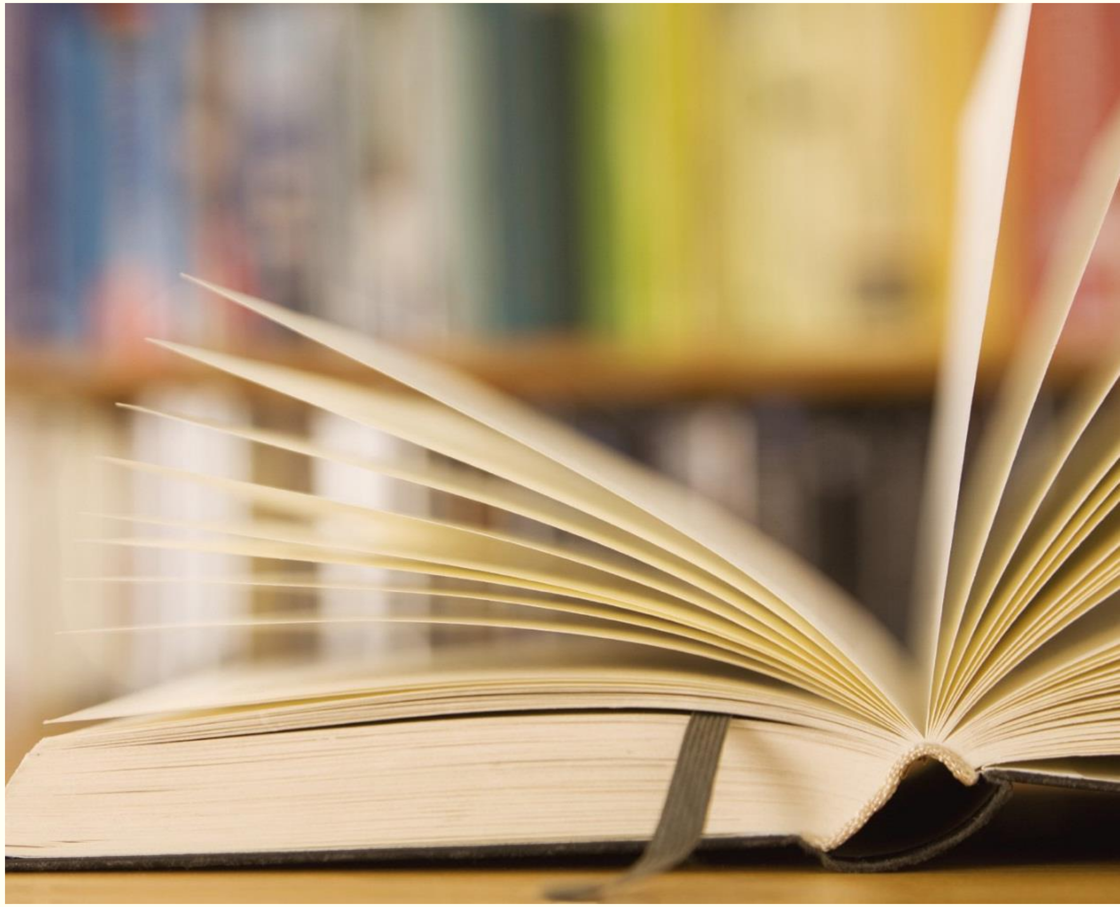
COMPARATIVE STUDY AND ANALYSIS OF DIMENSIONALITY REDUCTION TECHNIQUES



Joshua Ayanlowo | Mariama Musa | Steve Dalafu

Introduction

Machine learning practitioners are often confronted with high-dimensional feature spaces in which redundancy and noise impede modelling, visualization, and inference. Dimensionality reduction (DR) offers principled mappings from the original feature space to compact representations that retain salient structure. Foundational linear techniques such as Principal Component Analysis (PCA) minimize least-squares reconstruction error on centred data and provide variance-maximizing orthogonal bases, thereby delivering strong performance whenever data are well-approximated by linear subspaces. In contrast, non-linear DR methods seek embeddings that preserve manifold geometry beyond linear correlations, enabling faithful representations of curved or multi-modal structures that defy linear projection. (Gracia et al., 2014).

