Appendix: Stop Lying to Me: New Visual Tools to Choose the Most Honest Nonlinear Dimension Reduction

Jayani P. Gamage
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

June 30, 2025

Notation	Description
n, p, k	number of observations, variables, embedding dimension, 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
m	number of non-empty bins
b	number of hexagons in the grid
h	hexagonal id
l	side length
A	area
n_h	bin count
w_h	standardised bin counts

Table 1: Summary of notation for describing new methodology.

1 Generating the 2NC7 data

This data is constructed by simulating two clusters, each consisting of 1000 observations. The C-shaped cluster is generated from $\theta \sim U(\text{-}3\pi/2, 0)$, $X_1 = \sin(\theta)$, $X_2 \sim U(0, 2)$ (adding thickness to the C), $X_3 = \text{sign}(\theta) \times (\cos(\theta) - 1)$, $X_4 = \cos(\theta)$. Observations lie on a 2-D manifold in 7-D. The other cluster is from $X_1 \sim U(0, 2)$, $X_2 \sim U(0, 3)$, $X_3 = \text{-}(X_1^3 + X_2)$, and $X_4 \sim U(0, 2)$. It is also curved, but observations lie on a 3-D manifold in7-D. Three more variables, X_5, X_6, X_7, that are small amounts of pure noise are added. We would consider $T = (X_1, X_2, X_3, X_4)$ to be the geometric structure (true model) that we hope to capture.

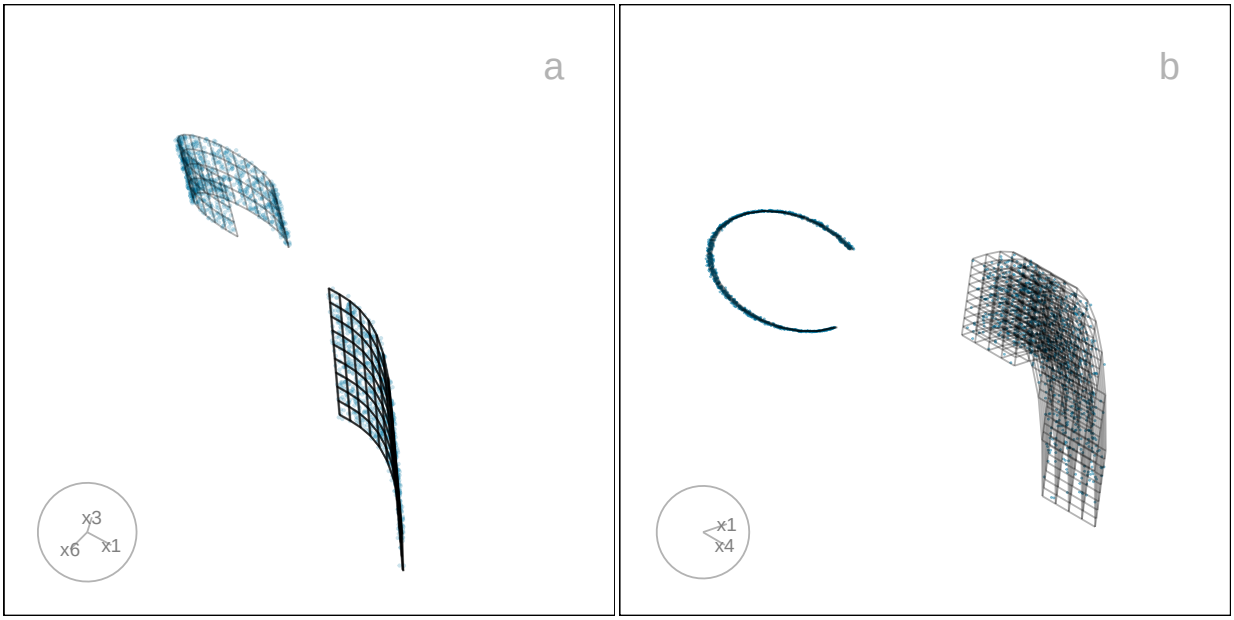

Figure 1: Two projections of the p-D true model overlaying the data are shown in a, b. Video of the langevitour animations is available at https://youtu.be/35TrnYJsUUI.

2 Computing hexagon grid configurations

Given range of embedding component, r_2, number of bins along the x-axis, b_1, and buffer proportion, q, hexagonal starting point coordinates, $s_1 = -q$, and $s_2 = -qr_2$. The purpose is to find width of the hexagon. a_1, and number of bins along the y-axis, b_2.

