CBMAP: Clustering-based manifold approximation and projection for dimensionality reduction

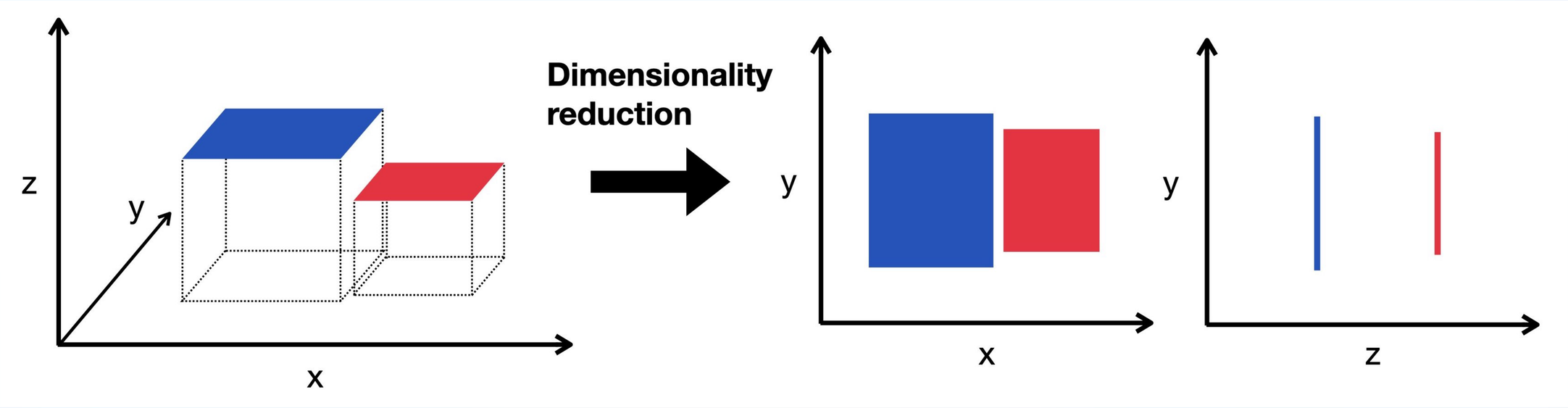
Author: Berat Doğan
Published Date: 16 Sep 2024
Presented by Chandini Saisri Uppuganti

Agenda

1. Introduction to Dimensionality Reduction
2. Existing Methods and Limitations
3. CBMAP: An Overview
4. Algorithm
5. Experimental Evaluation
6. Comparison with Existing Methods
7. Key Results and Insights
8. Conclusion

Introduction to Dimensionality Reduction

Reduce the number of input variables or features in a dataset while retaining its **essential information**. It transforms **high-dimensional data** into a lower-dimensional space to make it easier to analyze and visualize while improving **computational efficiency**.



Types of Dimensionality Reduction

1. *Feature Selection:* Selecting the most significant features.
2. *Feature Transformation:* Transforming data into a lower-dimensional space.

   Linear Methods: PCA, LDA.

   Nonlinear Methods: t-SNE, UMAP, TriMap, PaCMAP

Limitations:

1. Struggle with capturing nonlinear relationships.
2. Prioritize local structure over global structure.
3. Sensitive to hyperparameters (e.g., perplexity in t-SNE, n_neighbors in UMAP).
4. Computationally intensive for large datasets.

Method	Type	Strengths	Weakness	Use Cases
PCA	Linear	Computationally efficient, preserves global structure	Struggles with nonlinear relationships, lacks local structure preservation	Data compression, preprocessing for ML, visualizing linearly separable datasets
LDA	Linear	Maximizes class separability, ideal for classification tasks	Requires labeled data, assumes linear separability	Dimensionality reduction for supervised classification problems
t-SNE	Nonlinear (Manifold)	Excellent for visualizing high-dimensional data in 2D/3D, preserves local structure well	Computationally expensive, sensitive to hyperparameters (e.g., perplexity), poor global structure	Data visualization for clusters in high-dimensional datasets
UMAP	Nonlinear (Manifold)	Fast, preserves both local and some global structures, supports embedding test data	Sensitive to hyperparameters (n_neighbors, min_dist), less reliable on global structures	Visualizing complex datasets, neighbor-aware embedding tasks
TriMap	Nonlinear (Manifold)	Balances local and global structure, less sensitive to hyperparameters compared to t-SNE	Can still struggle with large datasets, moderate computational cost	Visualization of large datasets with a focus on preserving structure
PaCMAP	Nonlinear (Manifold)	Preserves pairwise distances, balances global and local structures	Sensitive to initialization, computationally expensive	Data visualization, particularly in biological or transcriptomics data