Velliangiri et al., (2019), examines DR in terms of computational efficiency, noise attenuation, and feature compactness, highlighting the importance of both feature extraction (transforming variables) and feature selection (identifying informative subsets) across applications ranging from medical imaging and bio-signals to text and remote sensing. DR is documented to reduce training time, remove redundancy, and improve classification accuracy, especially when integrated with deep learning or kernelized models.

Despite this progress, comparative evaluation of dimensionality reduction methods remains challenging because algorithms optimize different objectives (variance, neighborhood preservation, or probabilistic similarity) and exhibit trade-offs between global structure and local topology fidelity. Recent studies advocate multi-criteria evaluation frameworks that combine reconstruction accuracy, neighborhood preservation, and computational efficiency to assess the quality of transformed feature spaces and guide method selection.(Rastogi et al., 2023).

Introduction

In this context, our study contributes a controlled, end-to-end comparative analysis using three data regimes:

- linear data generated from latent linear mixtures
- non-linear data with curved manifold structure and element-wise non-linearities, and
- semi-linear data mixing linear and non-linear mechanisms.

We evaluate representative linear and non-linear methods across target dimensions, using a uniform reconstruction policy (native inverse transform when available, otherwise least-squares decoding from the embedding) and a panel of metrics. Our design aligns comparison to the generative assumptions – linear methods on linear data, manifold learners on non-linear data – and then tests robustness on semi-linear mixtures typical of real-world data

Problem Statement

High-dimensional datasets often contain redundant, correlated, or non-informative features that degrade model performance, and complicate visualization. While numerous dimensionality reduction (DR) algorithms exist, ranging from linear projections (e.g., PCA) to non-linear manifold learners (e.g., Isomap, t-SNE), their comparative performance depends strongly on the underlying data structure.

Real-world data rarely conforms strictly to linear or non-linear assumptions; instead, it often exhibits mixed or semi-linear characteristics. This raises critical questions:

- ❑ How do linear and non-linear DR methods perform when their assumptions align with the data-generating process?
- ❑ Which algorithms generalize best under semi-linear conditions?
- ❑ What metrics reliably capture reconstruction fidelity and structural preservation across varying target dimensions?

Addressing these questions is essential for guiding practitioners in selecting DR techniques that balance accuracy, interpretability, and computational efficiency.

Methodology

The study adopts a controlled experimental design implemented in Python and structured as follows:

1. DATA GENERATION

Three synthetic datasets were constructed to reflect distinct structural regimes:

- ❑ Linear dataset: Generated from five latent variables combined through linear mixing, augmented with correlated features and Gaussian noise.
- ❑ Non-linear dataset: Based on a Swiss-roll manifold with additional latent variables transformed via non-linear functions.
- ❑ Semi-linear dataset: Combines linear and non-linear features with correlated components and noise to mimic real-world complexity.

All datasets were centered to ensure comparability across algorithms.

Methodology

2. **ALGORITHMS EVALUATED**

Two families of DR methods were tested:

❑ Linear methods: PCA, Incremental PCA, Truncated SVD, Factor Analysis, Sparse PCA, MiniBatch Sparse PCA, Dictionary Learning, Gaussian and Sparse Random Projection.

❑ Non-linear methods: Isomap, Locally Linear Embedding (LLE), Spectral Embedding (Laplacian Eigenmaps), Kernel PCA (RBF kernel), and t-SNE (Barnes–Hut for 2–3D, Exact for higher dimensions).

Target dimensions: $k = \{2, 4, 6, 8\}$.

Methodology

3. **RECONSTRUCTION POLICY**

❑ If the algorithm provides inverse transform (e.g., PCA, KPCA), it was used.

❑ Otherwise, reconstruction employed a least-squares decoder:

$$\hat{x} = Z \cdot \mathrm{pinv}(Z) \cdot Xc$$

where Z is the low-dimensional embedding and Xc is the centered data. This ensures a uniform baseline for reconstruction across all methods.

