Supplementary Files

SIYI WANG

2023/07/20

1 Proofs

1.1 Reflection Spatial Depth

Proposition 1 (Symmetry)**.** *Consider two distinct points* $\mathbf{u}$ *and* $\mathbf{v}$*, and a point* $\mathbf{x}_i \in \mathbf{X}$*. Also, let* $\mathbf{u}^i$ *and* $\mathbf{v}^i$ *denote the points obtained by reflecting* $\mathbf{u}$ *and* $\mathbf{v}$*, respectively, about* $\mathbf{x}_i$*, that is* $\mathbf{u}^i = 2\mathbf{x}_i - \mathbf{u}$ *and* $\mathbf{v}^i = 2\mathbf{x}_i - \mathbf{v}$*. Then, we have that*

$$\mathbf{u} - \mathbf{v}^i = \mathbf{v} - \mathbf{u}^i,$$

and that

$$\|\mathbf{v}^i - \mathbf{u}\|^2 = 2\|\mathbf{v} - \mathbf{x}_i\|^2 + 2\|\mathbf{u} - \mathbf{x}_i\|^2 - \|\mathbf{u} - \mathbf{v}\|^2.$$

Proof. The reflection points $\mathbf{v}^i$ and $\mathbf{u}^i$ can be expressed as $2\mathbf{x}_i - \mathbf{v}$ and $2\mathbf{x}_i - \mathbf{u}$ respectively. Then we have

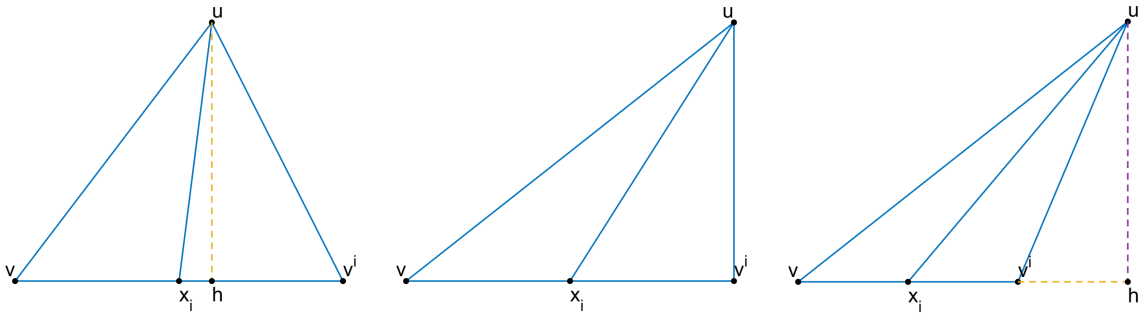
$$\begin{aligned} \mathbf{u} - \mathbf{v}^i &= \mathbf{u} - 2\mathbf{x}_i + \mathbf{v} \\ &= \mathbf{v} - (2\mathbf{x}_i - \mathbf{u}) \\ &= \mathbf{v} - \mathbf{u}^i. \end{aligned}$$

For expression simplicity, let a, b, c, and d represent $\|\mathbf{u} - \mathbf{v}\|$,$\|\mathbf{u} - \mathbf{x}_i\|$, $\|\mathbf{v} - \mathbf{x}_i\|$ and $\|\mathbf{u} - \boldsymbol{v}^{\boldsymbol{i}}\|$, i.e., the lengths of vectors $\mathbf{vu}$, $\mathbf{x}_i\mathbf{u}$, $\mathbf{vx}_i$, and $\mathbf{v^iu}$, respectively.

