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

May 13, 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 (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.

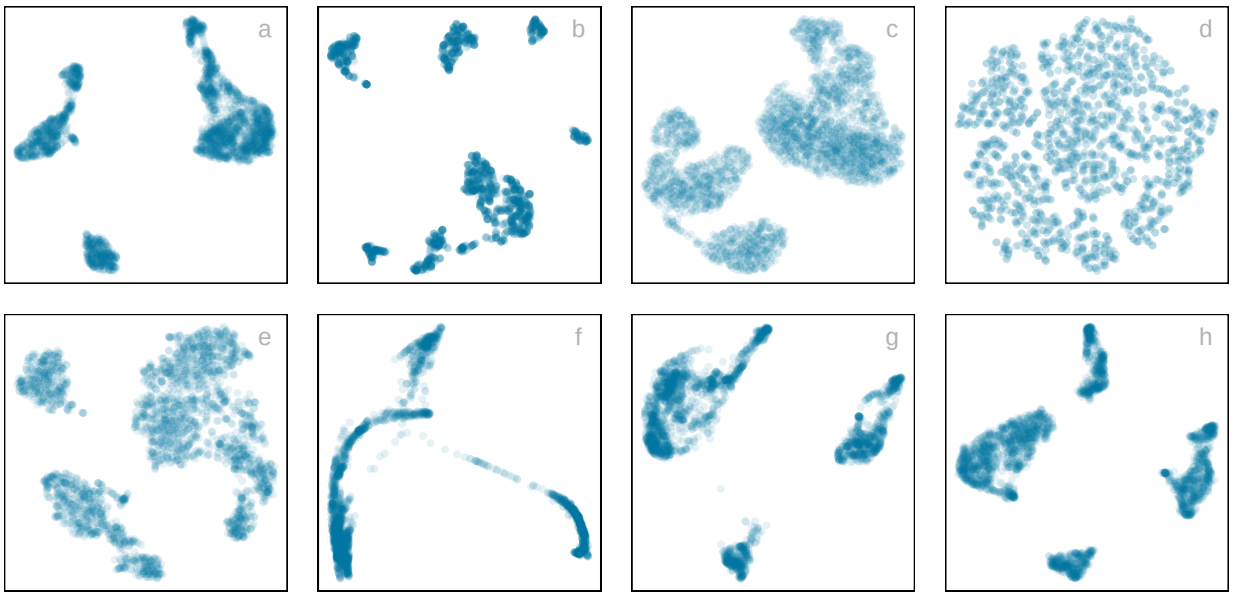

Figure 1: Six different NLDR representations of the same data. Different techniques and different parameter choices are used. Researchers may have seen any of these in their analysis of this data, depending on their choice of method, or typical parameter choice. Would they make different decisions downstream 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) (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 (Coifman et al. 2005) 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: (1) highly **separated clusters** (a, b, e, g, h) with the number ranging from 3-6; (2) **stringy branches** (f), and (3) **barely separated clusters** (c, d) which would **contradict** 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 Lee et al. (2021), broaden the scope by providing movies of linear projections, that provide views the data from all directions. Lee et al. (2021) provides an review of the main developments in tours. There are many tour algorithms implemented, with many available in the R package `tourr` (Wickham et al. 2011), and versions enabling better interactivity in `langevitour` (Harrison 2023) and `detourr` (Hart & Wang 2022). 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 Wickham et al. (2015) 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.

Wickham et al. (2015) 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 NLDR as a modeling technique might seem strange. It is 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. We can also imagine that the positioning of points in 2D represent the fitted values, that will have some prescribed position in p-D that can be compared with their observed values. This is the conceptual framework underlying the more formal versions of factor analysis (Jöreskog 1969) and multidimensional scaling (MDS) (Borg & Groenen 2005). (Note that, for this thinking the full p-D data needs to be available, not just the interpoint distances.)

We define the NLDR as a function g: $\mathbb{R}^{n\times p} \rightarrow \mathbb{R}^{n\times k}$, with (hyper-)parameters $\boldsymbol{\theta}$. The parameters, $\boldsymbol{\theta}$, depend on the choice of g, and can be considered part of model fitting in the traditional sense. Common choices for g include functions used in tSNE, UMAP, PHATE, TriMAP, PaCMAP, or MDS, although in theory any function that does this mapping is suitable.

With our goal being to make a representation of this 2D layout that can be lifted into high-dimensional space, the layout needs to be augmented to include neighbour information. A simple approach would be to triangulate the points and add edges. A more stable approach is to first hexagonally bin the data, reducing it from n to $m \leq n$ observations, and connect the bin centroids. This process serves to reduce some noisiness in the resulting surface shown in p-D. The steps in this process are shown in Figure 2, and documented below.

To illustrate the method, we use 7-D simulated data, which we call the "S-curve". It is constructed by setting $X_1 = \sin(a), X_2 = U(0,2), X_3 = \text{sign}(a) \times (\cos(a) - 1), \forall a \in [-3\pi/2, 3\pi/2]$. The remaining variables X_4, X_5, X_6, X_7 are all uniform error, with small variance. We would consider $T = (X_1, X_2, X_3)$ to be the true model.

