slide):
https://github.com/gustafssonlinnea/UMAP-vs-PaCMAP

[1] https://umap-learn.readthedocs.io/en/latest/
[2] https://pypi.org/project/pacmap/
[3] https://github.com/YingfanWang/PaCMAP/blob/master/evaluation/evaluation.py

Datasets

- **8 different datasets** of different dimensionalities, sizes and number of classes
  - Iris [1]
  - Digits [2]
  - MNIST [3]
  - FMNIST [4]
  - CIFAR-10 [5]
  - CIFAR-100 [5]
  - Spiral3D created by me
  - Circle2D created by me
    - 2D → 2D

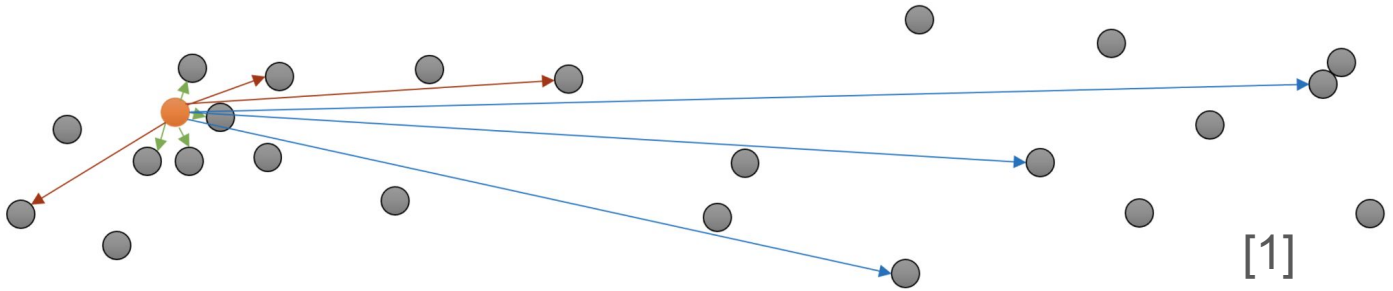
[1] https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html
[2] https://scikit-learn.org/stable/auto_examples/datasets/plot_digits_last_image.html
[3] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.
[4] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747.*
[5] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Hyperparameter tuning

Focus on trade-off between local and global structure preservation:

- UMAP
  - *n*, the number of neighbors to consider when approximating the local metric
- PaCMAP
  - *n*, the number of neighbors considered in the kNN graph

*The **green** point pairs*



[1]

$n \in \{5, 10, 20, 50, 100\}$

[1] *Dimension Reduction for Data Visualization*, Duke Course Notes, Cynthia Rudin

Evaluation

The final results = the best results of the hyperparameter search

(see which *n* in the visualizations)

