An Exploration of PaCMAP

Abdelrahman ElJamal
Emory Salaun
Ayman Sandouk

November 2025

1 Problem Definition

The goal of this work is to deepen our understanding of dimensionality reduction (DR) techniques, with a particular focus on PaCMAP (Pairwise Controlled Manifold Approximation Projection). We aim to study PaCMAN in detail, understand how it can be implemented on a dataset, apply it to real data and use appropriate evaluation metrics to assess its effectiveness. Through this, we hope to gain insight into how DR methods are chosen for a particular dataset, how they are applied and how their performance is evaluated

2 Introduction

Real-world data is often high dimensional, which can make it difficult to process, visualize, and interpret. In order to handle it adequately, its dimensionality needs to be reduced. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation with fewer dimensions. Ideally, the reduced representation should match the intrinsic dimensionality of the original data while preserving important structure and relationships.

Dimension reduction (DR) techniques have demonstrated impressive visualization performance on many complex datasets. However, they often involve trade-offs, such as losing aspects of either local structure or global structure.

With this work, we intend to answer the following questions:

- How can we decide which datasets can benefit from a particular DR technique?
  - How can we correctly load and prepare a dataset for a DR technique?
- How can PaCMAP and other DR techniques be applied to different datasets?
- How can we effectively evaluate a particular DR technique?
  - How can we evaluate the local structure preservation of a DR technique?
  - How can we evaluate the global structure preservation of it?
- How can we effectively compare between DR techniques?

In this work, we refer to Wang, Yingfan, et al.'s paper, which explores how dimensionality reduction methods work and deciphers some of them, such as PaCMAP, for data visualization. The work is published as a GitHub repository with clear instructions on how to implement this method using the PaCMAP Python library that can be found in the repository.

3 Dataset

4 Methodology

5 Evaluation Techniques

In order to evaluate how well PaCMAP worked on our datasets, we chose four separate metrics to cover a variety of aspects of how well the low-dimensional embedding preserves the structure of the original data. Our metrics include k-Nearest-Neighbors (kNN) accuracy, trustworthiness, continuity, and Mean Relative Rank Error (MRRE)

5.1 kNN Classification Accuracy

This metric measures how well class information is preserved after the DR technique is applied. We train a k-nearest neighbor classifier on the original dataset, record the accuracy on how well it predicts labels in the test set. We then apply the dimensionality-reduction method, train a new classifier on the reduced data, and record its accuracy as well. A new accuracy value that is close to the original accuracy indicates that the dimensionality-reduction method successfully preserved the class-related structure of the data.

5.2 Trustworthiness

Trustworthiness evaluates whether the generated low-dimensional space introduces any false neighbors, which are points that look close in the reduced embedding, but were not close in the original dataset. Trustworthiness has a scale from 0-1, with a high score meaning the method did not create many false neighbors, therefore better preserving the local structure.

5.3 Continuity

Continuity checks for the opposite of what trustworthiness was looking for. Instead of looking for false neighbors, it checks for missing neighbors, points that were close in the original space but became far apart after dimensional reduction. Continuity also has a scale from 0-1, with a high score meaning the DR method keeps true neighbors together, even after projecting the data into fewer dimensions.

5.4 Mean Relative Rank Error (MRRE)

MRRE measures how much the ranking of neighbors changes after DR. Unlike kNN accuracy, instead of looking at whether a point is inside or outside the k-nearest neighborhood, MRRE looks at how far its ranking shifts, where a point's rank refers to how close it is to a given sample compared to all other points (the closest point has rank 1, the second-closest has rank 2, ...). Lower MRRE values indicate that the reduced embedding preserves the original neighborhood ordering more accurately.

5.5 Qualitative Analysis of 2D Embeddings

In addition to the quantitative metrics described above, we also include a qualitative evaluation step by visually examining the two-dimensional PaCMAP embeddings for each dataset. While numerical scores capture specific aspects of neighborhood preservation, the visual layout of the embedded points provides an interpretable assessment of how well the method organizes the data.