Notation	Description
n, p, k, m	number of observations, variables, embedding dimension, number of non-empty bins, respectively
$\boldsymbol{X}, \boldsymbol{x}$	p-dimensional data (population, sample)
$\boldsymbol{y}$	k-dimensional layout
P	orthonormal basis, generating a *d-dimensional* linear projection of p-dimensional data
T	true model
g	functional mapping from p-D to k-D, especially as prescribed by NLDR
$\boldsymbol{\theta}$	(Hyper-) parameters for NLDR method
r	ranges of the embedding components
$C^{(j)}$	j-dimensional bin centers
(b_1, b_2)	number of bins in each direction
(a_1, a_2)	binwidths, distance between centroids in each direction
$(s_1,\ s_2)$	starting coordinates of the hexagonal grid
q	buffer to ensure hexgrid covers data, proportion of data range, 0-1
b, b'	total and non-empty hexagon bins in the grid
n_k	number of observations within the k^{th} hexagon

Table 1: Summary of notation for describing new methodology.

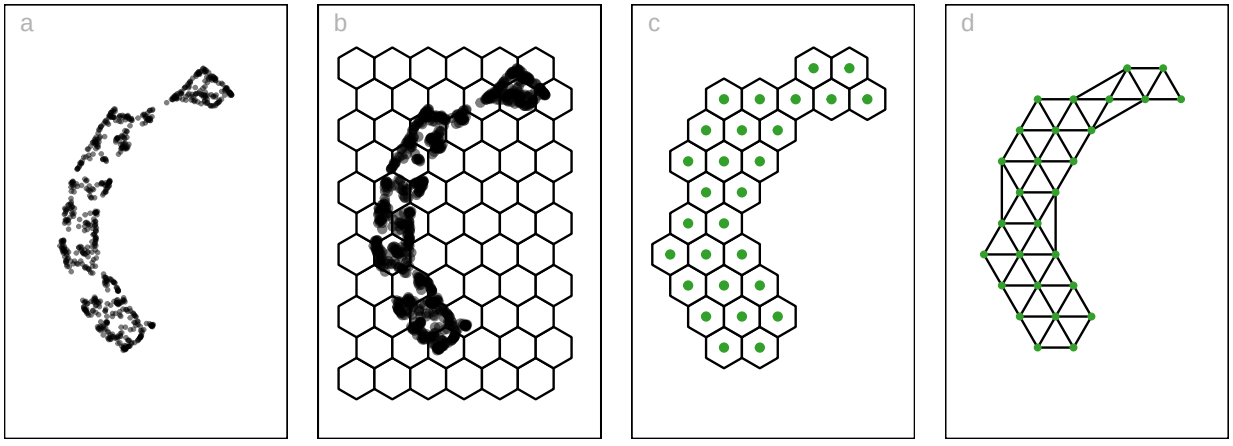

Figure 2: Key steps for constructing the model on the UMAP layout ($k = 2$) of the S-curve data: (a) data, (b) hexagon bins, (c) bin centroids, and (d) triangulated centroids.

3.1.1 Scaling the data

It is beneficial to define the algorithm on data having a standard scale. Here the variables are scaled to [0, 1], but the upper bound can incorporate the aspect ratio produced by the NLDR $(r_1, r_2, ..., r_k)$, by setting them to $(y_{1,\max}, y_{2,\max}, ..., y_{k,\max})$. When $k = 2$ which is assumed for hexagon binning, $y_{1,\max} = 1$ and $y_{2,\max} = \frac{r_2}{r_1}$, as observed in Figure 2.

3.1.2 Computing hexagon grid configurations

The 2D hexagon grid is defined by the number of bins in each direction (b_1, b_2), as given by the centroids of each hexagon (c_1, c_2) and the lower left position where the grid starts at (s_1, s_2), which correspond to the lowest left centroid. The values of s_i need to be below their respective minimum variable values, and could be a full bin lower, to allow a buffer (q) corresponding to a full hexagon width (a_1) and height (a_2) around the data. The values of b_i are variables to be computed that define the reduction in size of the data (n to m).

The value for b_2 is computed by fixing b_1. Considering the lower bound of the NLDR, $a_1 > 2 \times s_1$, and $a_1 > \frac{1-s_1}{b_1-1}$. Similarly, according to the upper bound of the NLDR, $a_1 > \frac{2(r_2-s_2)}{\sqrt{3}(b_2-1)}$, because $a_2 = \frac{\sqrt{(3)}}{2} a_1$ for regular hexagons. Therefore, $b_2 = \left\lceil 1 + \frac{2(r_2-s_2)(b_1-1)}{\sqrt{3}(1-s_1)} \right\rceil$.