Methodology

4. **EVALUATION METRICS**

Performance was assessed using:

- ❑ Frobenius norm: Total residual energy.
- ❑ Relative Frobenius norm: Scale-normalized error.
- ❑ Spectral norm: Largest singular value of residual (worst-case directional error).
- ❑ Mean Squared Error (MSE): Average entry-wise error.
- ❑ Reconstruction (R^2): Variance explained by reconstruction.

Methodology

5. **EXPERIMENTAL WORKFLOW**

For each dataset and target dimension:

Fit each DR algorithm.

Compute embeddings and reconstructions.

Calculate all metrics.

Aggregate results into:

Per-method tables (scores by dimension).

Average performance summaries.

Pivot tables for R^2.

Identify Top-3 performers per family (linear vs. non-linear) based on average R^2.

Re-evaluate these top methods on semi-linear data to test robustness.

Results Obtained

LINEAR DATA

On purely linear data, PCA and TruncatedSVD achieved the highest reconstruction fidelity across all dimensions, with average $R^2 \approx 0.94$ and near-perfect scores ($R^2 > 0.999$) at k ≥ 6. SparsePCA performed similarly but slightly lower, while Random Projection methods (GaussianRP, SparseRP) showed poor reconstruction, especially at k=2 ($R^2 \approx 0.09$).

These results confirm PCA’s theoretical least-squares optimality for linear variance structures and highlight that increasing k rapidly improves fidelity.

NON-LINEAR DATA

For non-linear manifold data, Isomap dominated with average $R^2 \approx 0.975$, outperforming SpectralEmbedding ($R^2 \approx 0.919$) and LLE ($R^2 \approx 0.916$). At k = 2, Isomap achieved $R^2 \approx 0.94$, while t-SNE and KernelPCA lagged significantly, confirming that t-SNE prioritizes local neighbourhood preservation for visualization rather than global reconstruction. Neighbourhood-based methods (Isomap, LLE) scaled well with dimension, reaching $R^2 \approx 0.98$ at k=8.

Results Obtained

SEMI-LINEAR DATA

On mixed data, linear methods (PCA, TruncatedSVD) dominated at higher dimensions ($R^2 \approx 0.998$ at k=8), indicating that linear variance components drive overall fidelity. However, at very low dimensions (k = 2), Isomap outperformed PCA ($R^2 = 0.637$ vs. $R^2 = 0.753$), suggesting non-linear structure matters when compression is extreme. SpectralEmbedding and LLE improved with dimension but never surpassed PCA beyond k=4.

KEY INSIGHTS

- ❑ For reconstruction/compression: PCA or TruncatedSVD are optimal
- ❑ for linear or semi-linear data; increasing kkk near intrinsic dimensionality yields rapid gains.
- ❑ For non-linear manifolds: Isomap is the best choice for balancing geometry preservation and reconstructability.
- ❑ For visualization: t-SNE should not be used for reconstruction; its low R^2 confirms its design for local neighborhood fidelity, not reversibility.
- ❑ Avoid Random Projections when fidelity matters, as they fail to capture dominant variance at low dimensions.