CBMAP: An Overview

- Utilizes clustering in high-dimensional space to identify clusters.
- Preserve both local and global structures which overcomes the limitations of other dimensionality reduction techniques.
- Provide reliable low-dimensional projections of test data.
- Reduce reliance on hyperparameters.
- Computationally efficient for large datasets.

CBMAP Algorithm

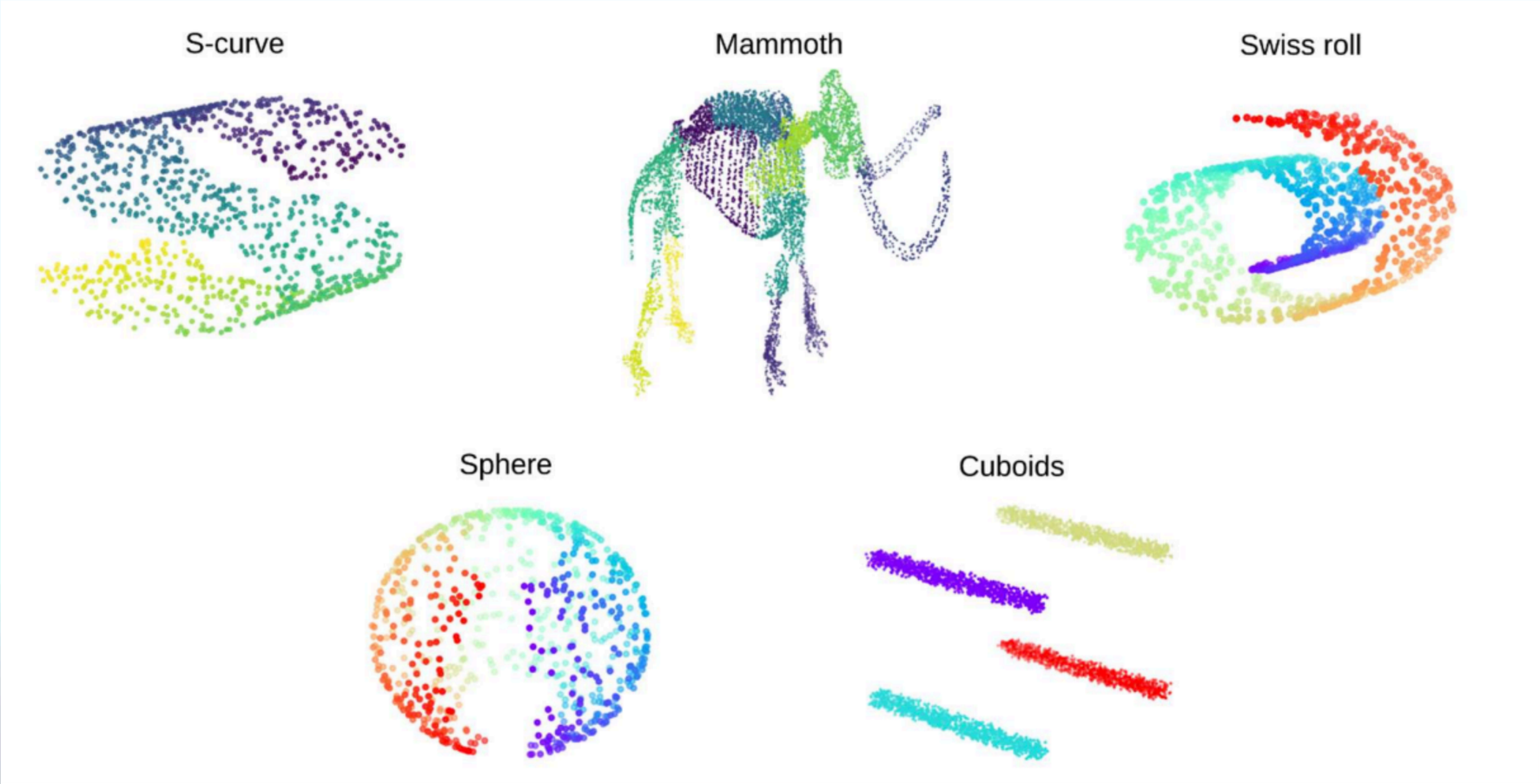
1. Clustering
2. Calculate Membership Values
3. Cluster Projection
4. Distance Calculation in Low-Dimensional Space
5. Generate Initial Low-Dimensional Points
6. Iterative Optimization (for iter =1 to max_iter)
7. Optimized low-dimensional representation