III. Evaluation

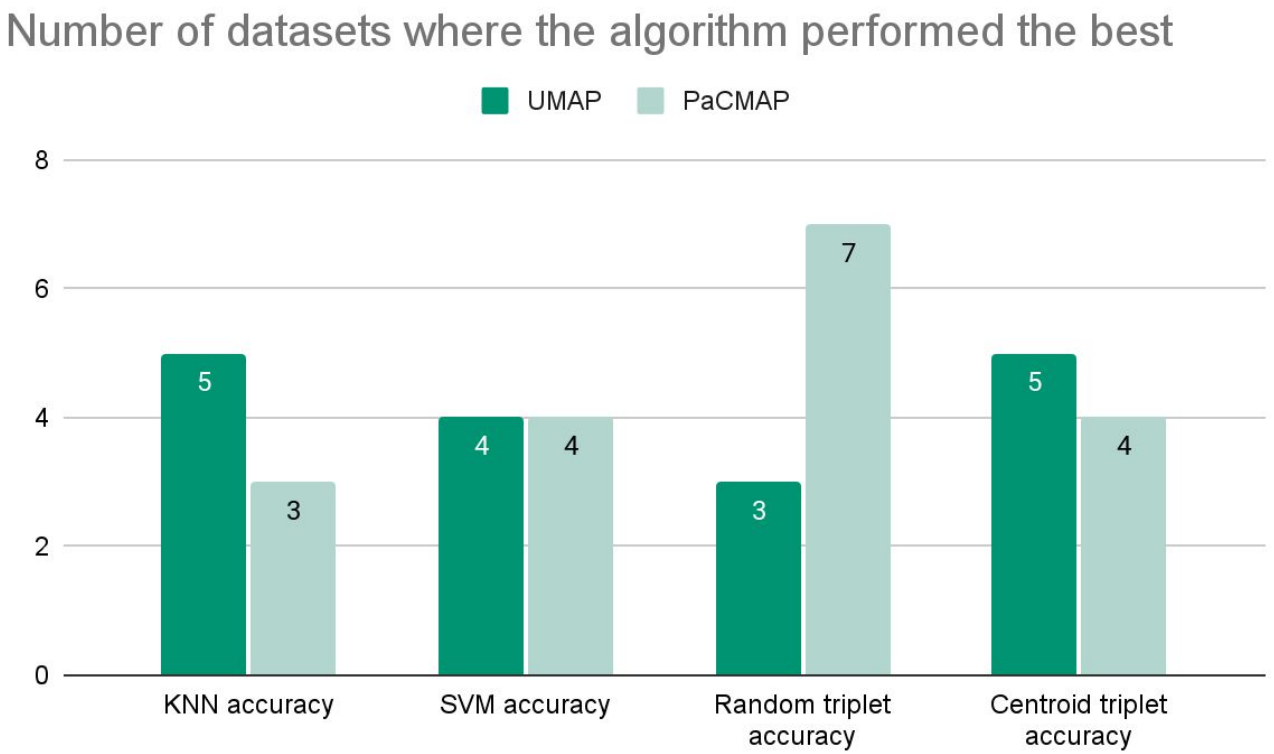
Local metrics: KNN and SVM accuracy

Dataset	Dimensions	Size	Classes	KNN: UMAP	KNN: PaCMAP	SVM: UMAP	SVM: PaCMAP
Iris	4	150	3	**0.980**	0.973	**0.973**	0.967
Digits	64	1 797	10	0.991	**0.992**	0.982	**0.984**
MNIST	784	70 000	10	0.969	**0.976**	0.967	**0.973**
FMNIST	784	70 000	10	**0.803**	0.773	0.742	**0.749**
CIFAR-10	3 072	60 000	10	**0.256**	0.235	0.253	**0.254**
CIFAR-100	3 072	60 000	100	**0.088**	0.068	**0.076**	0.074
Spiral3D	3	400	80	0.808	**0.818**	**0.648**	0.638
Circle2D	2	100	20	**0.830**	0.800	**0.800**	0.770

Global metrics: Random & centroid triplet accuracy

Dataset	Dimensions	Size	Classes	Random: UMAP	Random: PaCMAP	Centroid: UMAP	Centroid: PaCMAP
Iris	4	150	3	0.905	**0.927**	**1.000**	**1.000**
Digits	64	1 797	10	0.643	**0.683**	0.763	**0.786**
MNIST	784	70 000	10	0.588	**0.591**	**0.798**	0.696
FMNIST	784	70 000	10	**0.543**	0.540	**0.871**	0.859
CIFAR-10	3 072	60 000	10	**0.432**	**0.432**	**0.946**	0.937
CIFAR-100	3 072	60 000	100	**0.440**	**0.440**	**0.855**	0.851
Spiral3D	3	400	80	0.689	**0.700**	0.605	**0.682**
Circle2D	2	100	20	0.950	**0.990**	0.926	**0.968**

Comparison of results



For 3 of 4 metrics **they perform about equally well**,

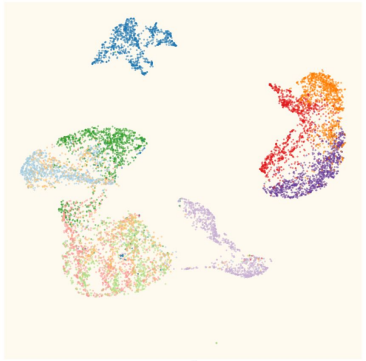
while for **random triplet accuracy, PaCMAP is remarkably better**.