Firstly, consider the scenario where $\mathbf{u} = \mathbf{v}$. In this case, $a = 0$ and $b^2 = c^2$. Evidently, $d^2 = (2c)^2 = 2c^2 + 2b^2$. Subsequently, assuming that the points $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{x}_i$ are not col-linear, they form a triangle. If $\triangle vuv^i$ is an isosceles triangle, with $a = d$, it is apparent that $b^2 + c^2 = a^2 = d^2$, and thus $d^2 = 2b^2 + 2c^2 - a^2$ holds. If $\triangle vuv^i$ is not an isosceles triangle, we consider three cases, each illustrated by a representative example in Figure 1:

- If $\angle vv^iu$ is an acute angle, we introduce an auxiliary line uh orthogonal to vv^i (see Figure 1a). Let z and y denote the lengths of x_ih and uh, respectively. Then, the following relationships hold:

$$\begin{aligned} z^2 + y^2 &= b^2, \\ (c + z)^2 + y^2 &= a^2, \\ (c - z)^2 + y^2 &= d^2. \end{aligned}$$



(a) Acute triangle (b) Right triangle (c) Obtuse triangle

Figure 1: Examples of three different types of $\triangle vuv^i$

As a consequence,

$$\begin{aligned} a^2 + d^2 &= 2c^2 + 2z^2 + 2y^2 = 2c^2 + 2b^2, \\ d^2 &= 2c^2 + 2b^2 - a^2. \end{aligned}$$

- If $\angle vv^iu$ is a right angle (Figure 1b), the following relationships are true:

$$\begin{aligned} c^2 + d^2 &= b^2, \\ 4c^2 + d^2 &= a^2. \end{aligned}$$

From which we derive:

$$\begin{aligned} 2b^2 - a^2 &= 2c^2 + 2d^2 - 4c^2 - d^2 = d^2 - 2c^2, \\ d^2 &= 2c^2 + 2b^2 - a^2. \end{aligned}$$

- If $\angle vv^iu$ is an obtuse angle, we proceed similarly to the acute angle case, introducing auxiliary lines (as depicted in Figure 1c). Let z and y denote the lengths of v^ih and uh, respectively. Then, we have:

$$\begin{aligned} z^2 + y^2 &= d^2, \\ (z + c)^2 + y^2 &= b^2, \\ (z + 2c)^2 + y^2 &= a^2. \end{aligned}$$

From which we derive:

$$\begin{aligned} 2b^2 - a^2 &= y^2 + z^2 - 2c^2 = d^2 - 2c^2, \\ d^2 &= 2c^2 + 2b^2 - a^2. \end{aligned}$$

Hence, in all cases, the relationship $d^2 = 2c^2 + 2b^2 - a^2$ holds, proving the lemma. □

1.2 Density-based metric

Notations in this section are the same as the paper, for convenience they are listed as below.

- η_X: The largest value of η such that all observations in the dataset are depth-connected.
- ζ: A certain cluster in the clustering results.
- η_ζ: The largest value of η such that ζ is decomposed into a single DBCA-cluster and outliers (without considering observations in other clusters).
- n_ζ: The number of points in ζ.
- $\mathbf{O}_\zeta$: The set of DBCA-outliers for the cluster ζ.
- $\mathcal{X}_\zeta$ The set of observations in cluster ζ.
- $\mathcal{J}_\zeta^i = \{j | \mathbf{x}_j \in \mathcal{X}_\zeta, \mathbf{x}_j \neq \mathbf{x}_i, \mathbb{S}_{ij} > \eta_\zeta\}$: The set of indices for observation $\mathbf{x}_i \in \mathcal{X}_\zeta$.
- $\mathbf{I}_\zeta := \{\mathbf{x}_i \in \mathcal{X}_\zeta : \mathbb{S}_{ij} < \eta_X \text{ for all } \mathbf{x}_j \in \mathbf{X} \setminus \mathcal{X}_\zeta\}$: The set of inner observations in cluster ζ.
- $\mathcal{H}_\zeta^i = \{h | \mathbf{x}_h \notin \mathcal{X}_\zeta, \mathbb{S}_{ih} > \eta_X\}$: The set of indices of observations from other clusters that are core neighbors of points in $\mathcal{X}_\zeta$ when $\eta = \eta_X$.

To estimate η, we use a method analogous to the grouping matrix approach in the DLCC algorithm, specifically examining η_X as an example. Initially, we test an η value to assess if it can interconnect all observations within the dataset. A logical starting point is:

$$\min_{\mathbf{x}_i} \max_{\mathbf{x}_j} \{\mathbb{S}_{ij} : i \neq j\}.$$

This represents the similarity value necessary to ensure that each observation is at least depth-connected to another observation. Any η value above it is impossible to connect all observations.