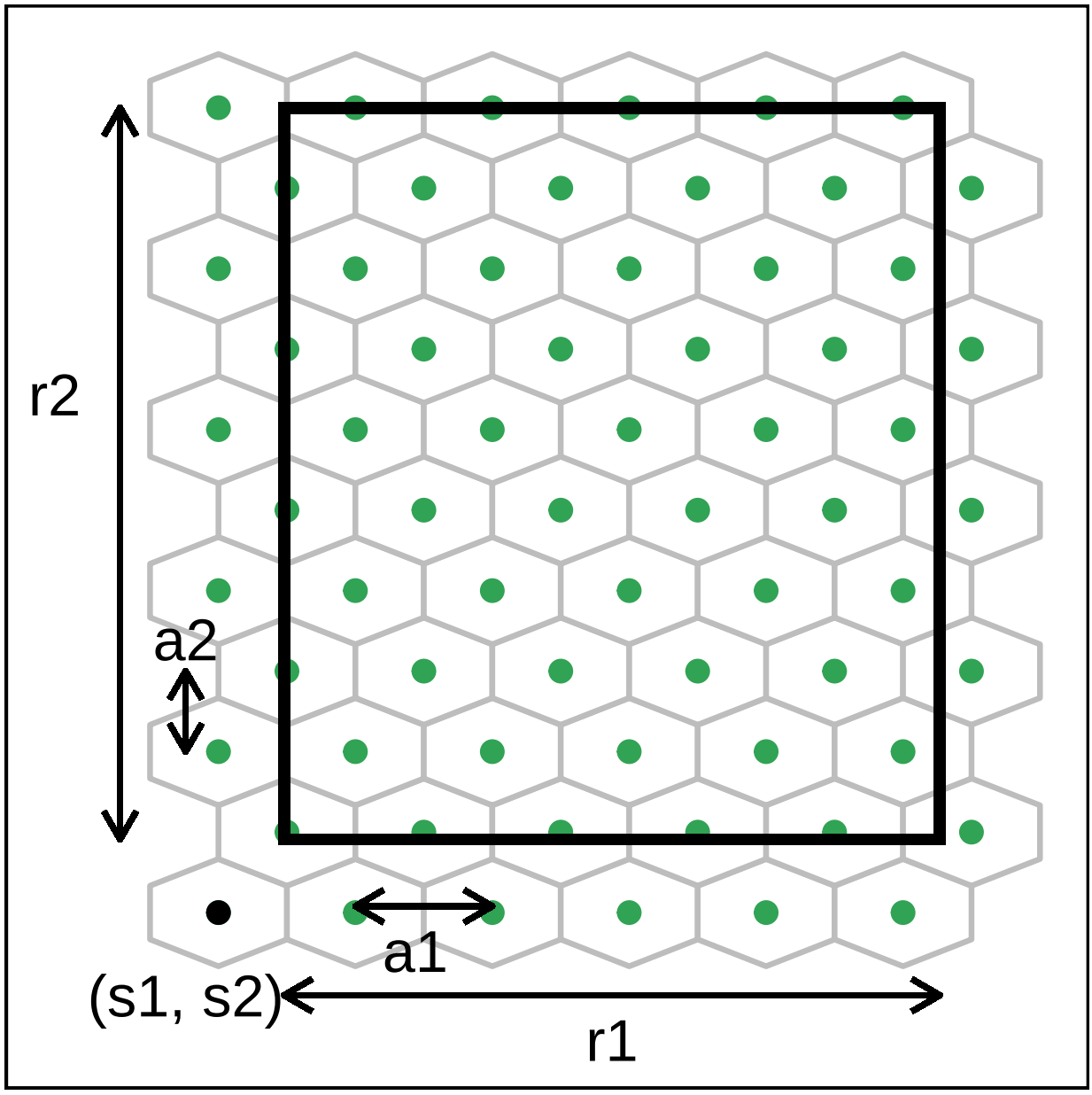

Figure 3: Hexagon binning parameters.

3.1.3 Binning the data

Points are allocated to the bin they fall into based on the nearest centroid. In situations where a point is equidistant from multiple centroids, tie-breaking rules are applied. If multiple centroids are in the same row, the point is assigned to the leftmost centroid. If multiple centroids are in different rows, the point is assigned to the bottom centroid.

3.1.4 Obtaining bin representations

Let $h : \mathbb{R}^{n \times k} \to \mathbb{R}^{m \times k}$ be the function that maps a point in k-D space to it's bin representation ($C^{(k)}$). This is the model points in k-D space.

When $k = 2$, one of the representations of a bin is the hexagonal centroid (Figure 2 (c)).

3.1.5 Generating edges

Delaunay triangulation on bin representations generates edges by specifying which bin representations should be connected to generate the model in k-D space.

When $k = 2$ Delaunay triangulation on $C^{(2)}$ generates the model in 2D space, which is a triangular mesh (Figure 2 (d)). It generates convex hulls of $C^{(2)}$ such that the circumcircle of every triangle in the triangulation contains no other points from $C^{(2)}$.

3.2 Displaying the model in p-D

The final step of the algorithm is *lifting* the k-D model back to p-D, completing the *model-in-the-data-space* picture. For this, define the set H^k to be the set of points that belong to centroid k. That is, $H^k = \{y \mid g(y) = (h_x^k, h_y^k)\}$. We use the p-D Euclidean mean of the points in H^k to map the centroid (h_x^k, h_y^k) to a point in p-D. Let the i-th component of the p-D mean be

$$f_i = \frac{1}{|H^k|} \sum_{y \in H^k} y_i,$$

with corresponding p-D vector $\vec{f}_i$. Then, $f(h \circ g)(x)$ maps a p-D point to the p-D model estimate, completing the model cycle.

