LMAP: Local PCA Models with Global MDS Embeddings

Oliver Kramer

Computational Intelligence Group
Department of Computer Science
University of Oldenburg, Germany
oliver.kramer@uni-oldenburg.de

Abstract. This paper introduces LMAP (Local PCA Models with Global MDS Embeddings), a geometry-based framework for nonlinear dimensionality reduction that couples local tangent modeling with global distance preservation. Landmark points define locally linear PCA charts that approximate the manifold's tangent structure, and classical multidimensional scaling (MDS) aligns these charts into a coherent low-dimensional embedding. The resulting atlas admits a closed-form out-of-sample extension via weighted blending of multiple local models, producing a smooth and reproducible mapping. LMAP provides an explicit local–global coupling that integrates interpretable tangent information with global geometric relationships. Experiments on synthetic manifolds examine the influence of landmark density and neighborhood size, demonstrating that LMAP yields smooth, globally consistent embeddings with competitive trustworthiness and low distortion.

1 Introduction

High-dimensional datasets often concentrate near low-dimensional manifolds, motivating dimensionality reduction techniques that map data from $\mathbb{R}^D$ to $\mathbb{R}^d$ while retaining meaningful geometric structure. Classical manifold learners such as Isomap [1], LLE [2], and LTSA [3] exploit locally linear relationships but tend to struggle with curvature, noise, and scalability. More recent approaches such as UMAP [4] improve efficiency through probabilistic neighborhood graphs and stochastic optimization, yet often compromise smooth global geometry and reproducibility due to their non-convex embedding objectives.

A complementary research direction employs local PCA to characterize tangent geometry on nonlinear manifolds. Early work used local PCA for reconstruction or patch-based unfolding [5, 6], while curvature-aware variants such as CA-PCA [7] quantify how curvature distorts covariance spectra in intrinsic-dimension estimation. Other approaches use local PCA and tangent-space comparisons for geometry-aware clustering [8], showing that tangent discrepancies capture meaningful structure but do not yield a global embedding or an alignment of local charts. Although these methods provide accurate *local* approximations, they lack a mechanism for integrating tangent charts into a coherent *global* representation or for extending the resulting embedding to new points in a closed-form manner.

In the following, LMAP is introduced, a manifold learning framework that couples local tangent modeling with global geometric alignment.

2 LMAP

LMAP approximates a nonlinear manifold by constructing a set of locally linear tangent charts around selected landmark points and aligning these charts into a coherent low-dimensional coordinate system. Taken together, these aligned charts form an atlas of the manifold: each chart provides a valid local parameterization, and their overlaps ensure a smooth transition between regions. The atlas structure enables both a globally meaningful embedding and a principled out-of-sample mapping based on interpolating between local models.

2.1 Landmark Sampling

Given a dataset $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, LMAP selects a subset of $m \ll n$ landmark points $C = \{c_j\}_{j=1}^m$ that act as centers of local coordinate charts. Landmarks are chosen to cover the data distribution while keeping the global alignment step computationally efficient. Because subsequent operations scale cubically in the number of landmarks, reducing the problem from n points to m representative centers decreases complexity from $O(n^3)$ to $O(m^3)$ without discarding essential geometric information. Each landmark serves as a "reference location" around which the manifold is locally approximated and later contributes one chart to the atlas.

2.2 Local Tangent Modeling via PCA

For every landmark c_j, a neighborhood

$$P_j = \{x_i \mid x_i \text{ is among the } k_{\text{local}} \text{ nearest neighbors of } c_j\}$$

is determined in the ambient space. Applying PCA to P_j yields a tangent basis $T_j \in \mathbb{R}^{D\times d}$ that captures the dominant directions of variation and provides a first-order approximation of the manifold around c_j. These PCA-based tangent charts supply locally valid coordinate systems that approximate the nonlinear geometry in each region, and their collection forms the local structure of the atlas.

To characterize the scale of each neighborhood, we compute

$$\sigma_j = \text{median}\big(\|x_i - c_j\| : x_i \in P_j\big),$$

which later determines the strength of interpolation between overlapping charts. Smaller σ_j emphasize locality, while larger values yield smoother transitions.

2.3 Global Alignment via MDS

To merge the locally linear charts into a common global coordinate system, LMAP constructs a k_{graph}-nearest-neighbor graph over the landmarks using pairwise distances $d_{ij} = \|c_i - c_j\|$. Shortest-path distances on this graph approximate inter-landmark geodesic structure and define a distance matrix D_L that captures coarse global geometry.

Classical MDS is then applied to D_L,

$$Y_C = \mathrm{MDS}(D_L) \in \mathbb{R}^{m \times d},$$

producing coordinates for all landmark centers in the embedding space. This step aligns all tangent charts into a single global coordinate system, ensuring that the atlas reflects consistent large-scale geometric relationships.