Geometric arguments give rise to the following constraints.

min a_1 s.t.

$$s_1 - \frac{a_1}{2} < 0, \tag{1}$$

$$s_1 + (b_1 - 1) \times a_1 > 1, \tag{2}$$

$$s_2 - \frac{a_2}{2} < 0, \tag{3}$$

$$s_2 + (b_2 - 1) \times a_2 > r_2. \tag{4}$$

Since a_1 and a_2 are distances,

$$a_1, a_2 > 0.$$

Also, $(s_1, s_2) \in$ (-0.1, -0.05) as these are multiplicative offsets in the negative direction.

Equation 1 can be rearranged as,

$$a_1 > 2s_1$$

which given $s_1 < 0$ and $a_1 > 0$ will *always* be true. The same logic follows for Equation 3 and substituting $a_2 = \sqrt{3}a_1/2$, and $s_2 =$ -qr_2 to Equation 3 can be written as,

$$a_1 > -\frac{4}{\sqrt{3}} qr_2$$

Also, substituting $a_2 = \sqrt{3}a_1/2$, $s_2 =$ -qr_2 and rearranging Equation 4 gives:

$$a_1 > \frac{2(r_2 + qr_2)}{\sqrt{3}(b_2 - 1)}. \tag{5}$$

Similarly, substituting $s_1 =$ -q Equation 2 becomes,

$$a_1 > \frac{(1+q)}{(b_1-1)}. \tag{6}$$

This is a linear optimization problem. Therefore, the optimal solution must occur on a vertex. So, by setting Equation 5 equals to Equation 6 gives,

$$\frac{2(r_2+qr_2)}{\sqrt{3}(b_2-1)} = \frac{(1+q)}{(b_1-1)}.$$

After rearranging this,

$$b_2 = 1 + \frac{2r_2(b_1-1)}{\sqrt{3}}$$

and since b_2 should be an integer,

$$b_2 = \left\lceil 1 + \frac{2r_2(b_1-1)}{\sqrt{3}} \right\rceil. \tag{7}$$

Furthermore, with known b_1 and b_2, by considering Equation 2 or Equation 4 as the *binding* or *active constraint*, can compute a_1.

If Equation 2 is active, then,

$$\frac{(1+q)}{(b_1-1)} < \frac{2(r_2+qr_2)}{\sqrt{3}(b_2-1)}.$$

Rearranging this gives,

$$r_2 > \frac{\sqrt{3}(b_2-1)}{2(b_1-1)}.$$

Therefore, if this equality is true, then

$$a_1 = \frac{(1+q)}{(b_1 - 1)},$$

otherwise,

$$a_1 = \frac{2r_2(1+q)}{\sqrt{3}(b_2 - 1)}.$$

3 Binning the data

Points are assigned to the bin they fall into based on the nearest centroid. If a point is equidistant from multiple centroids, it is assigned to the centroid with the smallest bin ID.