Furthermore, edges that exist between k-D representations should also generate edges in p-D by connecting p-D mapping of the corresponding k-D representations.

3.3 Measuring the fit

3.4 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
  - 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 our method to get a better representation

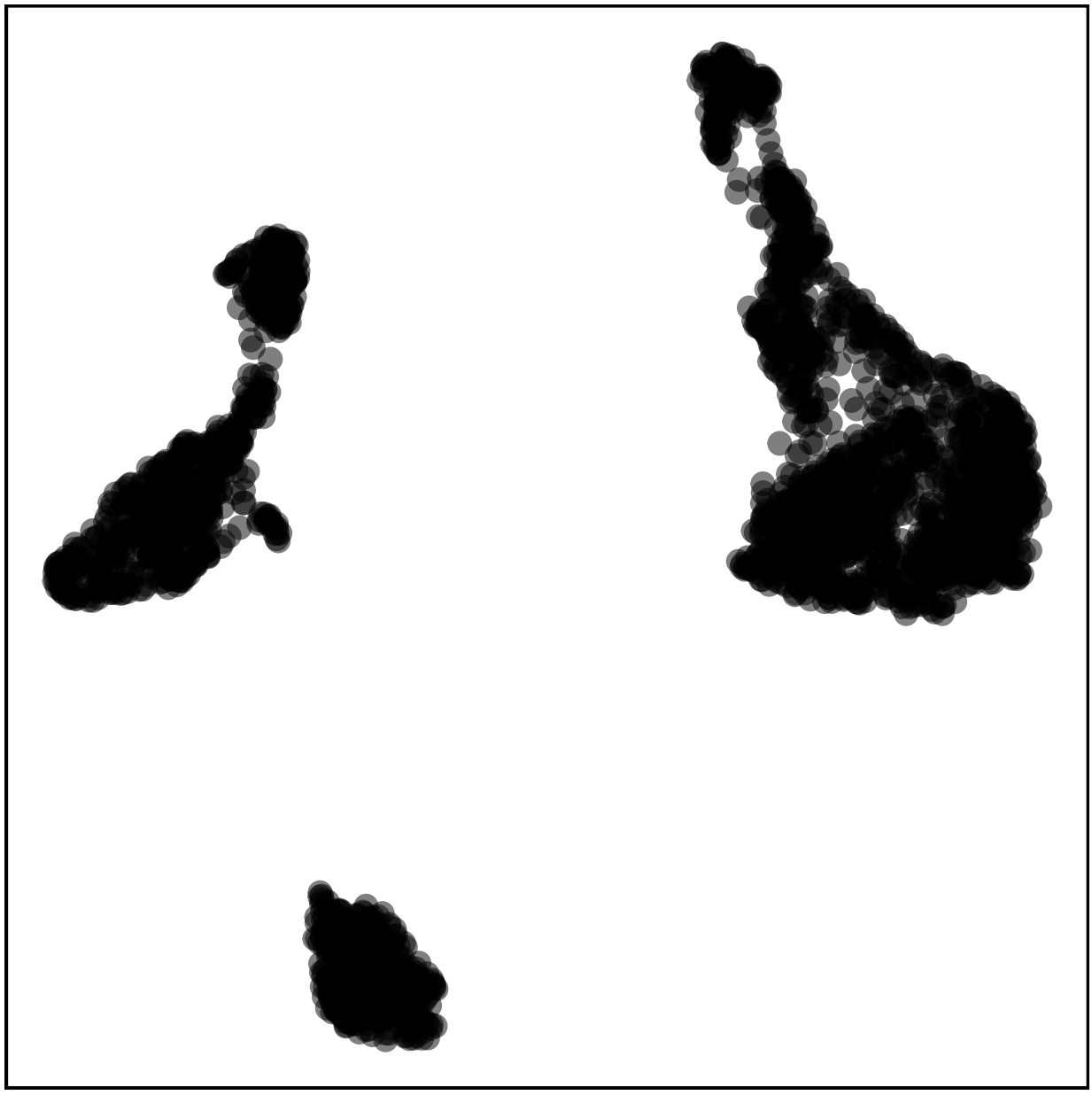

Figure 4: 2D layout from UMAP applied for the PBMC3k dataset. Is this a best representation of the original data?

