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, hexagonal binning, low-dimensional manifold, tour, data vizualization, model in the data space

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 (**?**), large-scale dimensionality reduction Using triplets (TriMAP) (**?**), 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.

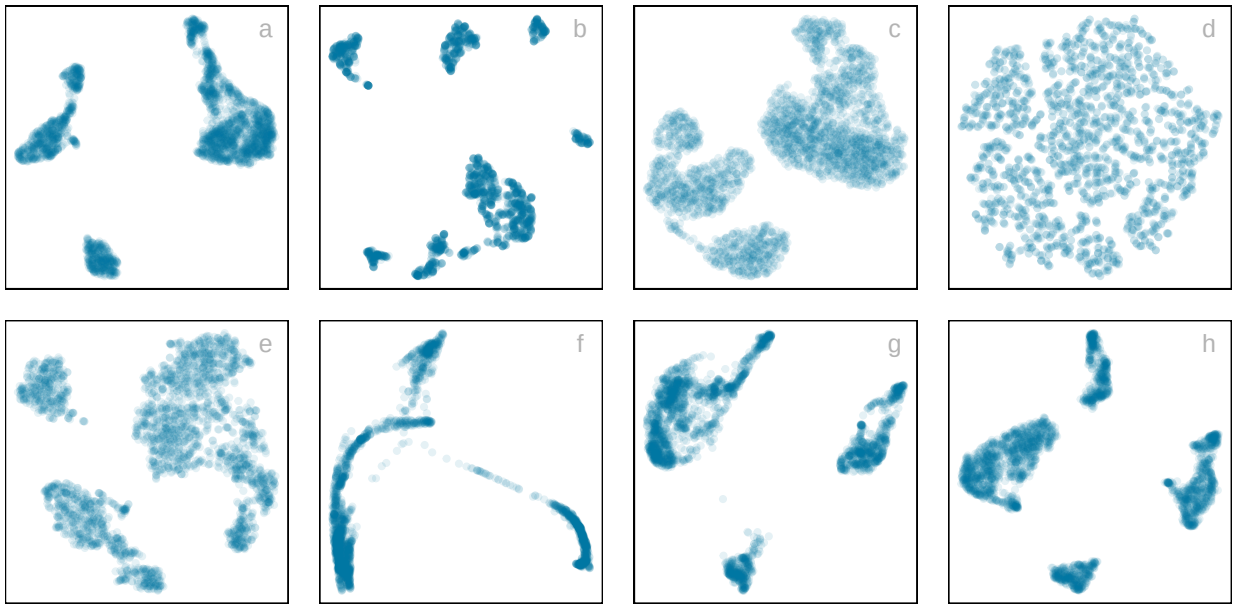

Figure 1: Six different NLDR representations of the same data. Different techniques and different parameter choices are used. Researchers may have seen any of these in their analysis of this data, depending on their choice of method, or typical parameter choice. Would they make different decisions downstream 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) (Borg & Groenen 2005), which includes principal components analysis (PCA) (Jolliffe 2011) 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 (Saeed et al. 2018) 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 (**?**) spreading to capture geometric shapes, that include both global and local structure.

The array of layouts in Figure 1 illustrate what can emerge from the choices of method and parameters, and the random seed that initiates the computation. Key structures interpreted from these views suggest:

- highly **separated clusters** (a, b, e, g, h) with the number ranging from 3-6,
- **stringy branches** (f), and
- **barely separated clusters** (c, d) which apparently **contradicts** the other representations.

It happens because these methods and parameter choices provide different lenses on the interpoint distances in the data.

The alternative approach to visualizing the high-dimensional data is to use linear projections. PCA is the classical approach, resulting in a set of new variables which are linear combinations of the original variables. Tours, defined by Asimov (1985), broaden the scope by providing movies of linear projections, that provide views the data from all directions. **?** provides an review of the main developments in tours. There are many tour algorithms implemented, with many available in the R package `tourr` (**?**), and versions enabling better interactivity in `langevitour` (**?**) and `detourr` (**?**). Linear projections are a safe way to view high-dimensional data, because they do not warp the space, so they are more faithful representations of the structure. However, linear projections can be cluttered, and global patterns can obscure local structure. The simple activity of projecting data from $p-D$ suffers from piling (Laa et al. 2022), where data concentrates in the center of projections. NLDR is designed to escape these issues, to exaggerate structure so that it can be observed. But as a result NLDR can hallucinate wildly, to suggest patterns that are not actually present in the data.

The solution is to use the tour to examine how the NLDR is warping the space. This approach follows what **?** describes as *model-in-the-data-space*. The fitted model should be

overlaid on the data, to examine the fit relative the spread of the observations. While this is straightforward, and commonly done when data is 2D, it is also possible in $p-D$, for many models, when a tour is used.

