kNN: From a Topological Data Analysis Perspective to a Topological Combinatorics Approach

MINH QUANG LE
HO CHI MINH CITY OPEN UNIVERSITY

kNN-graphs and kNN-complexes

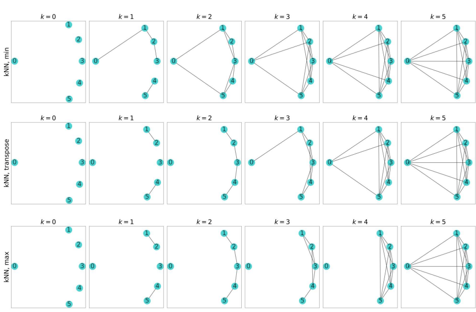
kNN-ordering: Let $\mathcal{V} = \{1, \dots, N\}$ enumerate a set of points $\mathcal{Y} = \{y^{(i)}\}_{i \in \mathcal{V}}$ with $y^{(i)}(t) \in \mathbb{R}^p$ in a (Euclidean) metric space and $\left\{f_{ij} = \|y^{(i)} - y^{(j)}\|_2\right\}$ be their pairwise distances. For each i, let k_{ij} denote the **nearest-neighbor order** of $y^{(j)}$ with respect to $y^{(i)}$ so that $y^{(j)}$ is the k_{ij}-th nearest neighbor of $y^{(i)}(k_{ii} = 0)$. The ordering is formally defined for each fixed i by $\{k_{ij}\}_{j=1}^N = \phi\left(\{f_{ij}\}_{j=1}^N\right)$, where ϕ: $\mathbb{R}^N \longrightarrow \mathbb{N}^N$ is the "argsort function".

kNN-neighborhood: Given a set of kNN orderings $\{k_{ij}\}$ for $i, j \in \mathcal{V}$, we define the **kNN neighborhoods** as the sets $\mathcal{N}_{ik} = \{j | k_{ij} \le k\}$ for each $i \in \mathcal{V}$ and $k \in \{0\} \cup \mathcal{V} \backslash \{N\}$.

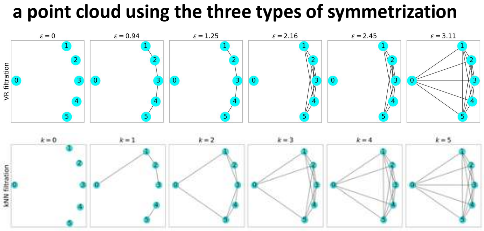
Symmetrized kNN Orderings and Neighborhoods: We define three types of **symmetrized kNN neighborhoods**---$\mathcal{N}_{ik}^{min}$, $\mathcal{N}_{ik}^{max}$, and $\mathcal{N}_{ik}^{trans}$---that use, respectively, the following three symmetrizations of kNN orderings k_{ij}: $\tilde{k}_{ij} = min(k_{ij}, k_{ji})$; $\bar{k}_{ij} = (k_{ij} + k_{ji})/2$ and $\hat{k}_{ij} = max(k_{ij}, k_{ji})$.

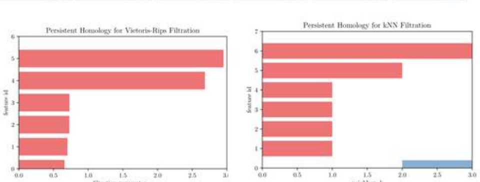
Filtrations persistent homology using kNN-complexes

Global kNN Filtration: Let $X_k, X_{k'} \in \{X_k^{min}, X_k^{ma}, X_k^{trans}\}$ be kNN complexes with the same symmetrization method. Then we define a **kNN-filtered simplicial complex:** $\{X_k \longrightarrow X_{k'}\}_{0 \le k \le k' \le N-1}$



Visualization of kNN-filtered simplicial complexes for a point cloud using the three types of symmetrization





Comparison of persistent homology for an example point cloud using two different filtrations kNN,min filtration; and a VR filtration.