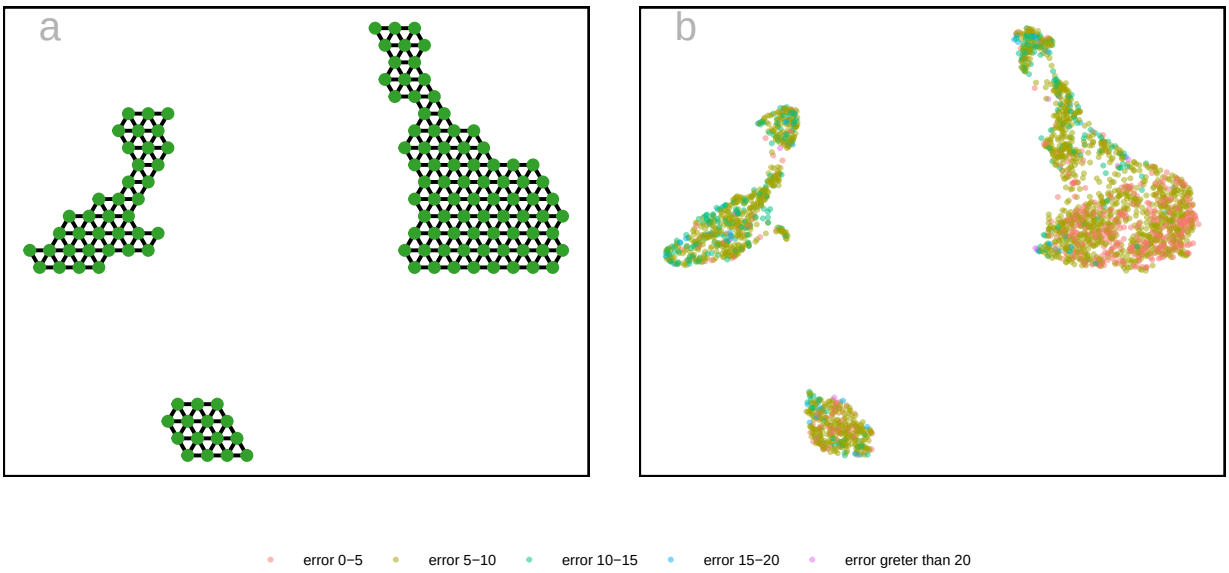

Figure 5: (a) Model generated in the 2D space overlaid on UMAP data, and (b) high-D model error in model space.

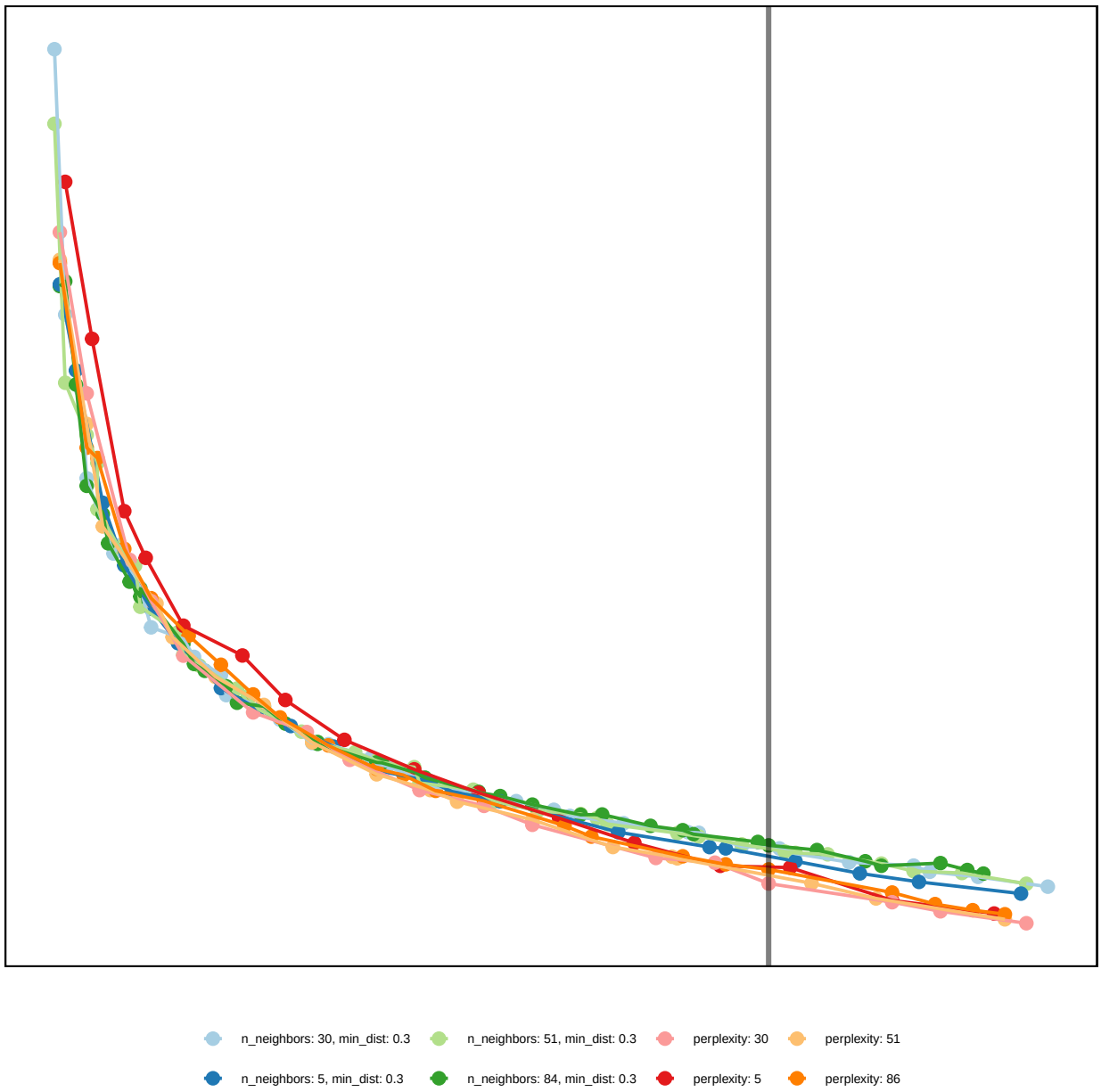

Figure 6: Absolute error from UMAP and tSNE applied to training PBMC3k dataset with diffferent parameter choices. What is the best parameter choice to create the model? The residual plot have a steep slope at the beginning, indicating that a smaller number of non-empty bins causes a larger amount of error. Then, the slope gradually declines or level off, indicating that a higher number of non-empty bins generates a smaller error. Using the elbow method, it was observed that when the number of non-empty bins is set to 137, the lowest error occurred with the parameters perplexity: 30.