? provides several examples of models overlaid on the data in $p-D$. In hierarchical clustering, a representation of the dendrogrom using points and lines can be constructed by augmenting the data with points marking merging of clusters. Showing the movie of linear projections reveals shows how the algorithm sequentially fitted the cluster model to the data. For linear discriminant analysis or model-based clustering the model can be indicated by $(p-1)-D$ ellipses. It is possible to see whether the elliptical shapes appropriately matches the variance of the relevant clusters, and to compare and contrast different fits. For PCA, one can display the $k-D$ plane of the reduced dimension using wireframes of transformed cubes. Using a wireframe is the approach we take here, to represent the NLDR model in $p-D$.

3 Method

3.1 What is the NLDR model?

At first glance, thinking of of NLDR as a model fitted to the data might seem strange. It is a model in the sense that it is a "a simplified representation or abstraction of a system, process, or phenomenon in the real world". The $p-D$ observations are the realization of the phenomenon, and the $k-D$ NLDR layout is the simplified representation. From a statistical perspective we can consider the distances between points in the $k-D$ layout to be variance that the model explains, and the (relative) difference with their distances in $p-D$ is the error, or unexplained variance. Abstractly, we can also imagine that the positioning of points in 2D represents fitted values, that will have some prescribed position in $p-D$ that can be compared with the observed value.

3.2 Notation

« XXX Jayani, please insert a table of notation »

Once we have notation, we need to have some math that gives precision to the paragraph above. (XXX Michael??)

3.3 Constructing the 2D model

3.4 Displaying the model in $p-D$

3.5 Measuring the fit

3.6 What can be learned

- Overview: Generate a form that maps the model, that is the interpoint distances. What is the model?
- Notation
- Create a representation of the model

Table 1: Notations used in this paper.

Notation	Description
n	number of observations
p	number of dimensions in high-D data
$X_{n\times p} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{bmatrix}^\top$	high-D data matrix
d	number of dimensions in embedding data, usually two
$Y_{n\times d} = \begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n\end{bmatrix}^\top$	embedding matrix
P_t	point along the geodesic path at time t
P_{t+1}	point along the geodesic path at the next time instant, which is $t+1$
y_max	maximum value of the scaled second embedding component
r_1	range of the first embedding component
r_2	range of the second embedding component
h_b	height of the hexagon
w_b	width of the hexagon
a_r	aspect ratio of the 2D layout before scaling
h_r	hexagon ratio
s	hexagonal size (radius of the outer circle surrounding the hexagon)
(x_s, y_s)	starting coordinates of the hexagonal grid
q_x	buffer amount along the x-axis
q_y	buffer amount along the y-axis
h_s	horizontal distance between adjacent hexagon bin centroids
v_s	vertical distance between adjacent hexagon bin centroids
d_x	horizontal spacing
d_y	adjusted vertical spacing for hexagonal coordinates
b_x	number of hexagon bins along the x-axis
b_y	number of hexagon bins along the y-axis
b	total number of hexagon bins in the grid
b'	number of hexagon bins without the empty bins
n_k	number of observations within k^{th} hexagon
$(h^k_{x_i}, h^k_{y_i})$	hexagonal grid coordinates of k^{th} hexagon
$C^{(2)}_k \equiv (C_{ky_1}, C_{ky_2})$	hexagonal bin centroid coordinates of k^{th} hexagon
$C^{(p)}_k \equiv (C_{kx_1}, ..., C_{kx_p})$	high-D mappings of 2D hexagonal bin centroids of k^{th} hexagon

  - 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
- Human behaviour, the desire to have more certainty, and a tendency to prefer the well-separated views
- Diagnostic app to explore differences in distances
- What might be useful enhancements

References

Asimov, D. (1985), 'The grand tour: A tool for viewing multidimensional data', *SIAM Journal on Scientific and Statistical Computing* **6**(1), 128–143.
URL: *https://doi.org/10.1137/0906011*

Borg, I. & Groenen, P. J. F. (2005), *Modern Multidimensional Scaling Theory and Applications*, Springer, New York.

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*

Jolliffe, I. (2011), *Principal Component Analysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1094–1096.
URL: *https://doi.org/10.1007/978-3-642-04898-2_455*

Laa, U., Cook, D. & Lee, S. (2022), 'Burning sage: Reversing the curse of dimensionality in the visualization of high-dimensional data', *J. Comput. Graph. Stat.* **31**(1), 40–49.
URL: *https://doi.org/10.1080/10618600.2021.1963264*

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

Saeed, N., Nam, H., Haq, M. I. U. & Muhammad Saqib, D. B. (2018), 'A survey on multidimensional scaling', *ACM Comput. Surv.* **51**(3).
URL: *https://doi.org/10.1145/3178155*

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*