If this initial η value successfully connects all observations based on the DBCA, then η_X is set to this value. If not, reflecting the principles of density-based clustering, we identify the highest similarity value between any two points across clusters under the current η setting. We then construct a matrix storing these values.

Following the methodology outlined on page 9 of the papers, this matrix undergoes iterative merging of columns and rows containing the largest significant (neither 0 nor 1) values. This process continues until the matrix is reduced to a single element. The η_X value is then the last significant value recorded in this reduction.

Properties of the DC metric:

Property 1: For each observation $\mathbf{x}_i \in \mathcal{X}_\zeta$, reducing the threshold from η_ζ to $\eta_{\zeta'}$ implies $\mu_{i'} \leq \mu_i$.

Property 2: For each observation $\mathbf{x}_i \in \mathcal{X}_\zeta$, reducing the threshold from η_ζ to $\eta_{\zeta'}$ implies $\tilde{\mu}_{i'} \leq \tilde{\mu}_i$.

Property 3: For any cluster ζ, the μ_i given by any outlier is less than η_ζ.

Property 4: For any cluster ζ, $H_\zeta \geq \eta_X$. Higher proportion of inner observations in ζ leads to smaller H_ζ.

Property 1. *For each observation $\mathbf{x}_i \in \mathcal{X}_\zeta$, reducing the threshold from η_ζ to $\eta_{\zeta'}$ implies $\mu_{i'} \leq \mu_i$. Additionally, if outliers are not considered, the weighted within-cluster similarity for observations in $\mathcal{X}_\zeta$ will either remain the same or decrease for smaller η values.*

Proof. Assume $\eta_{\zeta'} < \eta_\zeta$. For any observation $\mathbf{x}_i \in \mathcal{X}_\zeta$ that is not part of $\mathbf{o}_\zeta$, we have

$$\mu_i = \frac{Q_i}{n_{Q_i}}, \tag{1}$$

$$\mu_{i'} = \frac{Q_i + W_i}{n_{Q_i} + n_{W_i}}, \tag{2}$$

where $Q_i = \sum_{j \in \mathcal{J}_\zeta^i} \mathbb{S}_{ij}$, $n_{Q_i} = |\mathcal{J}_\zeta^i|$, $W_i = \sum\{\mathbb{S}_{ij} \mid \eta_{\zeta'} \leq \mathbb{S}_{ij} < \eta_\zeta\}$ and $n_{W_i} = |\{\mathbb{S}_{ij} \mid \eta_{\zeta'} \leq \mathbb{S}_{ij} < \eta_\zeta\}|$. If $n_{W_i} = 0$, then $\mu_i = \mu_{i'}$. For the case where $n_{W_i} > 0$ and W_i, Q_i, n_{W_i}, and

n_{Q_i} are all positive, we get

$$\begin{aligned} \eta_{\zeta'} \leq W_i/n_{W_i} < \ \eta_\zeta \ &\leq Q_i/n_{Q_i}, \qquad (3) \\ W_i n_{Q_i} \ &< \ Q_i n_{W_i}. \\ \mu_i - \mu_{i'} \ &= \ \frac{Q_i n_{W_i} - W_i n_{Q_i}}{n_{Q_i}(n_{Q_i} + n_{W_i})} > 0. \qquad (4) \end{aligned}$$

Therefore, $\mu_i > \mu_{i'}$ when $\eta_{\zeta'} < \eta_\zeta$.