Note: The differences in accuracy are often quite small.

Visualization examples: FMNIST dataset

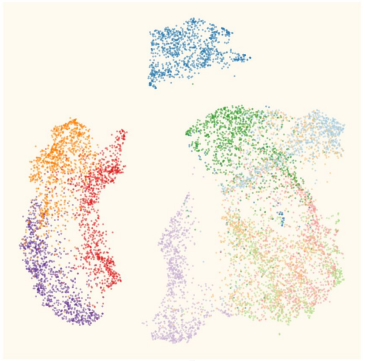
Dimensionality reduction of the FMNIST dataset: Embeddings with the best KNN accuracy

UMAP



$n = 5$

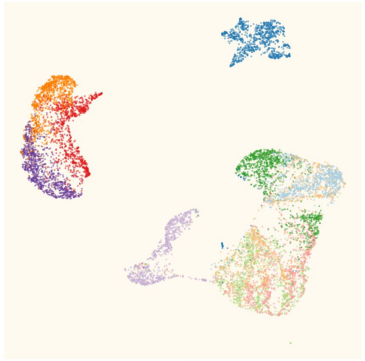
PaCMAP



$n = 5$

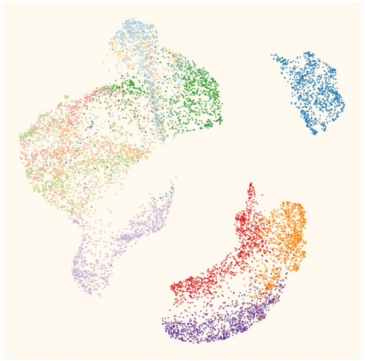
Dimensionality reduction of the FMNIST dataset: Embeddings with the best SVM accuracy

UMAP



$n = 10$

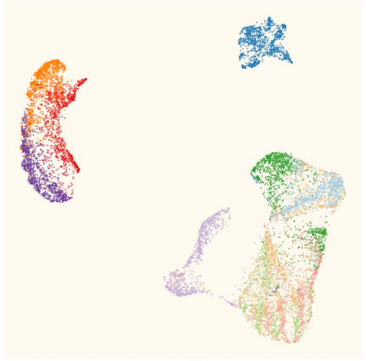
PaCMAP



$n = 10$

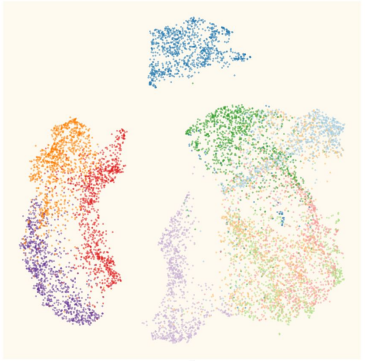
Dimensionality reduction of the FMNIST dataset: Embeddings with the best random triplet accuracy

UMAP



$n = 50$

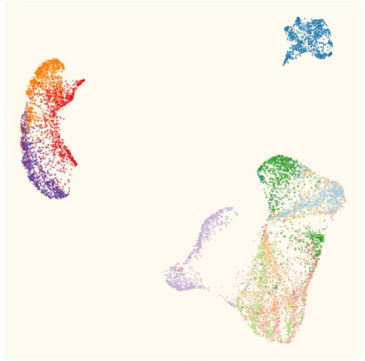
PaCMAP



$n = 5$

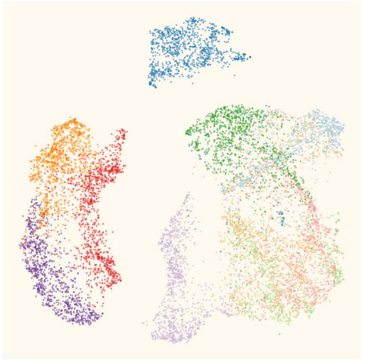
Dimensionality reduction of the FMNIST dataset: Embeddings with the best centroid triplet accuracy