Experimental Evaluation

1. Datasets
    - Toy datasets - Tests to preserve global and local structures
    - Real world datasets - Evaluate scalability, accuracy, and clustering performance.

2. Performance Metrics
    - Global Score (GS)
    - k-Nearest Neighbors Accuracy (ACC)

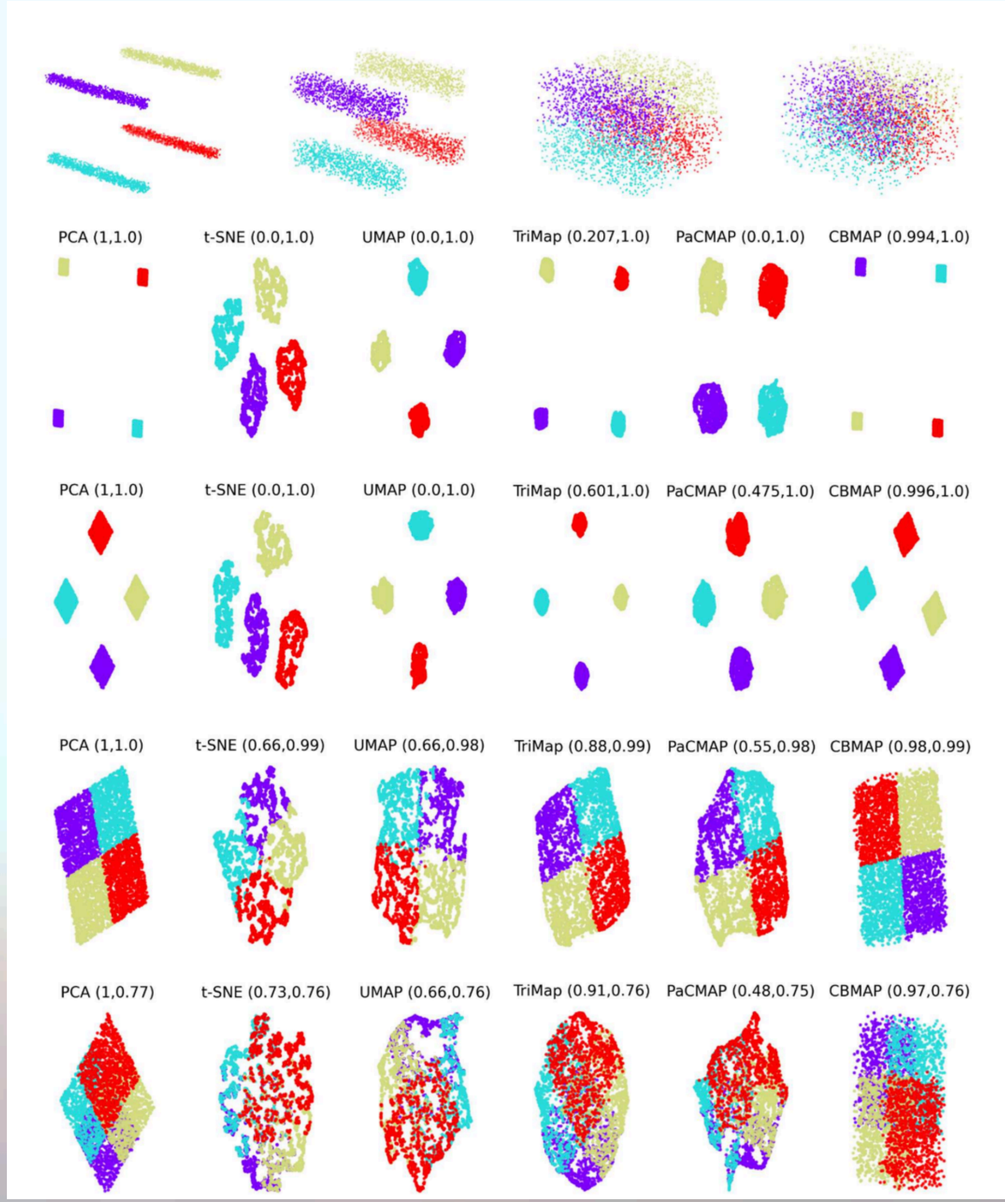
Toy datasets used in the experiments



A description of the real-world datasets used in the experiments.

Name	Size	Description
Iris	150×4	Three species of Iris flower (Iris setosa, Iris virginica and Iris versicolor) each consists of 50 samples. Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.
COIL-20	1440×1024	Gray-scale images of 20 objects in uniformly sampled orientations (5 degrees of rotation, 72 images per object). The size of each image is 32×32 pixels. Thus, each image is represented by a 1024-dimensional feature vector.
MNIST	70000×784	Images of handwritten digits (0–9) of size 28×28 each represented by a 784-dimensional feature vector.
Fashion MNIST	70000×784	Gray-scale images of clothing items such as t-shirt, pullover, bag, etc. of size 28×28 each represented by a 784-dimensional feature vector.
Duo 4Eq scRNA-seq	3994x100	Randomly selected B-cells, CD14 monocytes, naive cytotoxic T-cells, and regulatory T-cells were combined and filtered considering the average expression (log normalized), variability, and dropout effects. Next, the dimension is further reduced by PCA to 100.

Comparison with Existing Methods



Key Insights

Toy Datasets:

- CBMAP preserved both **global structure** and **local details** better than other methods.
- Higher GS values and clear visualization of structures (e.g., Mammoth's tusks in the Mammoth dataset).

Real-World Datasets:

- CBMAP outperformed PCA, UMAP, and t-SNE in global structure retention.
- Comparable or slightly lower ACC than t-SNE or UMAP, but reliable for test data projections.