2.4 Out-of-Sample Projection

General data points x_i are embedded by interpolating between multiple tangent charts. Let $\mathcal{N}(x_i)$ denote the q nearest landmarks. Each chart supplies an affine approximation $Y_C[j] + T_j^\top (x_i - c_j)$ of the embedded position, and these approximations are blended using Gaussian weights:

$$y_i = \frac{\sum_{j \in \mathcal{N}(x_i)} w_j(x_i) \left(Y_C[j] + T_j^\top (x_i - c_j)\right)}{\sum_{j \in \mathcal{N}(x_i)} w_j(x_i)}, \tag{1}$$

$$w_j(x_i) = \exp\Big(-\|x_i - c_j\|^2 / 2\sigma_j^2\Big), \tag{2}$$

where σ_j is the neighborhood scale of chart j.

This weighted blending provides a smooth transition across overlapping charts, yielding a continuous mapping from the ambient space to the global embedding defined by the atlas.

3 Experimental Analysis

This section presents an experimental evaluation of LMAP and examines the effects of its key hyperparameters

3.1 Landmark Density and Neighborhood Size

To analyze the influence of hyperparameters on geometric fidelity, we conduct a study varying the number of landmarks m, the local neighborhood size k_{local}, and the graph connectivity k_{graph} in the LMAP algorithm. Figure 1 shows the resulting embeddings and local tangent visualizations for nine parameter combinations. Local quality is measured using trustworthiness (TW) for neighborhood preservation, and global geometric accuracy using Sammon stress (SA) for distance distortion.

Increasing m improves global smoothness and manifold coverage, as more landmarks provide finer sampling of curvature and topology. Larger k_{local} values yield more stable but smoother tangent estimates, while smaller neighborhoods preserve sharper local variations. A higher k_{graph} enhances global consistency by connecting distant regions more effectively, reducing fragmentation at the cost

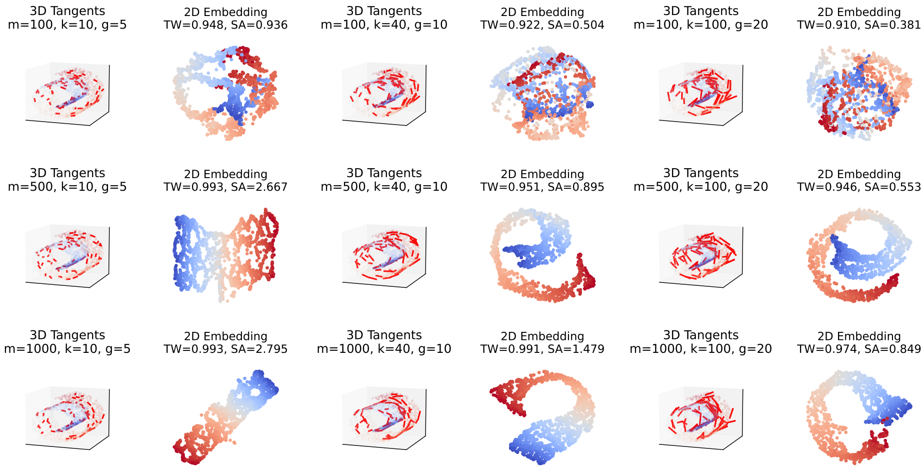

Fig. 1: Influence of the number of landmarks m, local neighborhood size k_{local}, and graph connectivity k_{graph} on the LMAP embedding. Each pair shows 3D local tangent models (left) and the resulting 2D embedding (right).

of slightly coarser local detail. Together, these parameters govern the trade-off between local geometric precision, curvature continuity, and global manifold coherence.

3.2 Quantitative Evaluation on High-Dimensional Data

To assess representational performance, two configurations of LMAP were evaluated: LMAP-1 ($m = 500$, $k_{\text{local}} = 10$, $k_{\text{graph}} = 5$) and LMAP-2 ($m = 1000$, $k_{\text{local}} = 40$, $k_{\text{graph}} = 10$). Table 1 reports TW and SA for PCA, MDS, t-SNE, UMAP, and both LMAP variants on three high-dimensional datasets.

Data	**PCA**	**MDS**	**LMAP-1**	**LMAP-2**	**t-SNE**	**UMAP**
Blobs	0.86/0.31	0.87/0.22	0.86/1.41	0.87/0.78	0.94/2.56	0.90/0.64
Cancer	0.87/0.29	0.89/0.19	0.88/0.95	0.87/0.33	0.95/2.82	0.92/0.57
Digits	0.82/0.61	0.89/0.29	0.88/2.22	0.90/1.59	0.99/3.66	0.97/0.54

Table 1: Comparison of PCA, MDS, LMAP-1, LMAP-2, t-SNE, and UMAP. Entries show TW/SA (higher TW, lower SA are better).