UMAP



$n = 100$

PaCMAP

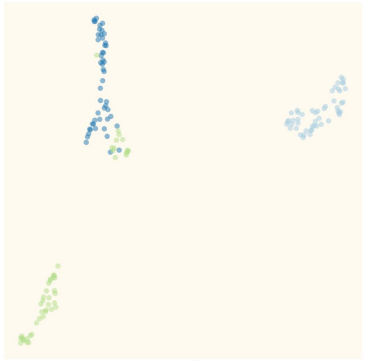


$n = 5$

Visualization examples: KNN accuracy

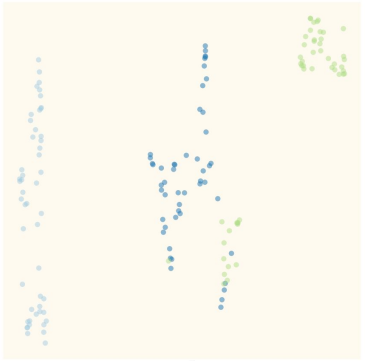
Dimensionality reduction of the Iris dataset: Embeddings with the best KNN accuracy

UMAP



$n = 10$

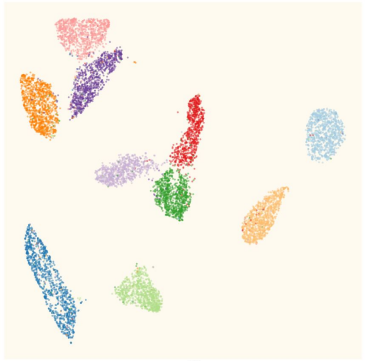
PaCMAP



$n = 5$

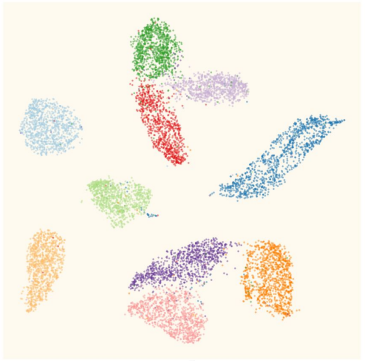
Dimensionality reduction of the MNIST dataset: Embeddings with the best KNN accuracy

UMAP



$n = 10$

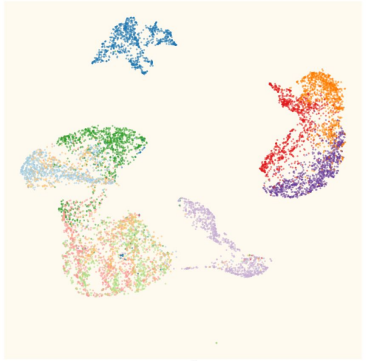
PaCMAP



$n = 5$

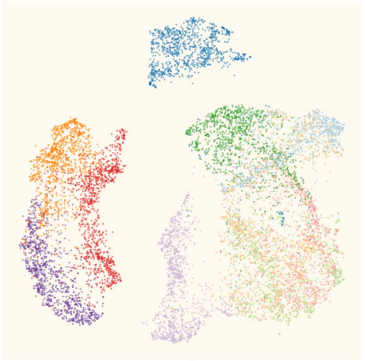
Dimensionality reduction of the FMNIST dataset: Embeddings with the best KNN accuracy