Stability for kNN-persistence diagrams

kNN Equivalence of Point Sets: Let F denote the neighbor-ordering function and consider two enumerated point sets $\mathcal{Y}, \mathcal{Z} \in \mathcal{A}_p^N$ such that $\mathcal{Y} = \{y^{(i)}\}_{i \in \mathcal{V}}$ and $\mathcal{Z} = \{z^{(i)}\}_{i \in \mathcal{V}}$. Point sets $\mathcal{Y}, \mathcal{Z}$ are said to be **kNN equivalent** if $F(\mathcal{Y}) = F(\mathcal{Z})$. Elements $y^{(i)}$ and $z^{(i)}$ are said to be **locally kNN equivalent** if $F_i(\mathcal{Y}) = F_i(\mathcal{Z})$.
Letting $\sim$ represent kNN equivalence, we further define $[\mathcal{Y}] = \{\mathcal{Y}' | \mathcal{Y}' \sim \mathcal{Y}\}$ as the kNN equivalence class of $\mathcal{Y}$.

kNN-preserving transformations: Consider an enumerated point set $\mathcal{Y} = \{y^{(i)}\}_{i=1}^N \in \mathcal{A}_p^N$, and let $F: \mathcal{A}_p^N \longrightarrow \mathbb{N}_+^{N \times N}$ and $F_i: \mathcal{A}_p^N \longrightarrow \mathbb{N}_+^{N \times N}$ be the full and row-wise kNN ordering functions. We define a function $h: \mathcal{A}_p^N \longrightarrow \mathcal{A}_p^N$ that takes the form $h(\mathcal{Y}) = \{h_1(y^{(1)}), \dots, h_N(y^{(N)})\}$ to be **global kNN preserving** for $\mathcal{Y}$ if $F(h(\mathcal{Y})) = F(\mathcal{Y})$. That is, the action of F on $\mathcal{Y}$ is invariant to the transformation h. Similarly, we define h to be **local kNN preserving** for point $y^{(i)} \in \mathcal{Y}$ if F_i is invariant to the transformation, i.e., $F_i(h(\mathcal{Y})) = F_i(\mathcal{Y})$.

Stability of kNN persistence diagrams: Consider two enumerated point clouds $\mathcal{Y}, \mathcal{Z} \in \mathcal{A}_p^N$ and the metric $\|\mathcal{Y} - \mathcal{Z}\|_\infty = max_i \|y^{(i)} - z^{(i)}\|$ over the space of enumerated point clouds. Letting F be the kNN ordering function, we define matrices $G = F(\mathcal{Y})$ and $H = F(\mathcal{Z})$ with entries G_{ij} and H_{ij}. Further denote their associated kNN persistence diagrams $\mathcal{D}_\mathcal{Y}$ and $\mathcal{D}_\mathcal{Z}$. Then there exists a uniform global bound of the form $d_B(\mathcal{D}_\mathcal{Y}, \mathcal{D}_\mathcal{Z}) \le \|G - H\|_\infty = max_{ij}|G_{ij} - H_{ij}|$; however, there does not exist a bound of the form: $d_B(\mathcal{D}_\mathcal{Y}, \mathcal{D}_\mathcal{Z}) \le \|G - H\|_\infty \nleq L\|\mathcal{Y} - \mathcal{Z}\|_\infty$ for some constant L.

Proof: The first inequality is satisfied by applying the original statement of the stability theorem: persistence diagrams are uniformly bound by the maximal difference between the functions that are filtered. That is, the difference in persistence diagrams is bound by the maximum difference in neighbor orderings.
However, the stability of persistent homology with respect to neighbor-ordering changes does not imply stability with respect to point-location changes. This occurs because kNN orderings are not stable to point perturbations.
To disprove the second inequality, we provide a counter example. Recall the sets of $N = 3$ points in below figure with locations $\mathcal{Y} = \{-1, -\varepsilon, 1\}$ and $\mathcal{Z} = h(\mathcal{Y}) = \{-1, \varepsilon, 1\}$, with $\varepsilon \in \mathbb{R}$. By construction, $\|G - H\|_\infty = 1$ for any small ε, but $\|\mathcal{Y} - \mathcal{Z}\|_\infty = 2\varepsilon$.
Suppose there did exist a uniform bound with Lipschitz constant L, then the points $\mathcal{Y}$ and $\mathcal{Z}$ with any $\varepsilon < 1/2L$ yields a contradiction since {one would must have both $\|G - H\|_\infty = 1$ and $\|G - H\|_\infty < 1$.

