Looking at Non-Linear Dimension Reductions as Models in the Data Space

Jayani P.G. Lakshika
Econometrics & Business Statistics, Monash University
and
Dianne Cook
Econometrics & Business Statistics, Monash University
and
Paul Harrison
MGBP, BDInstitute, Monash University
and
Michael Lydeamore
Econometrics & Business Statistics, Monash University
and
Thiyanga S. Talagala
Statistics, University of Sri Jayewardenepura

March 27, 2024

Abstract

Nonlinear dimension reduction (NLDR) techniques such as tSNE, and UMAP provide a low-dimensional representation of high-dimensional (high-D) data using non-linear transformation. The methods and parameter choices can create wildly different representations, making it difficult to decide which is best, or whether any or all are accurate or misleading. NLDR often exaggerates random patterns, sometimes due to the samples observed. But NLDR views have an important role in data analysis because, if done well, they provide a concise visual (and conceptual) summary of high-D distributions. To help evaluate the NLDR we have developed an algorithm to show the 2D NLDR model in the high-D space, viewed with a tour. One can see if the model fits everywhere or better in some subspaces, or completely mismatches the data. It is used to evaluate which 2D layout is the best representation of the high-D distribution and see how different methods may have similar summaries or quirks.

Keywords: high-dimensional data, dimension reduction, triangulation, hexagonal binning, low-dimensional manifold, manifold learning, tour, data vizualization

1 Introduction

- What's the problem:

Non-linear dimension reduction being used to summarise high-dimensional data.

- Summary of literature

Relevant high-d vis, NLDR history

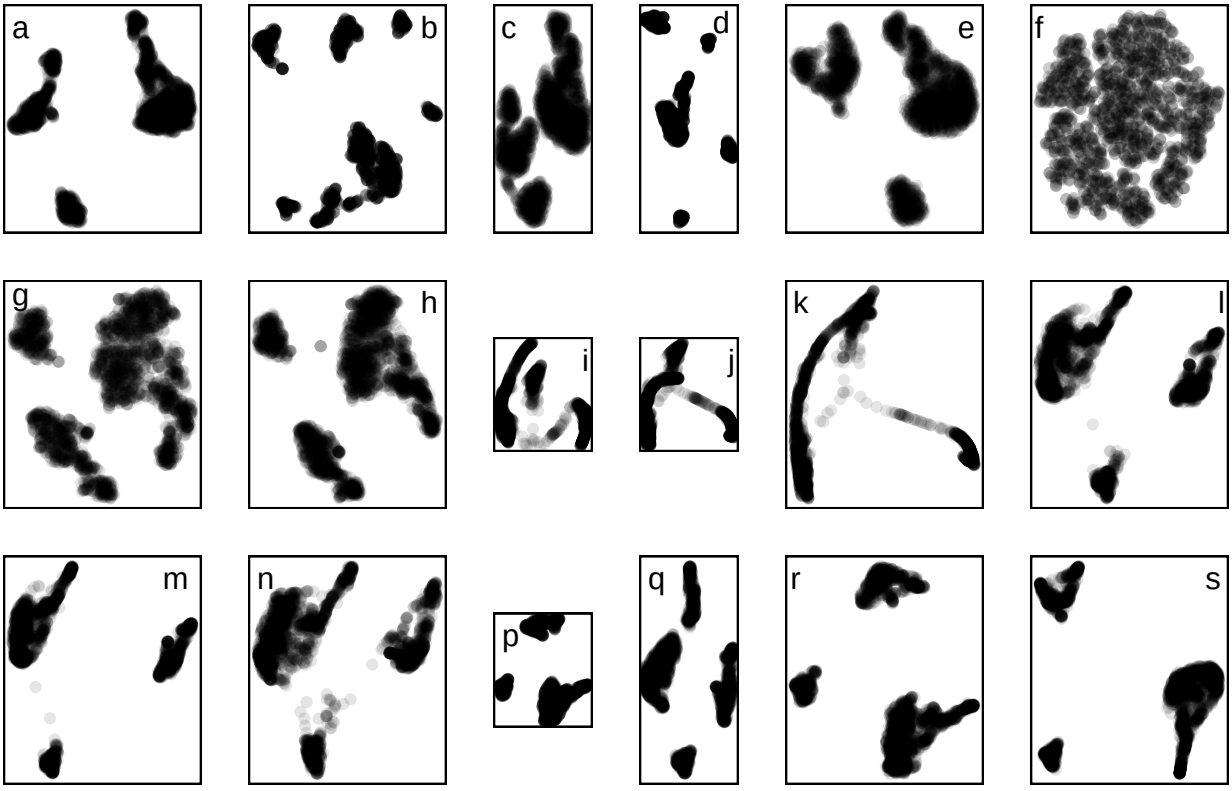

Figure 1: 2D layouts from different NLDR techniques and different hyperparameter choices applied for the PBMC3k dataset: (a) UMAP (n_neighbors = 30, min_dist = 0.3), (b) UMAP (n_neighbors = 5, min_dist = 0.01), (c) UMAP (n_neighbors = 15, min_dist = 0.99), (d) UMAP (n_neighbors = 15, min_dist = 0.03), (e) UMAP (n_neighbors = 84, min_dist = 0.99), (f) tSNE (perplexity = 5), (g) tSNE (perplexity = 51), (h) tSNE (perplexity = 92), (i) PHATE (knn = 2), (j) PHATE (knn = 5), (k) PHATE (knn = 10), (l) TriMAP (n_inliers = 12, n_outliers = 4, n_random = 3), (m) TriMAP (n_inliers = 30, n_outliers = 4, n_random = 10), (n) TriMAP (n_inliers = 5, n_outliers = 2, n_random = 2), (p) PaCMAP (n_neighbors = 10, init = pca, MN_ratio = 0.5, FP_ratio = 2), (q) PaCMAP (n_neighbors = 30, init = random, MN_ratio = 0.9, FP_ratio = 5), (r) PaCMAP (n_neighbors = 5, init = pca, MN_ratio = 0.1, FP_ratio = 1), and (s) PaCMAP (n_neighbors = 30, init = random, MN_ratio = 0.1, FP_ratio = 1). Is there a best representation of the original data or are they all providing equivalent information?

2 Method

- Create a representation of the model
- Algorithm in 2D
  - Parameters
  - Tuning
- Showing model in high-d
- What is learned about simulated examples

3 Applications

3.1 pbmc

```
- NLDR view used to illustrate clusters
- Use our method to assess is it a reasonable representation
- Demonstrate that it is not
- Illustrate how to use out method to get a better representation
```

3.2 digits: 1

```
- NLDR is used to illustrate different ways 1's are drawn
- Use our method to assess is it a reasonable representation
- Demonstrate that it is, except for the anomalies
```

4 Discussion

- Summarise contributions
- Explain where it is expected or not expected to work, eg higher dimensional relationships
- What might be useful enhancements

References