UMAP



$n = 5$

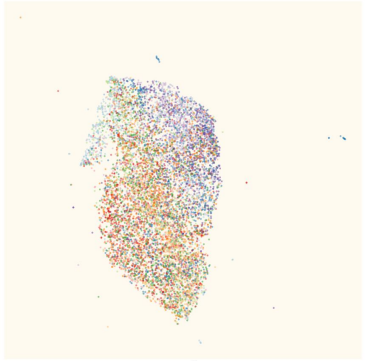
PaCMAP



$n = 5$

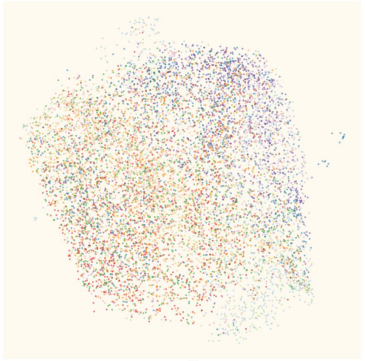
Dimensionality reduction of the CIFAR-10 dataset: Embeddings with the best KNN accuracy

UMAP



$n = 5$

PaCMAP

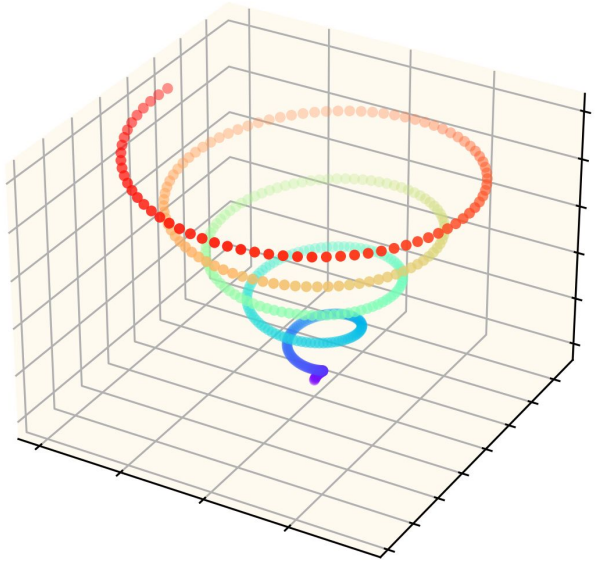


$n = 10$

Spiral3D dataset: Local preservation

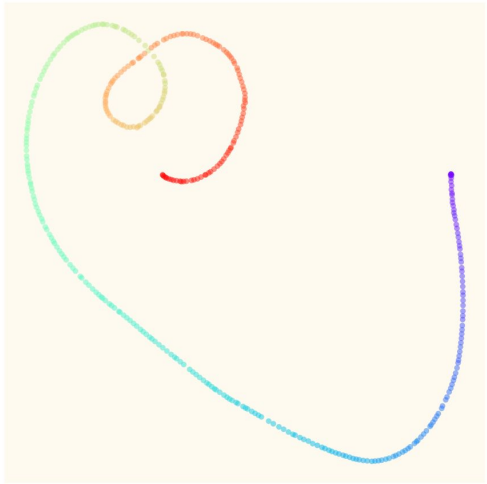
Same plot for SVM accuracy

Spiral3D



Dimensionality reduction of the Spiral3D dataset: Embeddings with the best KNN accuracy