Key Results

1. CBMAP retains the cluster topology and relative positions of the clusters to one another after dimensionality reduction
2. CBMAP’s ability to preserve local and global structural details was further confirmed by well-known benchmark toy datasets.
3. CBMAP’s performance on real-world benchmark datasets.
4. CBMAP allows dimensionality reduction for unseen test data

Computational time required for each method to project datasets into two-dimensional space.

	Time elapsed for each method (sec)					
Datasets	**PCA**	**t-SNE**	**UMAP**	**TriMap**	**PaCMAP**	**CBMAP**
Cuboids	0.017093	10.390414	11.859233	4.563494	2.893685	3.976686
S-curve	0.002586	2.359065	4.097905	1.433038	0.727454	0.952427
Mammoth	0.006881	19.003116	12.485656	9.597256	7.467309	9.990367
Swiss roll	0.002085	9.990367	3.522110	1.299816	0.733597	1.388023
Sphere	0.005694	2.763791	2.985690	0.983152	0.877690	0.939784
Iris	0.000508	1.369776	4.061811	0.309516	0.180864	0.606335
Coil-20	0.129517	3.824632	4.862549	2.802890	3.041111	3.126181
MNIST	1.123673	319.183082	37.511525	71.488276	48.22733	102.376454
Fashion MNIST	1.156796	326.614965	47.136507	70.206722	47.103224	102.332053
Duo 4Eq scRNA-seq	0.052447	13.919151	14.205324	4.211889	2.761129	4.424303

Conclusion

- Preservation of Global and Local Structures
- Scalability and Efficiency
- Minimal Hyperparameter Dependence
- Nearly parameter-free, reducing the need for extensive tuning.

Limitations

- **Dependency on k-Means Clustering:**
  - Performs best with normalized or normally distributed data.
  - Struggles with non-normal data distributions (e.g., count data).
- **Gaussian Membership Function:**
  - May not optimally handle skewed data distributions.
  - Increasing clusters helps but may require alternative membership functions for optimal performance.