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 29, 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

Non-linear dimension reduction (NLDR) is popular for making a convenient low-dimensional ($k-D$) representation of high-dimensional ($p-D$) data. Recently developed methods include t-distributed stochastic neighbor embedding (tSNE) (van der Maaten & Hinton 2008), uniform manifold approximation and projection (UMAP) (McInnes & Healy 2018), potential of heat-diffusion for affinity-based trajectory embedding (PHATE) algorithm (Moon et al. 2019), large-scale dimensionality reduction Using triplets (TriMAP) (Amid & Warmuth 2022), and pairwise controlled manifold approximation (PaCMAP) (Wang et al. 2021). However, the representation generated can vary dramatically from method to method, and with different choices of parameters or random seeds made using the same method (Figure 1). The dilemma for the analyst is then, **which representation to use**. The choice might result in different procedures used in the downstream analysis, or different inferential conclusions. The research described here provides new visual tools to aid with this decision.

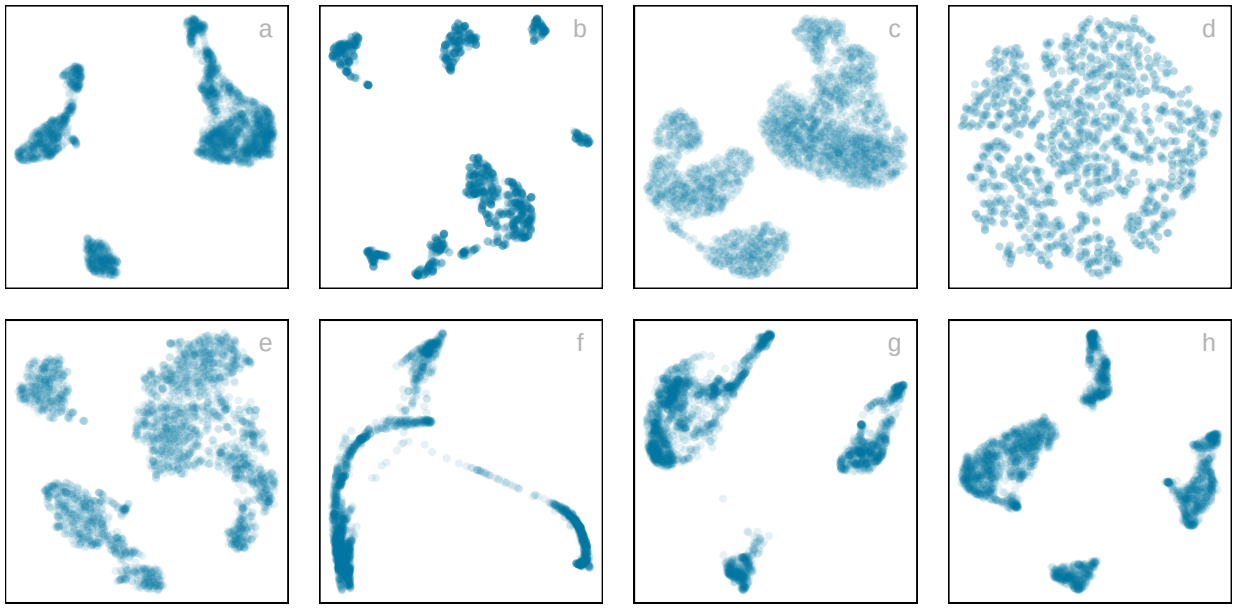

Figure 1: Six different NLDR representations of the same data. Different techniques and different parameter choices are used. Researchers may have seen any of these in their analysis of this data, depending on their choice of method, or typical parameter choice. Would they make different decisions downstream in the analysis depending on which version seen? Which is the most accurate representation of the structure in high dimensions?

The paper is organised as follows. Section 2 provides a summary of the literature on NLDR, and high-dimensional data visualization methods. Section 3 contains the details of the new methodology, including simulated data examples. Two applications illustrating the use of the new methodology for bioinformatics and image classification are in Section 4. Limitations and future directions are provided in Section 5.

2 Background

Historically, $k-D$ representations of $p-D$ data have been computed using multidimensional Scaling (MDS), which includes principal components analysis as a special case. The $k-D$ representation can be considered to be a layout of points in $k-D$ produced by an embedding procedure that maps the data from $p-D$. In MDS, the $k-D$ layout is constructed by minimizing a stress function that differences distances between points in $p-D$ with potential distances between points in $k-D$. Various formulations of the stress function result in non-metric scaling (Kruskal 1964) and isomap (Silva & Tenenbaum 2002). Challenges in working with high-dimensional data, including visualization, are outlined in Johnstone & Titterington (2009).

Many new methods for NLDR have emerged in recent years, all designed to better capture specific structures potentially existing in $p-D$. Here we focus on five currently popular techniques, tSNE, UMAP, PHATE, TriMAP and PaCMAP. tNSE and UMAP can be considered to produce the $k-D$ minimizing the divergence between two distributions, where the distributions are modeling the inter-point distances. PHATE, TriMAP and PaCMAP are examples of diffusion processes (Coifman et al. 2005) spreading to capture geometric shapes, that include both global and local structure.

- Tours
- Model-in-the-data-space: how can we represent the model, eg plane for PCA, grid of values for classification boundaries.

3 Method

- Overview: Generate a form that maps the model, that is the interpoint distances. What is the model?
- Notation
- Create a representation of the model
  - using hex-binning in 2D,
  - parameters,
  - tuning,
  - pre-processing
- How does this map to the representation in high-d
  - Centroids,
  - Edges
- Measuring fit
  - Fitted values
  - Error calculation
- What is learned about simulated examples
  - Interesting organisation of points in UMAP
  -

4 Applications

4.1 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
```

4.2 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
```

5 Discussion

- Summarise contributions
- Explain where it is expected or not expected to work, eg higher dimensional relationships
- Diagnostic app to explore differences in distances
- What might be useful enhancements

References

Amid, E. & Warmuth, M. K. (2022), 'Trimap: Large-scale dimensionality reduction using triplets'.

Coifman, R., Lafon, S., Lee, A., Maggioni, M., Nadler, B., Warner, F. & Zucker, S. (2005), 'Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps', *Proceedings of the National Academy of Sciences of the United States of America* **102**, 7426–31.

Johnstone, I. M. & Titterington, D. M. (2009), 'Statistical challenges of high-dimensional data', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1906), 4237–4253.
URL: *https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2009.0159*

Kruskal, J. B. (1964), 'Nonmetric multidimensional scaling: a numerical method', *Psychometrika* **29**(2), 115–129.

McInnes, L. & Healy, J. (2018), 'Umap: Uniform manifold approximation and projection for dimension reduction', *ArXiv* **abs/1802.03426**.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S. A., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G. & Krishnaswamy, S. (2019), 'Visualizing structure and transitions in high-dimensional biological data', *Nature Biotechnology* **37**, 1482 – 1492.

Silva, V. & Tenenbaum, J. (2002), 'Global versus local methods in nonlinear dimensionality reduction', *Advances in neural information processing systems* **15**.

van der Maaten, L. & Hinton, G. E. (2008), 'Visualizing data using t-sne', *Journal of Machine Learning Research* **9**, 2579–2605.

Wang, Y., Huang, H., Rudin, C. & Shaposhnik, Y. (2021), 'Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization', *Journal of Machine Learning Research* **22**(201), 1–73. **URL:** *http://jmlr.org/papers/v22/20-1061.html*