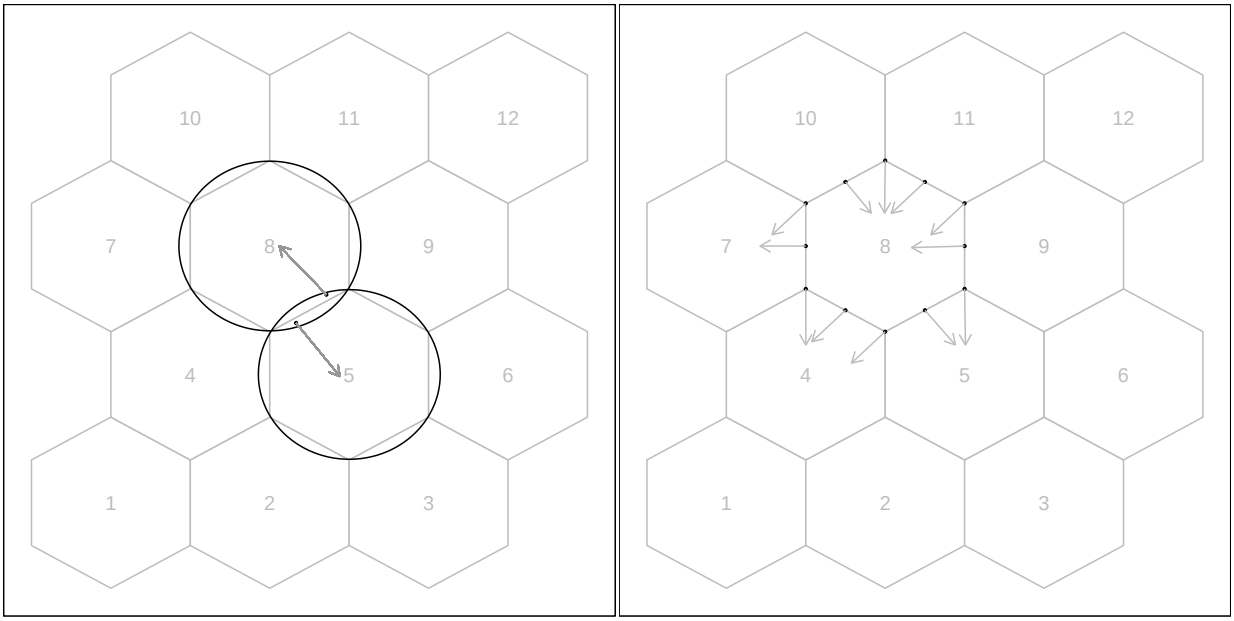

Figure 2: Binning the data. Points are assigned to the nearest centroid. If a point is equidistant from multiple centroids, assigned to the centroid with the smallest bin ID.

4 Area of a hexagon

The area of a hexagon is defined as $A = 3\sqrt{3}l^2/2$, where l is the side length of the hexagon.

l can be computed using a_1 and a_2.

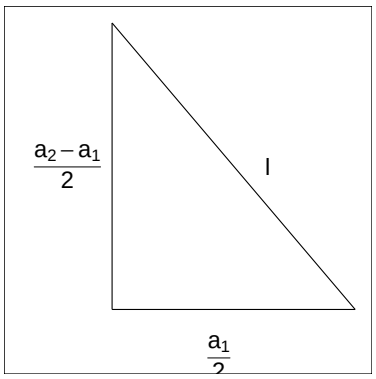

Figure 3: The components of the right triangle illustrating notation.

By applying the Pythagorean theorem, we obtain,

$$l^2 = \left(\frac{a_1}{2}\right)^2 + \left(\frac{a_2 - l}{2}\right)^2.$$

Next, rearranging the terms, we get,

$$l^2 - \left(\frac{a_2 - l}{2}\right)^2 = \left(\frac{a_1}{2}\right)^2,$$

$$\left[l - \left(\frac{a_2 - l}{2}\right)\right]\left[l + \left(\frac{a_2 - l}{2}\right)\right] = \left(\frac{a_1}{2}\right)^2,$$

$$3l^2 + 2a_2 l - (a_1^2 + a_2^2) = 0.$$

Finally, by solving the quadratic equation, we compute,

$$l = \frac{-2a_2 \pm \sqrt{4a_2^2 - 24[-(a_1^2 + a_2^2)]}}{6},$$

$$l = \frac{-a_2 \pm \sqrt{a_2^2 - 6[-(a_1^2 + a_2^2)]}}{3},$$

where $l > 0$.

5 Single-cell gene expression: comparison with results of scDEED recommendations

As we were writing this paper Xia et al. (2023) appeared proposing a new method called scDEED helping to assess the validity of a 2-D embedding. scDEED calculates a reliability score for each cell embedding based on the similarity between the cell's 2-D embedding neighbors and its neighbors prior to embedding. A low reliability score suggests a dubious embedding. It can help in the deciding on optimal hyper-parameters. Here we illustrate how our method compares with the results from scDEED.

Note that Xia et al. (2023) uses a different PBMC dataset than that used by Chen et al. (2024), shown by us in the main paper example, which is why this comparison is shown here and not in the main paper. Their data contains $31,021$ cells including cell type labels, and the gene expression levels were in the unit of log-transformed UMI count per $10,000$. They focused on three sequencing methods (inDrops, DropSeq, and SeqWell) and four common cell types Cytotoxic T cell, CD4+T cell, CD14+ Monocyte, and B cell. Pre-processing follows the process in Xia et al. (2023) again using the Human Peripheral Blood Mononuclear Cells (PBMC) data.

For illustration purposes, we only selected cells generated with inDrops ($n = 5858$ cells). Also, Xia et al. (2023) used first 9 principal components to generate the UMAP and tSNE with default hyper-parameters. The objective is to assess the optimized layout by scDEED, and if it does not accurately represent the three clusters with small separations of the PBMC dataset, then select a reasonable 2-D layout.