Across datasets, both LMAP variants achieve trustworthiness scores in the range 0.86–0.90, clearly below those of t-SNE and UMAP but higher than PCA and roughly on par with MDS. This indicates that LMAP preserves local neighborhoods reasonably well, but not at the level of stochastic neighbor-embedding methods. In terms of global structure, LMAP-2 substantially improves over

LMAP-1, reducing Sammon stress by roughly 40–50% across datasets. Nevertheless, its global distortion remains higher than PCA and MDS, which achieve consistently lower SA values. UMAP shows moderately low distortion, while t-SNE exhibits the highest distortions. Overall, LMAP-2 provides a meaningful trade-off between local and global structure: it preserves local neighborhoods better than PCA and MDS while maintaining notably lower global distortion than t-SNE. However, LMAP does not surpass UMAP or MDS in either metric, positioning it as a transparent mid-spectrum method rather than a replacement for state-of-the-art neighbor embeddings.

3.3 Comparison with Baseline Methods

Both variants are also compared with classical linear and nonlinear embedding techniques, PCA, MDS, t-SNE, and UMAP, on three nonlinear benchmark manifolds: Swiss Roll, Y-Branches, and Sparse Corridor. Figure 2 illustrates the re-

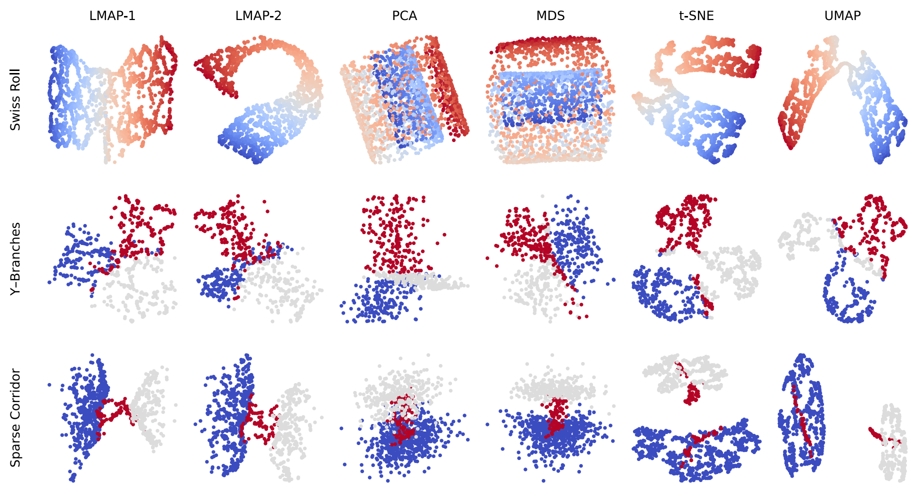

Fig. 2: Comparison of LMAP-1 and LMAP-2 with PCA, MDS, t-SNE, and UMAP on the Swiss Roll, Y-Branches, and Sparse Corridor datasets. LMAP produces globally coherent embeddings balancing local fidelity and global geometry.

sults, where each row corresponds to one dataset and each column to an embedding method. The first LMAP configuration uses fewer landmarks and smaller neighborhoods, while the second employs denser local models and higher graph connectivity. Both variants produce smooth embeddings that unfold the manifolds coherently, bridging the gap between global linear methods (PCA, MDS) and stochastic neighbor embeddings (t-SNE, UMAP), which emphasize local detail but often distort global geometry.

4 Conclusions

This paper introduces LMAP, a geometric framework for nonlinear dimensionality reduction that combines PCA-based tangent charts with global MDS alignment to obtain smooth embeddings with coherent local and global structure. The method further provides a closed-form out-of-sample extension via weighted blending of multiple tangent charts, enabling a continuous and reproducible mapping from the ambient space to the embedding.

Illustrative experiments on synthetic manifolds show that LMAP yields globally consistent embeddings with competitive quality. Trustworthiness is not perfect, as the blending of overlapping local neighborhoods slightly smooths and merges local structures, but this behavior is inherent to atlas-based constructions that prioritize continuity over strict neighbor preservation.

A particularly promising direction is a kernelized global alignment variant, in which both local charts and landmark relationships are expressed through kernel distances. Such a "Kernel-LMAP" formulation may capture complex nonlinear geometry more faithfully while preserving the framework's transparency and stability.

The open-source implementation of LMAP is available here.

References

[1] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[2] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[3] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004.

[4] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[5] Kitty Mohammed and Hariharan Narayanan. Manifold learning using kernel density estimation and local principal components analysis. *arXiv preprint arXiv:1709.03615*, 2017.

[6] Jonas Nordhaug Myhre, Matineh Shaker, Mustafa Devrim Kaba, Robert Jenssen, and Deniz Erdogmus. A generic unfolding algorithm for manifolds estimated by local linear approximations. In *CVPRW*, pages 3735–3743, 2020.

[7] Anna C. Gilbert and Kevin O'Neill. Ca-pca: Manifold dimension estimation, adapted for curvature. *SIAM Journal on Mathematics of Data Science*, 7(1):355–383, 2025.

[8] Emmanuel Arias-Castro, Gilad Lerman, and Teng Zhang. Spectral clustering based on local pca. *Journal of Machine Learning Research*, 18:1–57, 2017.