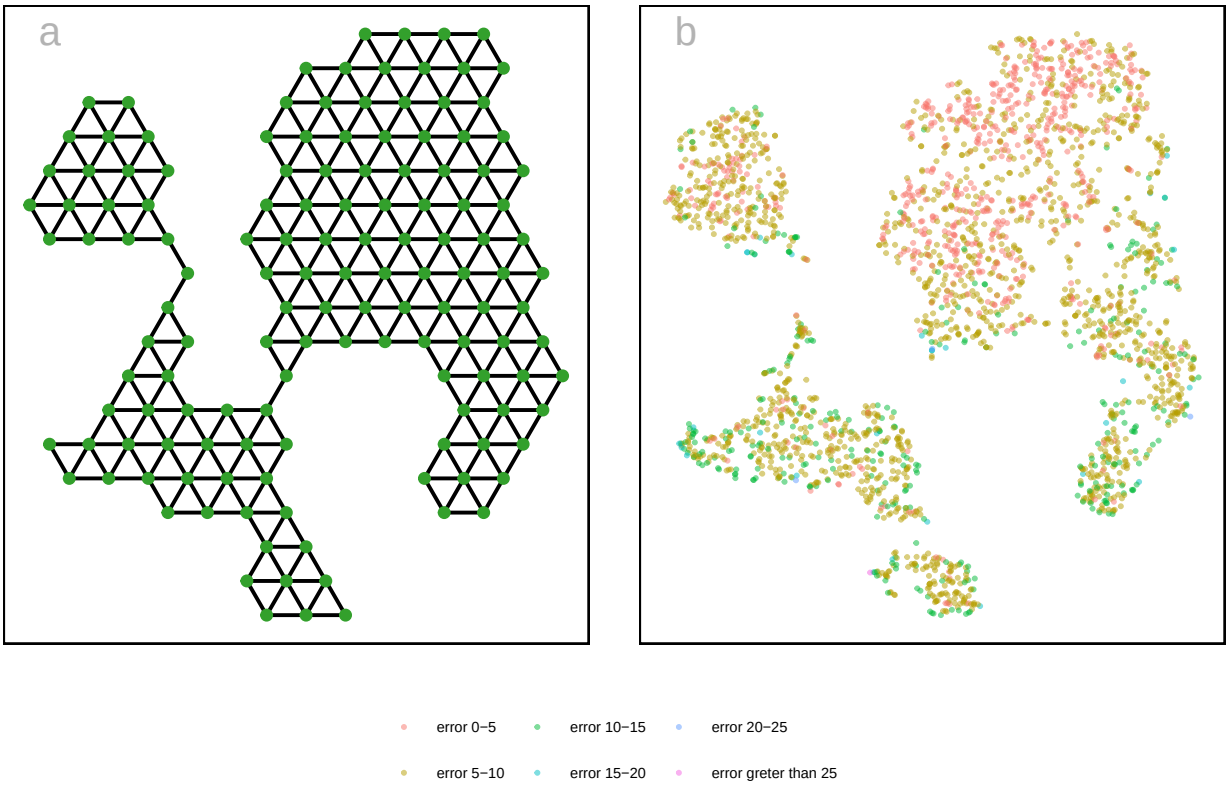

Figure 7: (a) Model generated in the 2D space overlaid on tSNE data, and (b) high-D model error in model space.