The layout a (Figure 4) is generated from the hyper-parameters suggested by Chen et al. (2024), and the layout b (Figure 4) is with suggested hyper-parameters by scDEED to be more accurate. The layout a and b contain 46 and 83 dubious cells respectively.

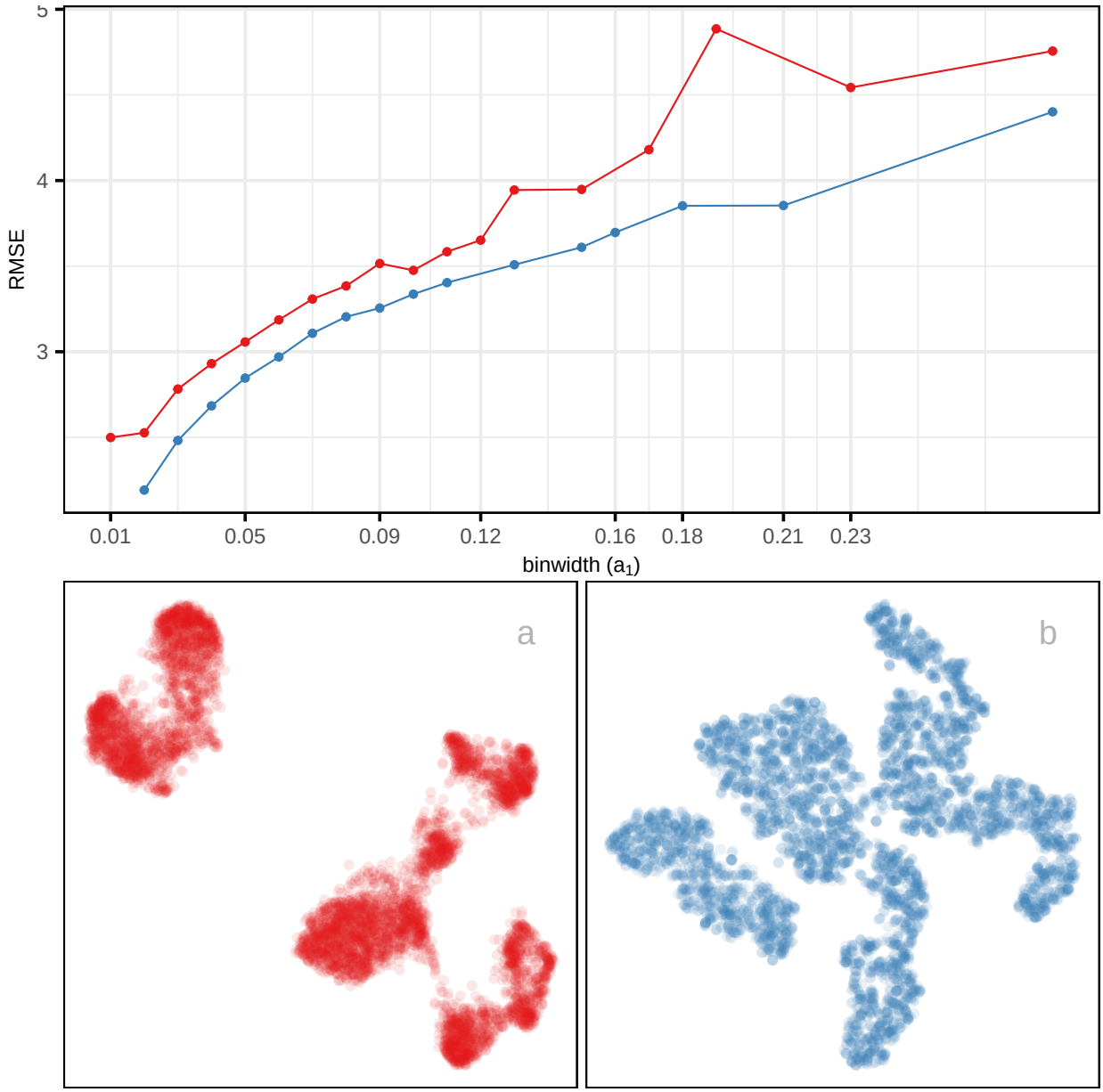

Figure 4: Assessing which of the 2 NLDR layouts on the PBMC3k data is the better representation using RMSE for varying binwidth (a_1). Color used for the lines and points in the left plot and in the scatterplots represents NLDR layouts (a, b). Of the two, layout b is optimal across all binwidths making it the best choice.