Convergence in Norm Implies Convergence in kNN Topology: Let $\mathcal{Y}(t) = \{y^{(i)}(t) | t \in V\}$, with $y^{(i)}(t), t \in \mathbb{R}^p$ be an element of a sequence of point clouds in which each point converges in norm to some limit, $\lim_{t \to \infty} \|y^{(i)}(t) - y^{(i)}\| = 0$ for each $i \in V$.
We assume the points $y^{(i)}$ to be non-degenerate (i.e., we assume $\|y^{(i)} - y^{(j)}\|_2 \ne \|y^{(i)} - y^{(k)}\|_2$ for any $k \ne j$). Define $\mathcal{Y} = \{y^{(i)} | i \in \mathcal{V}\} \in \mathcal{A}_p^N$ as that limit, and assume uniform convergence $\mathcal{Y}(t) \to \mathcal{Y}$
Then $\mathcal{Y}(t) \to \mathcal{Y}$ in kNN topology, and there exists a t^* such that $F(\mathcal{Y}(t)) = F(\mathcal{Y})$ for any $t > t^*$, where F is the neighbor-ordering

Nearest-neighbor path homology

NN-path chain complex: We define the nearest-neighbor path in the point cloud data $\mathcal{V}$ as $p = x_0 x_1 \dots x_n$, where $x_i \in \mathcal{N}(x_{i-1})$, i.e., x_i is a nearest neighbor of x_{i-1} and $x_i \ne x_j, i \ne j$. Let C_n be the vector space over $\mathbb{K} = \mathbb{F}_2$ generated by all such n-nearest-neighbor paths.

$$C_n = \mathbb{K}\langle x_0 x_1 \dots x_n \rangle | x_i \in \mathcal{N}(x_{i-1}), x_i \in \mathcal{V}, i = 1,2, \dots, n$$

The boundary operator:
$\partial_n: C_n \to C_{n-1}$ is defined on basis elements by: $\partial_n = \sum_{i=1}^{n-1} \partial_n^i$ where $\partial_n^i\ (x_0 x_1 \dots x_n) = x_0 \dots x_{i-1} \hat{x}_i x_{i+1} \dots x_n$ where $x_{i+1} \in \mathcal{N}(x_{i-1})$, and 0 otherwise.

Theorem: The boundary operator ∂_n satisfies $\partial_n^2 = 0$.

Starting point and decomposition: We define a chain complex, called a nearest neighbor (NN-)chain complex, and denoted by C_l^a ($a \in X$ is a starting point).
In general, we denote $C_l^{a,k}$ as a set of paths γ starting at a point $a \in X$, $x_{i+1} \in N_k(x_i)$ with $x_{i+1}, x_i \in \gamma$
Also, let $A^{(1)}$ be the set of starting points within nearest neighbor; $A^{(k)}$ be the set of starting points within k —nearest neighbor.
We have a descending sequence as follows:

$$A^{(1)} \supseteq A^{(2)} \supseteq \cdots \supseteq A^{(l)}$$

Also, we define: $C_l^{(1)} = \coprod_{a \in A^{(1)}} C_l^a$
And $C_l^{(k)} = \coprod_{a \in A^{(k)}} C_l^{a,k}$
Lemma:

$$\exists i_k \in [1, l]: C_l^{(1)}\ |_{A^{(i_k)}} = C_l^{(i_k)}$$

In NN-path, we can define a starting point as follows: a point $a \in X$ is a starting point of a path γ if
either $\exists b \in \gamma: b \in \mathcal{N}(a)$ and $\nexists c \in \gamma: a \in \mathcal{N}(c)$
or $\exists (a, b, c) \in \gamma\ s.t.\ abc$ is an equilateral triangle (a form of Δ)

Lemma: Given a point cloud data X, and supposed that $|X| \ge 4$, then the cycle, if existing, will only be in the form of Δ

(NN)-preserving function: given 2 (finite) point cloud data X, Y, we say $f: X \longrightarrow Y$ a NN-preserving function if $y \in \mathcal{N}(x)$ then $f(y) \in \mathcal{N}(f(x))$.

Noticed that NN-preserving function does not guarantee existing cycles in path is preserved.

Stability theorem

Height function:
Given FPD with a metric d, and $\gamma \in C_l^a$(a is a starting point, l is the length), we define the height function of γ, denoted $h(\gamma)$ as follows: $h(\gamma) = \sum_{i=1}^l d(x_{i-1}, x_i)$.

Stability theorem: fixed the length l, let $f, g: X \to \mathbb{R}$ define as follows: $f(a) = \begin{cases} \min_\gamma\{h(\gamma) | \gamma(0) = a, l(\gamma) = l\} \\ 0; otherwise \end{cases}$; $g(a) = \begin{cases} \min_\gamma\{h(\gamma) | \gamma(0) = a, l(\gamma) = l + 1\} \\ f(a), otherwise \end{cases}$. If for all starting points $a \in A$, we can extend one more step for all γ in definition f at least ε in their height function, then the Bottleneck distance between their Persistence Diagrams induced from these functions approach to 0 as ϵ is sufficiently small.