To determine whether $J_\zeta > \tilde{J}_{\zeta'}$, let us start with a simple case of only two observations in cluster ζ, as demonstrated in the equation below. (Noted, the expression of $\tilde{J}_{\zeta'}$ is introduced to distinguish it from $J_{\zeta'}$, as different η values indicate different clustering outcomes. Here, $\tilde{J}_{\zeta'}$ is calculated for observations within the cluster ζ, corresponding to the η_ζ value.)

$$\begin{aligned} J_\zeta \ &= \ \frac{\sum Q_i}{\sum n_{Q_i}}, \qquad (5) \\ \tilde{J}_{\zeta'} \ &= \ \frac{\sum (Q_i + W_i)}{\sum (n_{Q_i} + n_{W_i})}. \qquad (6) \end{aligned}$$

To examine if $J_\zeta > \tilde{J}_{\zeta'}$, we can begin with a simple case with only two observations in a cluster ζ, which is shown in (7).

$$\begin{aligned} &\frac{Q_1 + Q_2}{n_{Q_1} + n_{Q_2}} - \frac{Q_1 + W_1 + Q_2 + W_2}{n_{Q_1} + n_{W_1} + n_{Q_2} + n_{W_2}} = \qquad (7) \\ &\frac{n_{W_1} Q_1 - n_{Q_1} W_1 + n_{W_2} Q_2 - n_{Q_2} W_2 + n_{W_2} Q_1 - n_{Q_1} W_2 + n_{W_1} Q_2 - n_{Q_2} W_1}{(n_{Q_1} + n_{Q_2})(n_{Q_1} + n_{W_1} + n_{Q_2} + n_{W_2})}, \end{aligned}$$

where n_{W_1} and n_{W_2} are non-zero. Following (3), $n_{W_1} Q_1 - n_{Q_1} W_1 > 0$ and $n_{W_2} Q_2 - n_{Q_2} W_2 > 0$. Moreover, all $Q_i/n_{Q_i} \geq \eta_\zeta$ and all $W_j/n_{W_j} < \eta_\zeta$ with $n_{W_j} \neq 0$; Thereby, even when $i \neq j$, $n_{Q_i} W_j < n_{W_j} Q_i$ holds. Then there are $n_{W_2} Q_1 - n_{Q_1} W_2 > 0$ and $n_{W_1} Q_2 - n_{Q_2} W_1 > 0$. Hence,

$$\frac{Q_1 + Q_2}{n_{Q_1} + n_{Q_2}} - \frac{Q_1 + W_1 + Q_2 + W_2}{n_{Q_1} + n_{W_1} + n_{Q_2} + n_{W_2}} > 0. \qquad (8)$$

Similarly, if $Q_1 + Q_2$, $n_{Q_1} + n_{Q_2}$, $W_1 + W_2$ and $n_{W_1} + n_{W_2}$ are written as $Q_{1\sim2}$, $n_{Q_{1\sim2}}$, $W_{1\sim2}$ and $n_{W_{1\sim2}}$, it is straightforward to show $J_\zeta - \tilde{J}_{\zeta'} > 0$ when there are more points in

the cluster. Therefore, $J_\zeta > \tilde{J}_{\zeta'}$. Moreover, $J_\zeta = \tilde{J}_{\zeta'}$ if and only if $\sum n_{W_i} = 0$.

So, we have $J_\zeta \geq \tilde{J}\zeta'$ for $\eta_\zeta > \eta_{\zeta'}$. □

Property 2. *For each observation $\mathbf{x}_i \in \mathcal{X}_\zeta$, reducing the threshold from η_ζ to $\eta_{\zeta'}$ implies $\tilde{\mu}_{i'} \leq \tilde{\mu}_i$. Additionally, if core neighbors of all $\mathbf{x}_i \in \mathcal{X}_{\zeta'}$ are also in $\mathcal{X}_{\zeta'}$, then the weighted between cluster similarity for observations in $\mathcal{X}_\zeta$ will either remain the same or decrease for smaller η values.*

Proof. Let $\eta_\zeta > \eta_{\zeta'} \geq \eta_X$. Assume that for every $\mathbf{x}_j$ satisfying $\mathbb{S}_{ij} \geq \eta_{\zeta'}$, $\mathbf{x}_j$ belongs to $\mathcal{X}_{\zeta'}$. Then for any $\mathbf{x}_i \in \mathcal{X}_\zeta$, there is,

$$\begin{aligned}\{\mathbf{x}_j \mid \mathbb{S}_{ij} \geq \eta_{\zeta'}\} &= \{\mathbf{x}_v \mid \mathbb{S}_{iv} \geq \eta_\zeta\} + \\ &\quad \{\mathbf{x}_h \mid \eta_{\zeta'} \leq \mathbb{S}_{ih} < \eta_\zeta\},\end{aligned}$$

where $\mathbf{x}_j \in \mathcal{X}_{\zeta'}$, $\mathbf{x}_v \in \mathcal{X}_\zeta$ and $\mathbf{x}_h \notin \mathcal{X}_\zeta$, which indicates that with a higher η, some observations may be assigned to other clusters.