The RMSE vs binwidth (a_1) plot (Figure 4) illustrates that our approach would suggest that scDEED is correct here, that layout b is more accurately reflecting the cluster structure in the PBMC data.

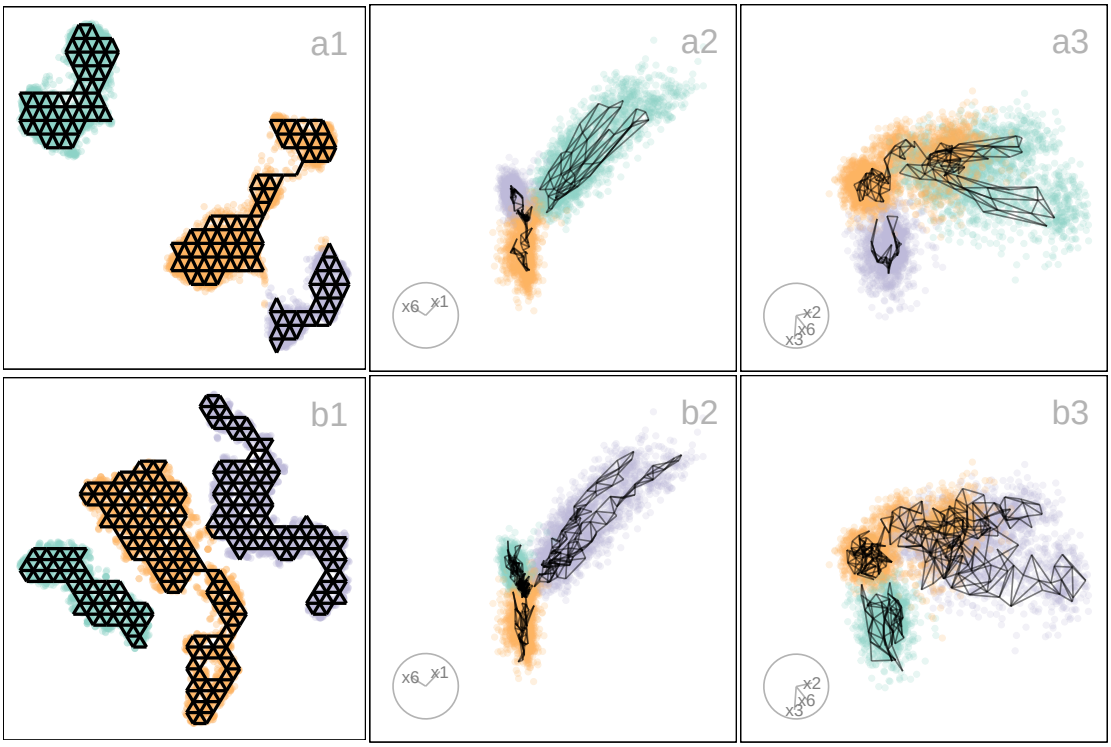

Figure 5: Compare the published 2-D layout (Figure 4 a) made with UMAP and the 2-D layout selected (Figure 4 b) made with tSNE by RMSE plot (Figure 4). The two plots on the right show projections from a tour, with the models overlaid. The published layout a suggested three separated clusters with two of them are close, but this is not present in the data. While there may be three clusters they are not well-separated. The difference in model fit also indicates this: the published layout a does not capture the nonlinear structure of the clusters like the model generated from layout b. This supports the choice that layout b is the better representation of the data, because it shows close clusters. Videos of the langevitour animations are available at https://youtu.be/ffiB4MGWyn8 and https://youtu.be/e7XNL18co1c respectively.

References

Chen, Z., Wang, C., Huang, S., Shi, Y. & Xi, R. (2024), 'Directly Selecting Cell-type Marker Genes for Single-cell Clustering Analyses', *Cell Reports Methods* **4**(7), 100810. https://www.sciencedirect.com/science/article/pii/S2667237524001735.

Xia, L., Lee, C. & Li, J. J. (2023), 'scDEED: A Statistical Method for Detecting Dubious

2D Single-cell Embeddings', *bioRxiv* . https://www.biorxiv.org/content/early/2023/04/25/2023.04.21.537839.