Conclusion

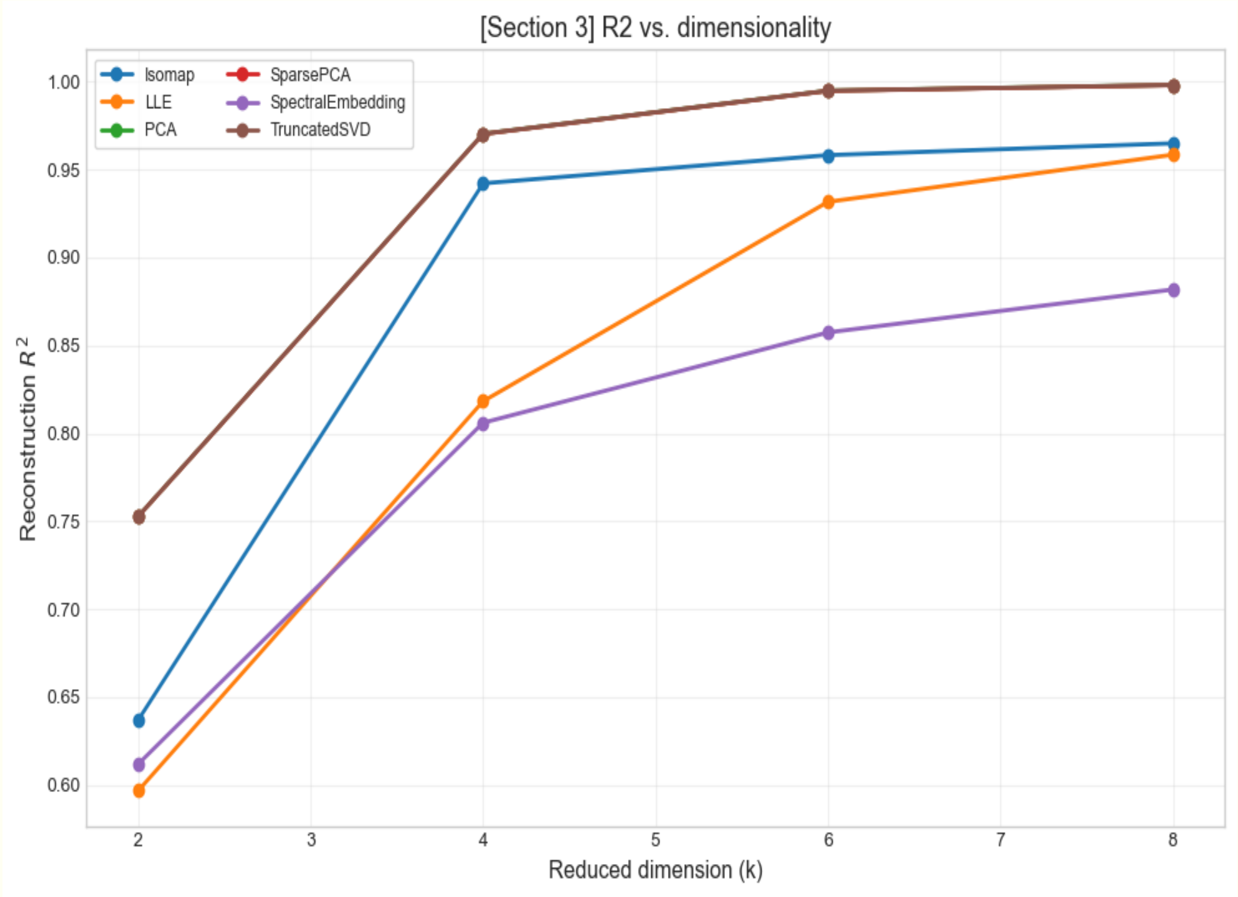
In conclusion, for real-world semi-linear data, combining a strong linear method with a robust non-linear method offers complementary strengths. Practitioners should select DR techniques based on data structure, target dimension, and goal using multi-metric evaluation rather than single-score rankings.