$$\tilde{\mu}_{i'} = \frac{V_i}{n_{V_i}}, \tag{9}$$

$$\tilde{\mu}_i = \frac{V_i + L_i}{n_{V_i} + n_{L_i}}, \tag{10}$$

$$\tilde{H}_{\zeta'} = \frac{\sum V_i + n_{I_1}\eta_X}{\sum n_{V_i} + n_{I_1}}, \tag{11}$$

$$H_\zeta = \frac{\sum V_i + \sum L_i + n_{I_2}\eta_X}{\sum n_{V_i} + \sum n_{L_i} + n_{I_2}}, \tag{12}$$

where $V_i = \sum\{\mathbb{S}_{ih} \mid \eta_{\zeta'} > \mathbb{S}_{ih} \geq \eta_X\}$, $n_{v_i} = |\{\mathbf{x}_h \mid \eta_{\zeta'} > \mathbb{S}_{ih} \geq \eta_X\}|$, $n_{I_1} = |\{\mathbf{I}_{\zeta'}\}|$, $L_i = \sum\{\mathbb{S}_{ih} \mid \eta_{\zeta'} \leq \mathbb{S}_{ih} < \eta_\zeta\}$, $n_{L_i} = |\{\mathbf{x}_h \mid \eta_{\zeta'} \leq \mathbb{S}_{ih} < \eta_\zeta\}|$ and $n_{I_2} = |\{\mathbf{I}_{\zeta'}\}|$. Similar to the proof of Property 1, it can be shown that $\tilde{\mu}_{i'} - \tilde{\mu}_i < 0$ when $n_{L_i} > 0$ and $\tilde{\mu}_{i'} = \tilde{\mu}_i$ when $n_{L_i} = 0$.

For the case $\sum n_{L_i} > 0$, consider

$$\frac{\sum V_i + n_{I_1}\eta_X}{\sum n_{V_i} + n_{I_1}} - \frac{\sum V_i + \sum L_i + n_{I_2}\eta_X}{\sum n_{V_i} + \sum n_{L_i} + n_{I_2}} = \frac{\sum V_i \sum n_{L_i} - \sum L_i \sum n_{V_i} + n_{I_1}(\sum n_{L_i}\eta_X - \sum L_I) + (n_{I_1} - n_{I_2})(\sum n_{V_i}\eta_X - \sum V_i)}{(\sum n_{V_i} + n_{I_1})(\sum n_{V_i} + \sum n_{L_i} + n_{I_2})}.$$

Because $\eta_X \leq \sum V_i / \sum n_{V_i} < \eta_{\zeta'} \leq \sum L_i / \sum n_{L_i}$, it follows that $\sum V_i \sum n_{L_i} - \sum L_i \sum n_{V_i} < 0$ and $n_{I_1}(\sum n_{L_i}\eta_X - \sum L_I) < 0$. Since $\eta_{\zeta'} < \eta_\zeta$, $n_{I_1} \geq n_{I_2}$ (With

a larger η, less number of inner observations). Given that $\sum \eta_X \leq \sum L_I / n_{L_i}$, it follows that $(n_{I_1} - n_{I_2})(\sum n_{V_i}\eta_X - \sum V_i) \leq 0$. Therefore, $\tilde{H}_{\zeta'} < H_\zeta$ when $\sum n_{L_i} > 0$.

If $\sum n_{L_i}$ equals 0, then H_ζ equals $\tilde{H}_{\zeta'}$. So, we have $H_\zeta \geq \tilde{H}\zeta'$ for $\eta_\zeta > \eta_{\zeta'}$.

□

Properties 3 and 4 naturally follow from the definitions of outliers and inner observations.