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

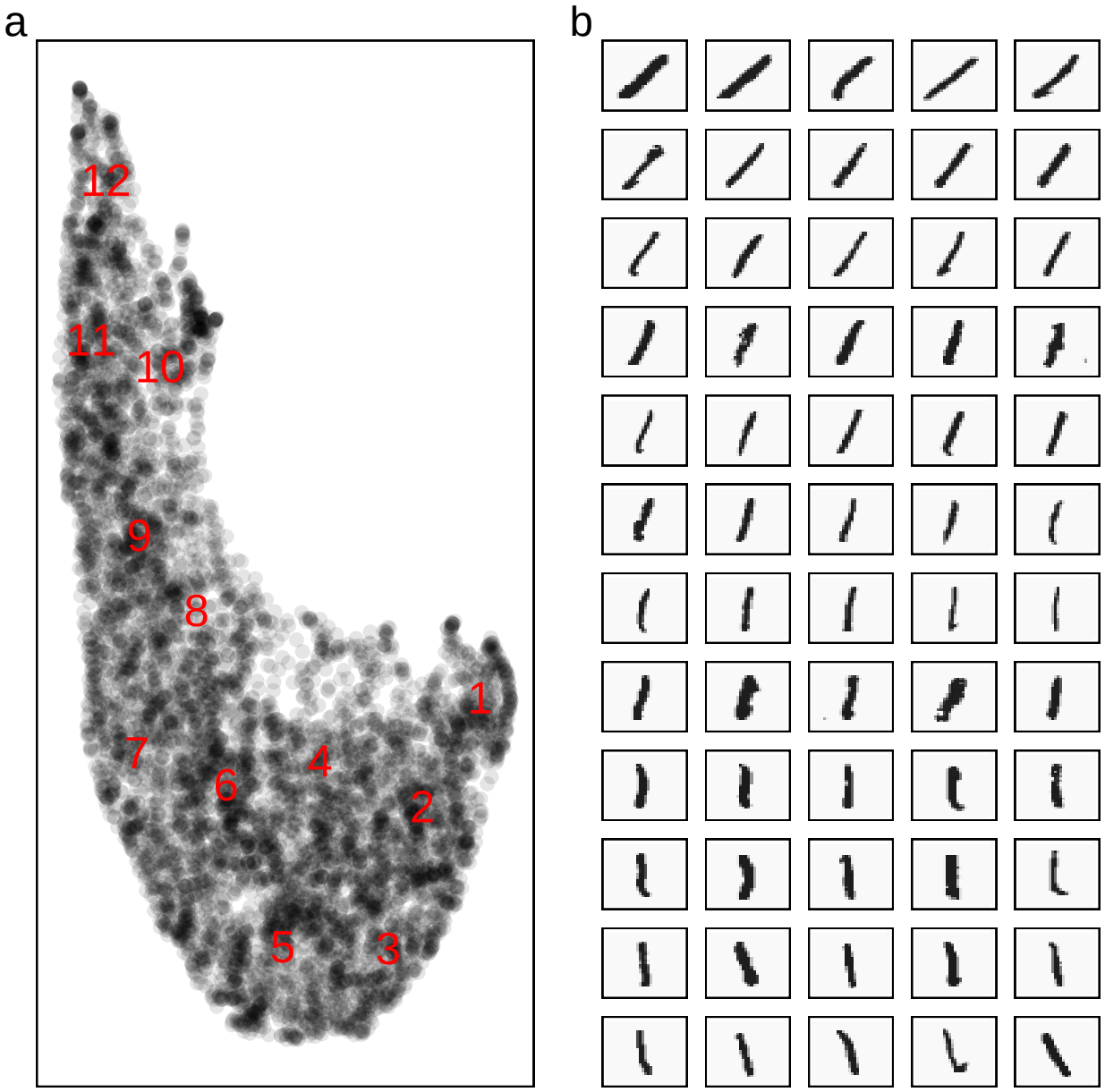

Figure 8: (a) 2D layout from PaCMAP applied for the digit 1 of the MNIST dataset. Is this a best representation of the original data?, and (b) Images of the handwritten digit 1. The angle of the digit 1 varies along non-linear structure.

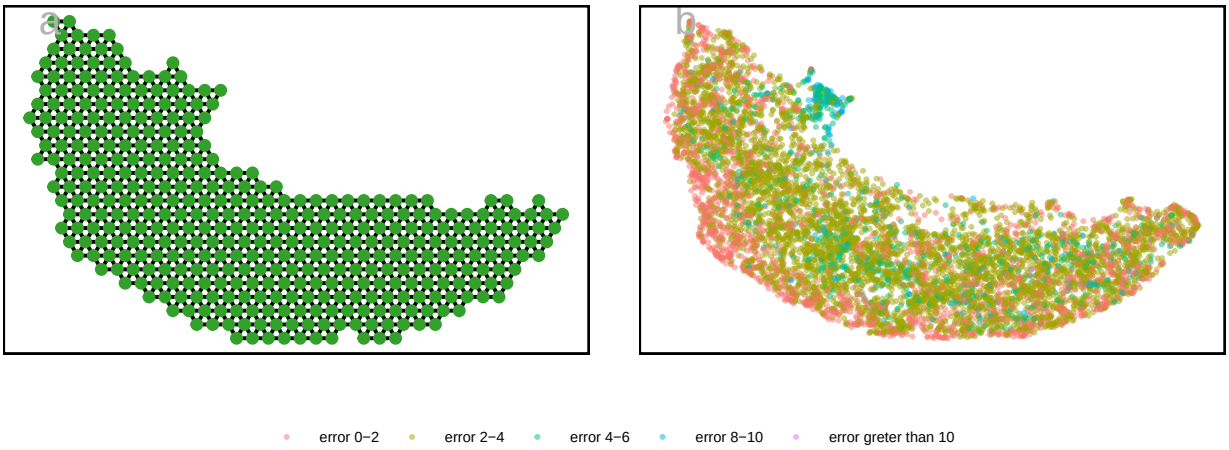

Figure 9: (a) Model generated in the 2D space overlaid on PaCMAP data, and (b) high-D model error in model space.