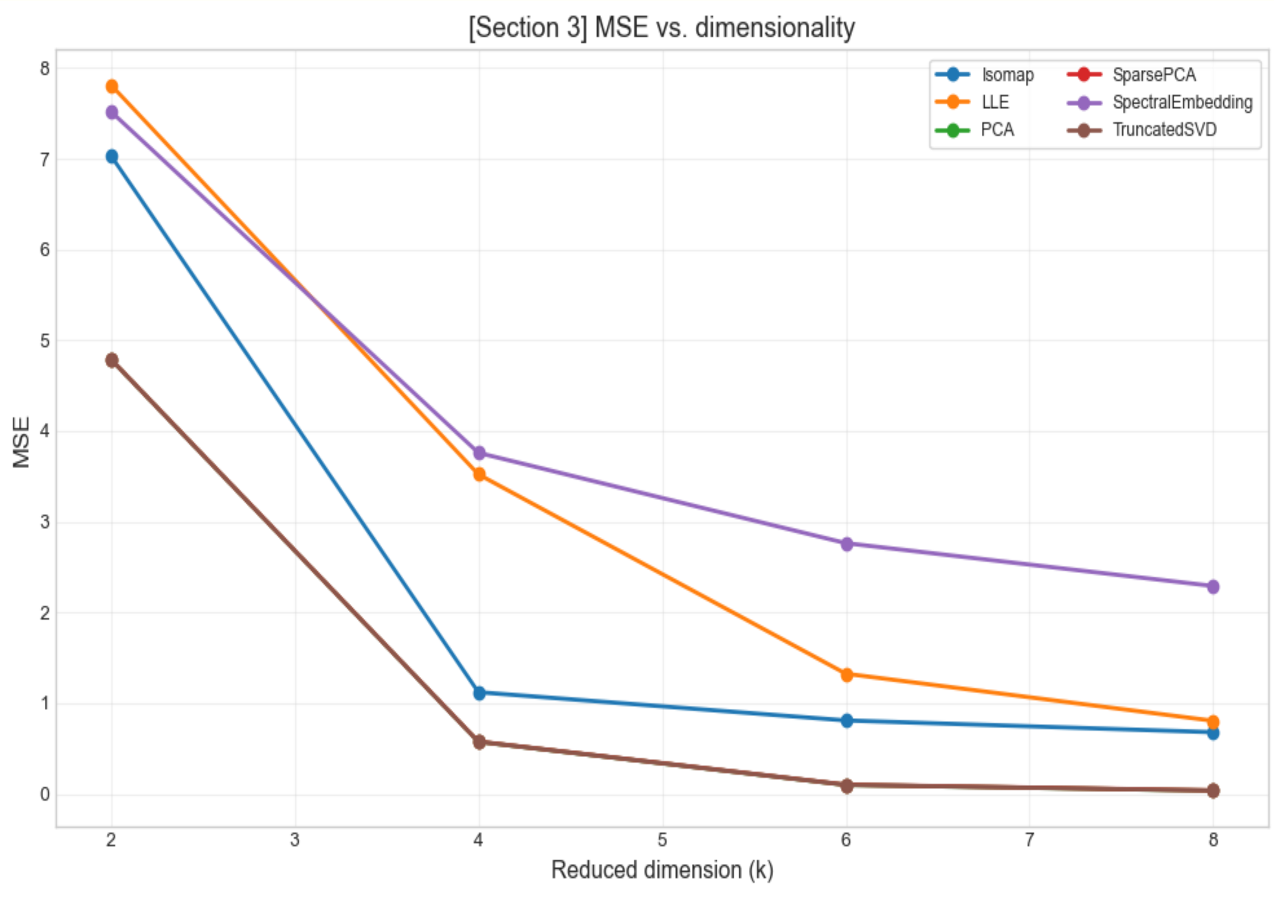
Tables and Figures

R^2 Score by Method and Dimension (Top-6)

DR Method	2	4	6	8
Isomap	0.636991	0.942148	0.958215	0.964908
LLE	0.596763	0.818228	0.931631	0.958429
PCA	0.752849	0.970330	0.994867	0.998119
SparsePCA	0.752775	0.970235	0.994769	0.998021
SpectralEmbedding	0.611908	0.805988	0.857372	0.881761
TruncatedSVD	0.752849	0.970330	0.994867	0.998119

Tables and Figures





References

❖Gracia, A., Gonzalez, S., Robles, V., Menasalvas, E., (2014). *A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality*.

http://dx.doi.org/10.1016/j.ins.2014.02.068

❖S.Velliangiri, S.Alagumuthukrishnan, S. Iwin, (2019*). A Review of Dimensionality Reduction Techniques For Efficient Computation*.

https://doi.org/10.1016/j.procs.2020.01.079

❖Rastogi, A.K., Taterh, S., Kumar, B.S., (2023). *Dimensionality Reduction Algorithms in Machine Learning: A Theoretical and Experimental Comparison.*

https://doi.org/10.3390/engproc2023059082

Thank You