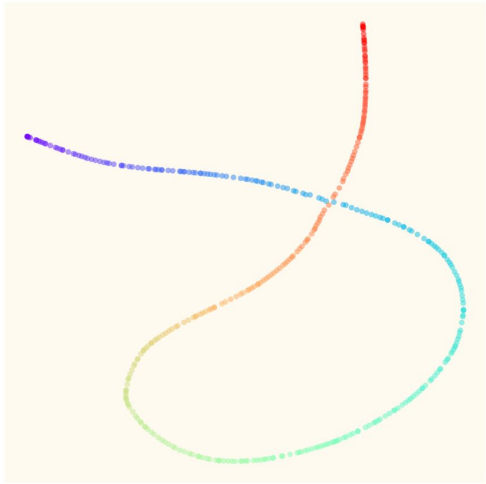
UMAP



$n = 10$

KNN accuracy: 0.808
SVM accuracy: **0.648**

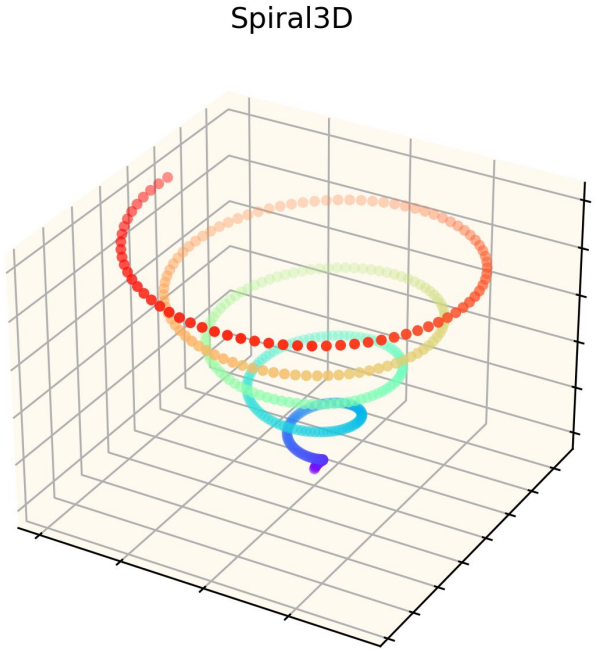
PaCMAP

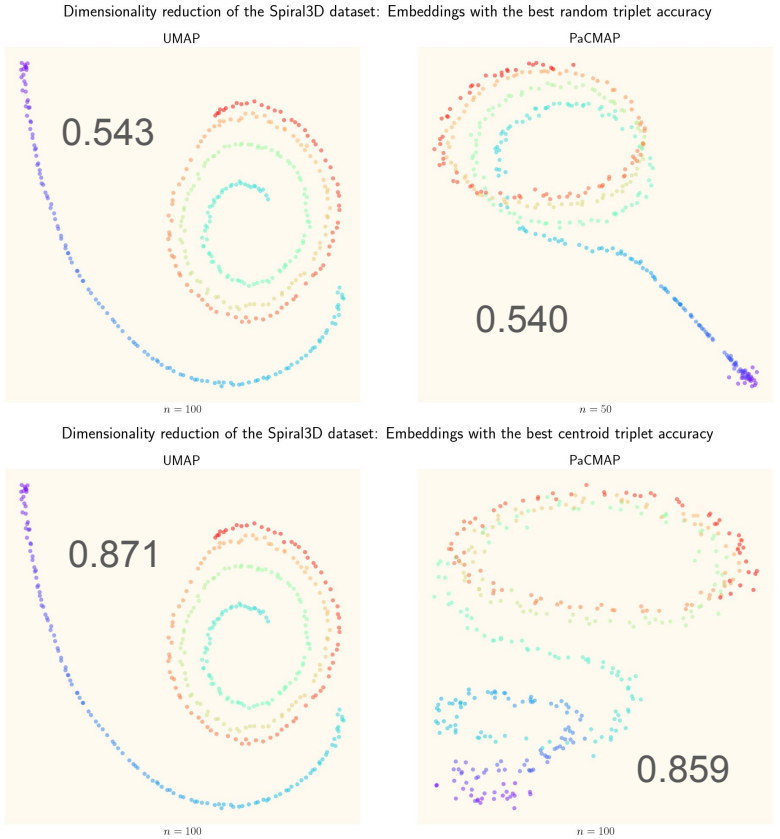


$n = 5$

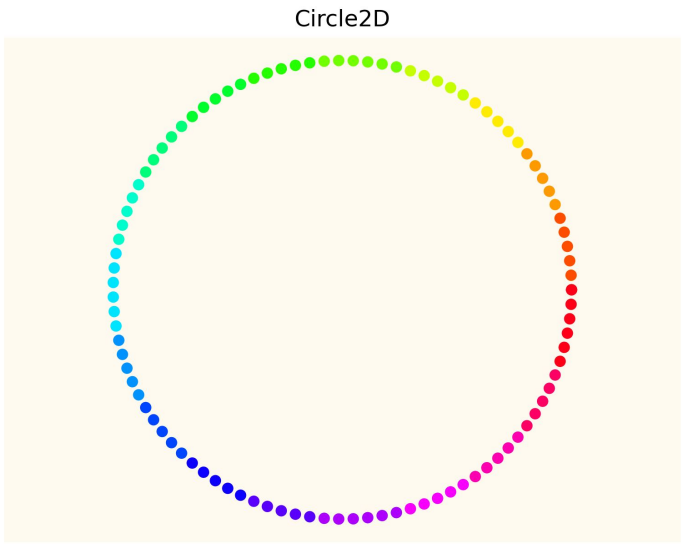
KNN accuracy: **0.818**
SVM accuracy: 0.638

