

# Appendix: 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

January 17, 2025

Notation	Description
$n, p, k$	number of observations, variables, embedding dimension, respectively
$\mathbf{X}, \mathbf{x}$	$p$ -dimensional data (population, sample)
$\mathbf{y}$	$k$ -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
$\theta$	(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

Table 1: Summary of notation for describing new methodology.

## 1 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  $s_2 = -q \times r_2$ . The purpose is to find width of the hexagon.  $a_1$ , and number of bins along the y-axis,  $b_2$ .

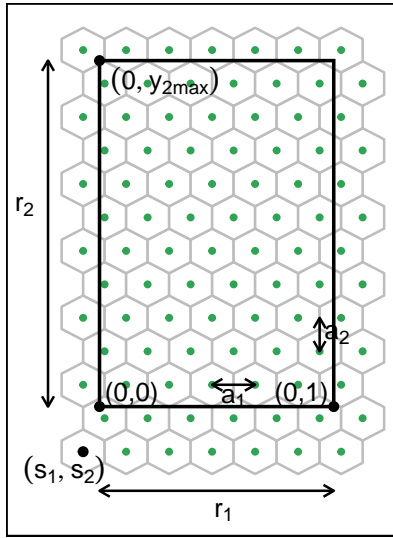


Figure 1: The components of the hexagon grid illustrating notation.

Geometric arguments give rise to the following constraints.

$\min a_1$  s.t.

$$s_1 - \frac{a_1}{2} < 0, \quad (1)$$

$$s_1 + (b_1 - 1) \times a_1 > 1, \quad (2)$$

$$s_2 - \frac{a_2}{2} < 0, \quad (3)$$

$$s_2 + (b_2 - 1) \times a_2 > r_2. \quad (4)$$

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  $a_2 = \frac{\sqrt{3}}{2}a_1$ , and  $s_2 = -q \times r_2$  to Equation 3 can be written as,

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

Also, substituting  $a_2 = \frac{\sqrt{3}}{2}a_1$ ,  $s_2 = -q \times r_2$  and rearranging Equation 4 gives:

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

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

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

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,

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

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

## 2 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 lowest hexagonal bin ID.

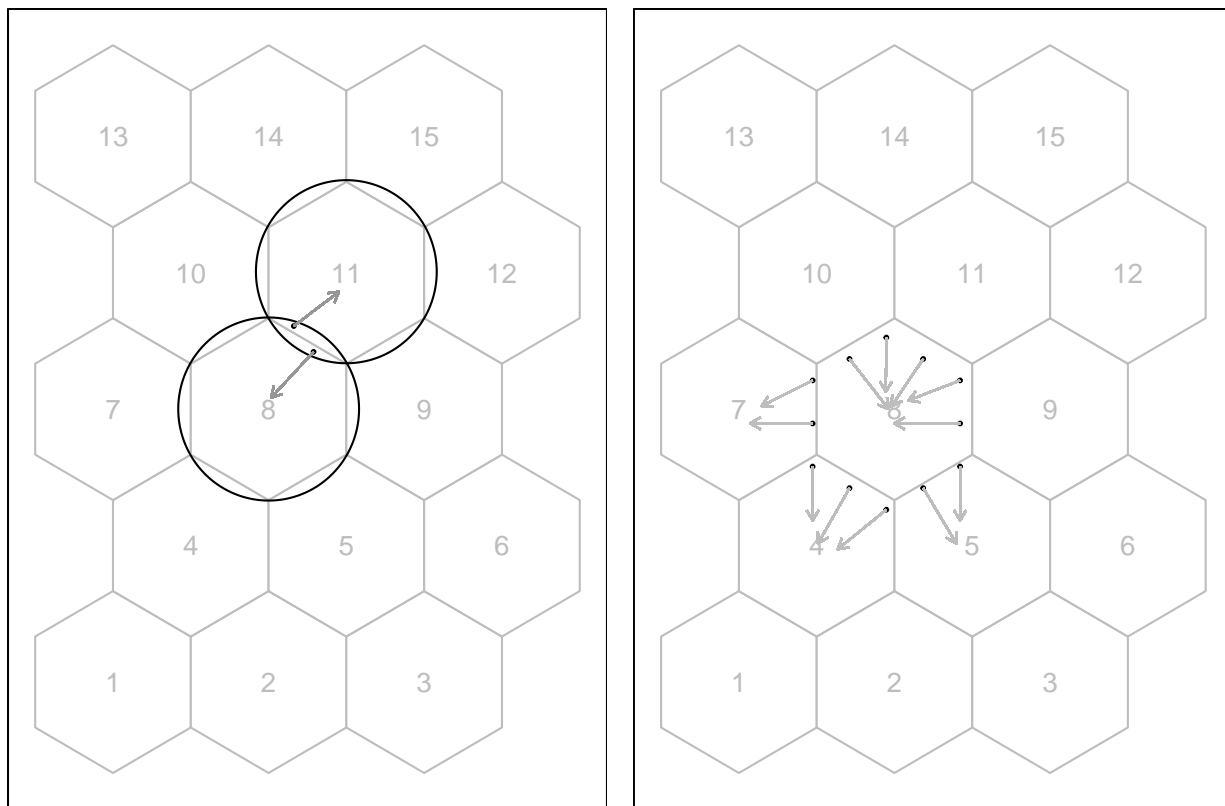


Figure 2: Binning the data. Points are assigned to the nearest centroid. If a point is equidistant from multiple centroids, assigned to the lowest centroid.

### 3 Area of a hexagon

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

can be computed using  $a_1$  and  $a_2$ .

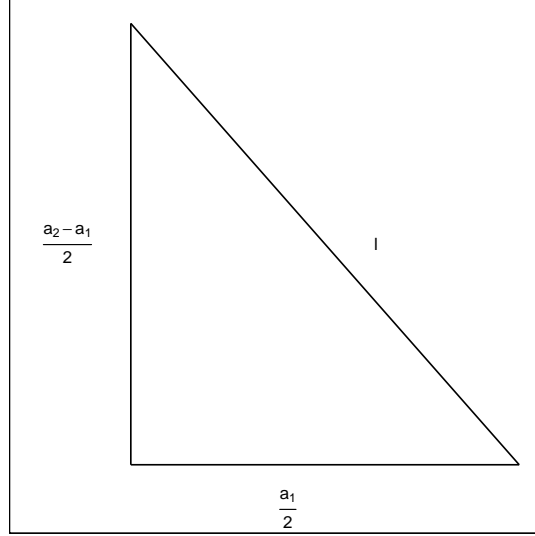


Figure 3: The components of the right triangle illustrating notation.

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,

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

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

$$3l^2 + 2a_2l - (a_1^2 + a_2^2) = 0.$$

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  $l > 0$ .

$\rightarrow$

$\rightarrow$