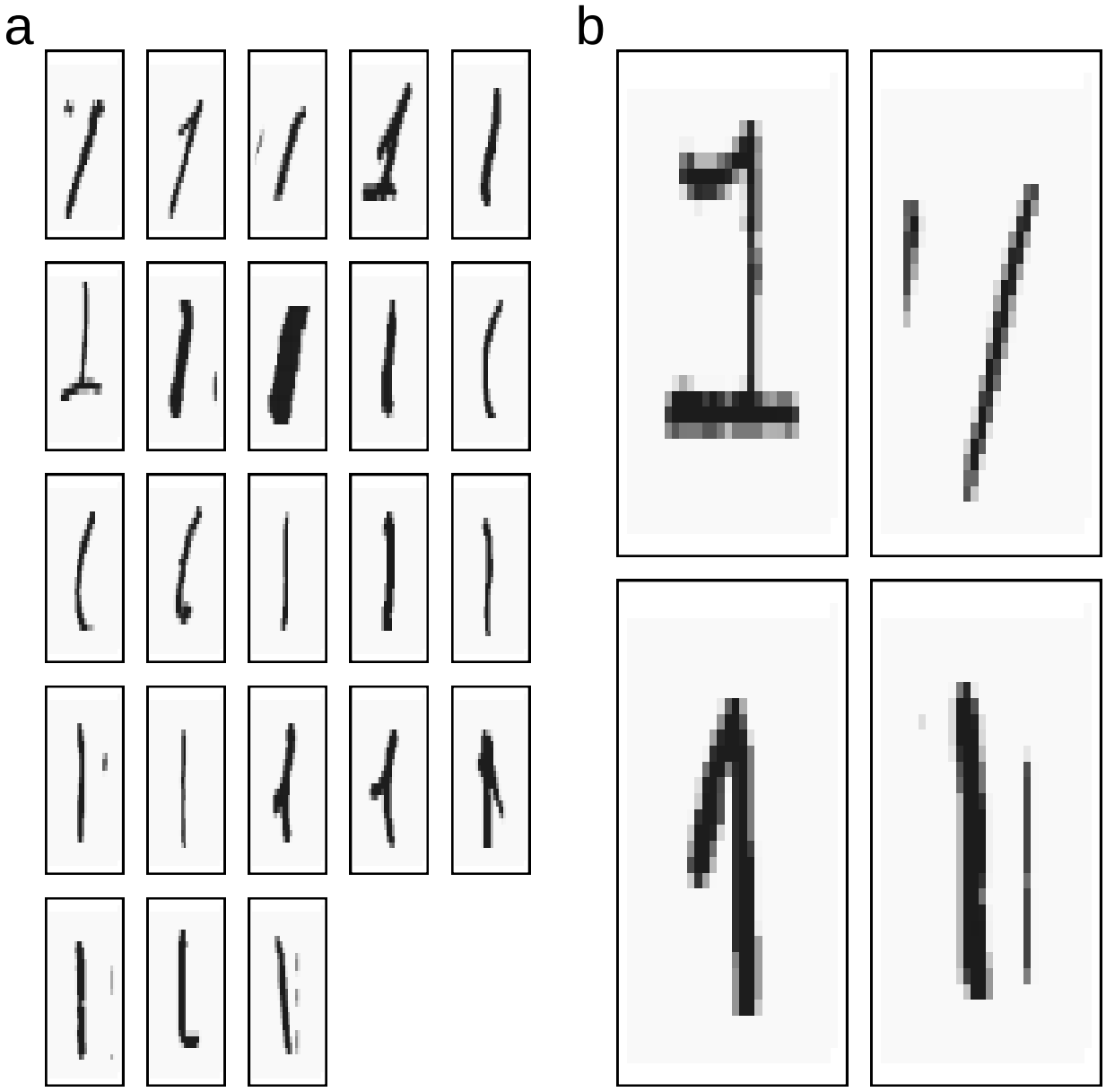

Figure 10: (a) Images of handwritten digit 1 which occur large model error within the non-linear strcuture, and (b) Images of handwritten digit 1 which occur large model error outside the non-linear structure.

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
- Predicting new observations in k-D

- Extending layouts beyond k-D, when 2D is clearly inadequate.
- 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'.

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

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.

Harrison, P. (2023), 'langevitour: Smooth interactive touring of high dimensions, demonstrated with scrna-seq data', *The R Journal* **15**, 206–219. https://doi.org/10.32614/RJ-2023-046.

Hart, C. & Wang, E. (2022), *detourr: Portable and Performant Tour Animations*. R package version 0.1.0.
URL: *https://casperhart.github.io/detourr/*

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*

Jöreskog, K. G. (1969), 'A general approach to confirmatory maximum likelihood factor analysis', *Psychometrika* pp. 183–202.
URL: *https://doi.org/10.1007/BF02289343*

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*

Lee, S., Cook, D., da Silva, N., Laa, U., Wang, E., Spyrison, N. & Zhang, H. S. (2021), 'A review of the state-of-the-art on tours for dynamic visualization of high-dimensional data'.

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.

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*

Wickham, H., Cook, D. & Hofmann, H. (2015), 'Visualizing statistical models: Removing the blindfold', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **8**(4), 203–225.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11271*

Wickham, H., Cook, D., Hofmann, H. & Buja, A. (2011), 'tourr: An r package for exploring multivariate data with projections', *Journal of Statistical Software* **40**(2), 1—18.
URL: *http://www.jstatsoft.org/v40/i02/*