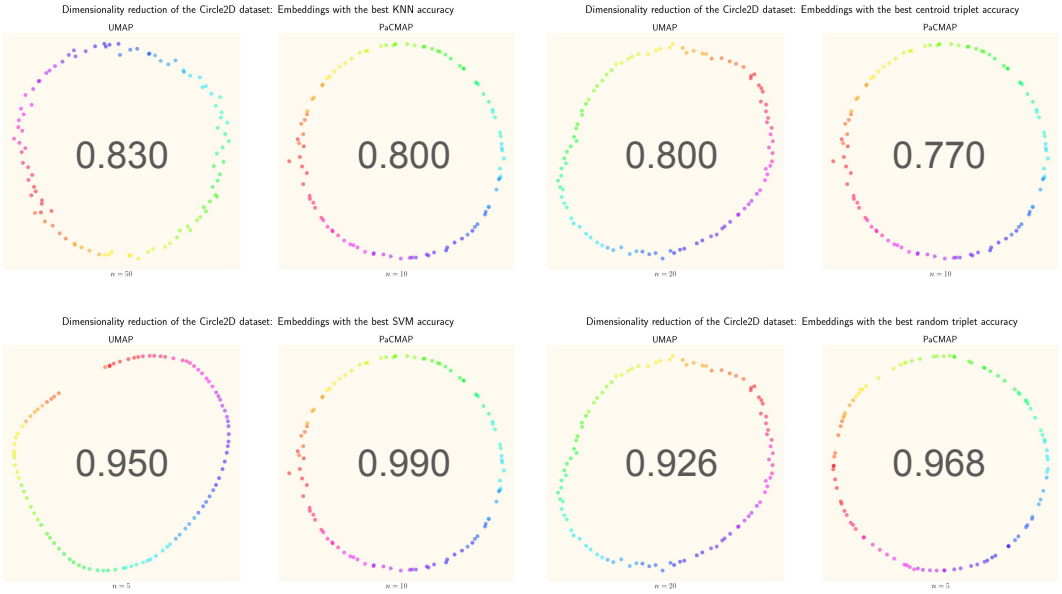
Spiral3D dataset: Global preservation





Circle2D dataset





Conclusion

How does the performance of **UMAP** compare to **PaCMAP** in **local and global structures**?

The **evaluation** will be based on ***KNN accuracy*** and ***SVM accuracy*** for **local structure preservation**,

and ***random triplet accuracy*** and ***centroid triplet accuracy*** for **global structure preservation**.

- **In my study, the algorithms generally perform equally well, except for random triplet accuracy where PaCMAP was better**
- **However, the method would need to be refined → Future work**

Future work

- **More careful hyperparameter search**
  - Coarse to fine search
  - Wider range (now I only perform coarse search in range 5-100)
  - Experiment with more hyperparameters
    - Initialization
    - Number of optimization epochs
    - Mid-near and further point ratio
    - Etc.
- Due to randomness in algorithms and evaluation: **Average several measurements** to get the metric values
  - Now, I only use one random seed and measure once
  - …but it takes a long time to run big datasets, so that's for another larger project
- Experiments on a **vast number of datasets** of different sizes etc. to be able to generalize conclusions

Learn more about the project

Slides

GitHub repository

UMAP paper

PaCMAP paper

Linnéa Gustafsson, linneag2@kth.se
DD2470 Advanced Topics in Visualization and Computer Graphics
KTH Royal Institute of Technology