By plotting each dataset in the reduced 2D space and coloring points by class, we can directly observe cluster separability, overlap between classes, and global geometric structure. Datasets with well-preserved structure typically exhibit distinct, compact clusters that reflect the original class organization, whereas heavy overlap or diffuse regions indicate structural distortions introduced during the dimensionality-reduction process.

This qualitative inspection does not replace quantitative metrics, but it complements them by offering a interpretable view of the embedding's behavior. It allows us to visually confirm patterns suggested by kNN

accuracy, trustworthiness, continuity, and MRRE, and it often reveals structure—such as cluster shape or global arrangement—that is not fully captured by numerical measures alone.

6 Results

6.1 kNN Classification Accuracy

Table 1: KNN classification accuracy before and after PaCMAP embedding for each dataset.

Dataset	Stage	K=1	K=3	K=5	K=7	K=9
Coil20	Before	0.9931	0.9722	0.9340	0.8958	0.8889
	After	0.8438	0.8333	0.8264	0.8229	0.8333
MNIST	Before	0.9691	0.9705	0.9688	0.9694	0.9659
	After	0.9429	0.9598	0.9609	0.9612	0.9612
Olivetti	Before	0.9375	0.8750	0.8625	0.8250	0.7500
	After	0.4750	0.4500	0.5250	0.4750	0.4750

As shown in Table 1, the MNIST dataset exhibits the smallest performance reduction, maintaining more than 94% accuracy for all k values in the embedded space. This indicates that MNIST's digit classes remain well-separated even after dimensionality reduction, suggesting strong preservation of both global and local structure by PaCMAP.

Coil20 demonstrates a moderate accuracy drop (from 99.31% to 84.38% at $k = 1$). While the general class structure is preserved, fine-grained distinctions between visually similar objects become less separable in two dimensions.

In contrast, the Olivetti Faces dataset shows the most substantial degradation, with accuracy dropping from 93.75% to 47.50% at $k = 1$. Facial identity datasets require high-dimensional feature information to distinguish between individuals, and this structure is not well preserved when compressed to just two dimensions.

Overall, the results in Table 1 highlight that PaCMAP is highly effective on datasets with strong natural clustering (such as MNIST), moderately effective on structured object datasets (such as Coil20), and significantly less effective on datasets with subtle, high-dimensional class boundaries (such as Olivetti Faces). This reinforces the need to select dimensionality reduction methods and parameters with respect to the complexity of the dataset being analyzed.

6.2 Trustworthiness

6.3 Continuity

Table 2: Continuity values for each dataset at different neighborhood sizes K. Higher values indicate better preservation of global neighborhood structure.

Dataset	K=1	K=3	K=5	K=7	K=9
Coil20	0.9968	0.9938	0.9936	0.9928	0.9919
MNIST	0.9901	0.9874	0.9854	0.9837	0.9822
Olivetti	0.9740	0.9638	0.9448	0.9293	0.9180

Table 2 shows the Continuity values for Coil20, MNIST, and Olivetti across different neighborhood sizes K. All three datasets exhibit high Continuity scores, indicating that the PaCMAP embedding preserves a substantial portion of the global neighborhood structure. As expected, Continuity decreases slightly as K

increases, since larger neighborhoods make the metric more sensitive to global distortions introduced by the embedding.

Among the datasets, Coil20 achieves the highest Continuity values, remaining above 0.99 for all K. This suggests that PaCMAP maintains strong global relationships within this dataset, consistent with its relatively clean class structure and moderate size. MNIST also achieves high Continuity, with values above 0.98 across all K, demonstrating that global organization of digit classes is largely preserved even after projecting the data to two dimensions.

Olivetti shows the lowest Continuity scores, although the values remain relatively high overall (0.918–0.974). This reduction is consistent with the increased difficulty of organizing facial identity data in a low-dimensional space, where global relationships between individuals are more complex and less well separated.

Overall, these results indicate that PaCMAP effectively preserves global structural information across all datasets, with the strongest performance on Coil20 and MNIST.

6.4 Mean Relative Rank Error (MRRE)

Table 3 presents the MRRE values for MNIST, Coil20, and Olivetti. As expected, the absolute magnitude of MRRE differs substantially across datasets due to differences in dataset size. MNIST, with nearly 70,000 samples, naturally produces much larger MRRE values than Coil20 and Olivetti, since even small rank shifts correspond to changes across thousands of possible neighbor positions. In contrast, Coil20 (1,440 samples) and Olivetti (400 samples) yield considerably smaller MRRE values simply because the ranking space is far more limited.

These results reinforce that MRRE is most informative when used to *compare different dimensionality reduction techniques on the same dataset*, rather than for cross-dataset comparison. Within each dataset, the decreasing MRRE trend as K increases follows the expected pattern: smaller neighborhoods are more sensitive to local distortions introduced by dimensionality reduction, while larger neighborhoods are more tolerant of rank shifts.

Table 3: MRRE values for each dataset at different neighborhood sizes K. Lower values indicate better preservation of neighbor rank structure.

Dataset	**K=1**	**K=3**	**K=5**	**K=7**	**K=9**
MNIST	591.47	427.09	350.31	305.13	273.48
Coil20	3.70	3.98	3.32	2.98	2.75
Olivetti	8.28	6.75	6.63	6.41	6.09

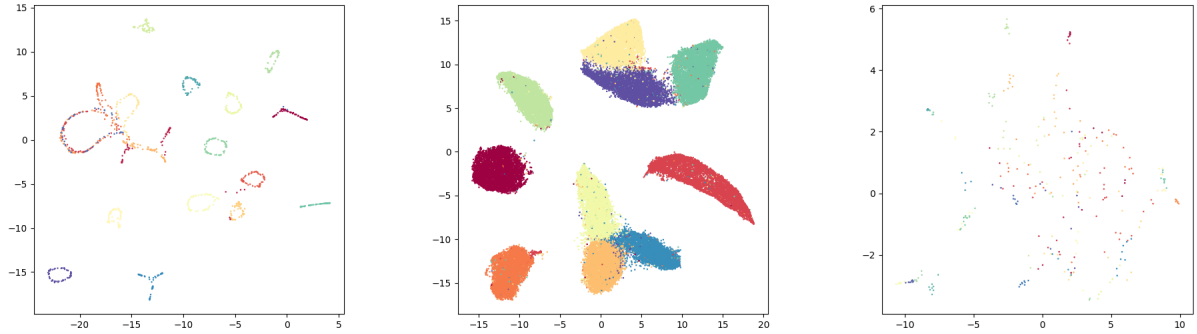
6.5 Qualitative Analysis of 2D Embeddings

Figure 1a, Figure 1b, and Figure 1c present the two-dimensional PaCMAP embeddings for Coil20, MNIST, and Olivetti respectively. These visualizations provide an intuitive understanding of how well PaCMAP preserves structure beyond what quantitative metrics alone can show.

The Coil20 embedding exhibits clearly separated, compact clusters, each corresponding to a different object category. The clusters form smooth, continuous loops that reflect the rotational nature of the dataset, and minimal overlap is observed between classes. This agrees with the high continuity values and the moderate post-embedding kNN performance, indicating that global and local structure are both well preserved.

The MNIST embedding also shows strong cluster structure, with distinct regions corresponding to each digit class. While some visually similar digits (such as 4 and 9, or 3 and 8) exhibit partial overlap, most clusters are dense and well formed. This visual clarity aligns with MNIST's high trustworthiness and continuity scores, as well as its relatively small drop in kNN accuracy after dimensionality reduction.

In contrast, the Olivetti embedding appears far more diffuse, with clusters that are less compact and more interwoven. Individual identities do not form tight groups, and many points lie between clusters or in areas of overlap. This visual fragmentation is consistent with the lower kNN accuracy and higher MRRE values for Olivetti, reflecting the difficulty of preserving facial identity information in only two dimensions.



(a) Coil20 (b) MNIST (c) Olivetti

Figure 1: PaCMAP embeddings for Coil20, MNIST, and Olivetti datasets.

Overall, the qualitative embeddings reinforce the quantitative evaluation: PaCMAP preserves structure effectively for Coil20 and MNIST, while Olivetti remains challenging due to its high intrinsic dimensionality and subtle class distinctions. Together, these visual and numerical results provide a comprehensive assessment of PaCMAP's performance across datasets of varying complexity.

7 Conclusion