Skeleton Clustering: Dimension-Free Density-based Clustering

Zeyu Wei *(zwei5@uw.edu)*, Yen-Chi Chen *(yenchic@uw.edu)*

University of Washington, Department of Statistics

Abstract

- Introduce a **skeleton clustering framework** that combines various clustering approaches.
- Propose multiple **density-based similarity measures** that scale well with dimensions.
- Prove the consistency of the sample estimates of the proposed similarity measures.
- Use simulations and real data to show the reliability and usefulness of our method in different scenarios.

Motivation

- **Task:** Cluster high-dimensional data with unbalanced groups and complex cluster shapes.
- Density-based clustering **advantages**: can handle irregular shapes; nice interpretation and estimation based on the underlying PDF. **Limitation:** not suitable for high-dimensional data due to curse of dimensionality.
- **Intuition:** Borrow the idea of **merging a large number of clusters** from (Peterson et al., 2018; Baudry et al., 2010).Also propose **density-based similarity measures** suited for high-dimensional settings.

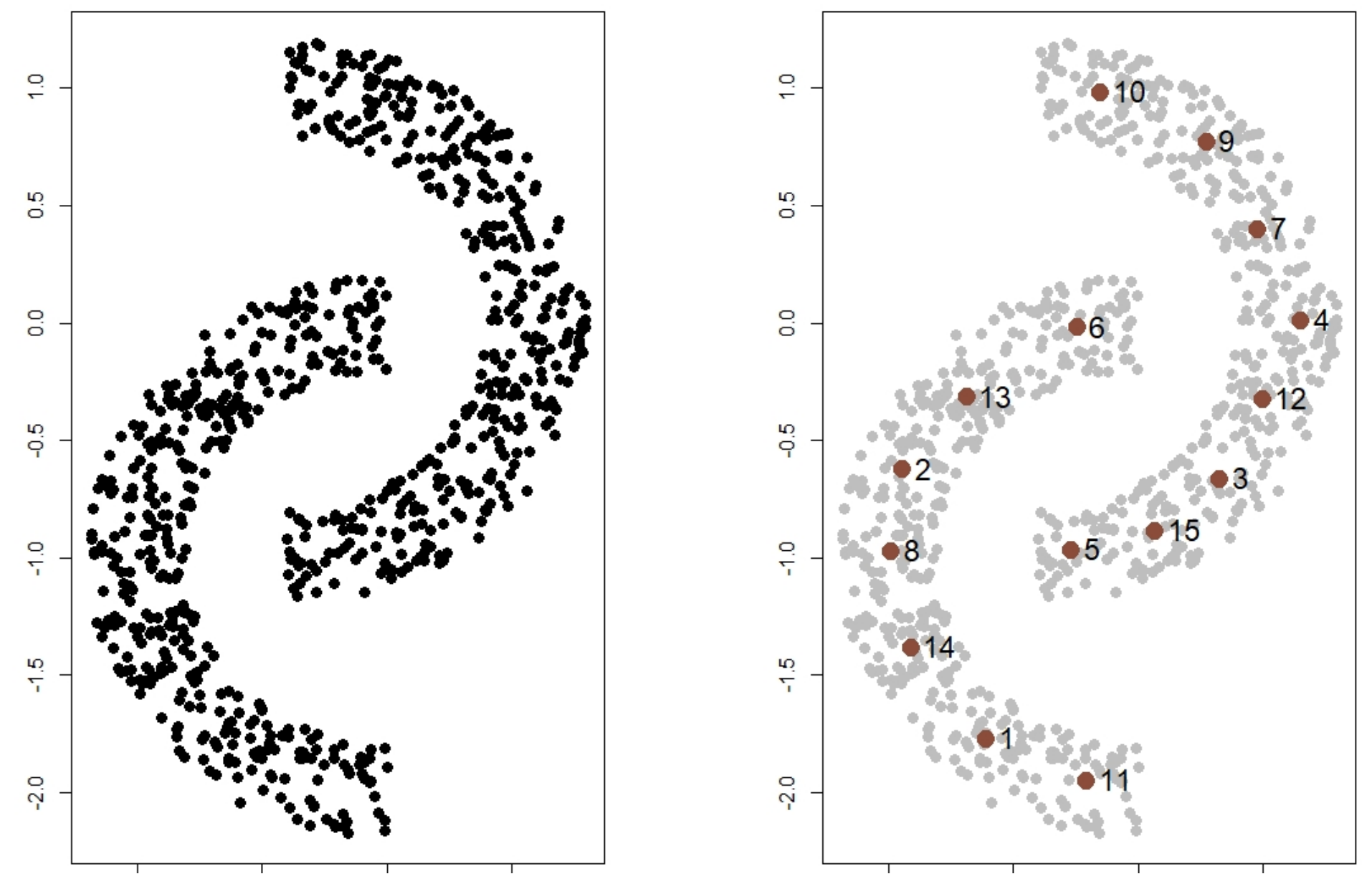
Skeleton Clustering Framework

Algorithm 1 Skeleton Clustering

```
Input: Observations $X_1, \cdots, X_n$; final number of clusters $S$.
1. Knot construction. Perform $k$-means clustering with a large number of $k$; the centers are the knots.
2. Edge construction. Apply approximate Delaunay triangulation to the knots.
3. Edge weights construction. Add weights to each edge using Voronoi density, Face density, or Tube density approach.
4. Knots segmentation. Use linkage criterion to segment knots based on the edge weights into $S$ groups.
5. Assignment of labels. Assign cluster labels to each observation based on which knot-group of the nearest knot.
```

Knots Construction

Knots are constructed to give a concise representation of the data structure. In practice we use k-Means to choose $k = [\sqrt{n}]$ knots, where n is the number of samples.



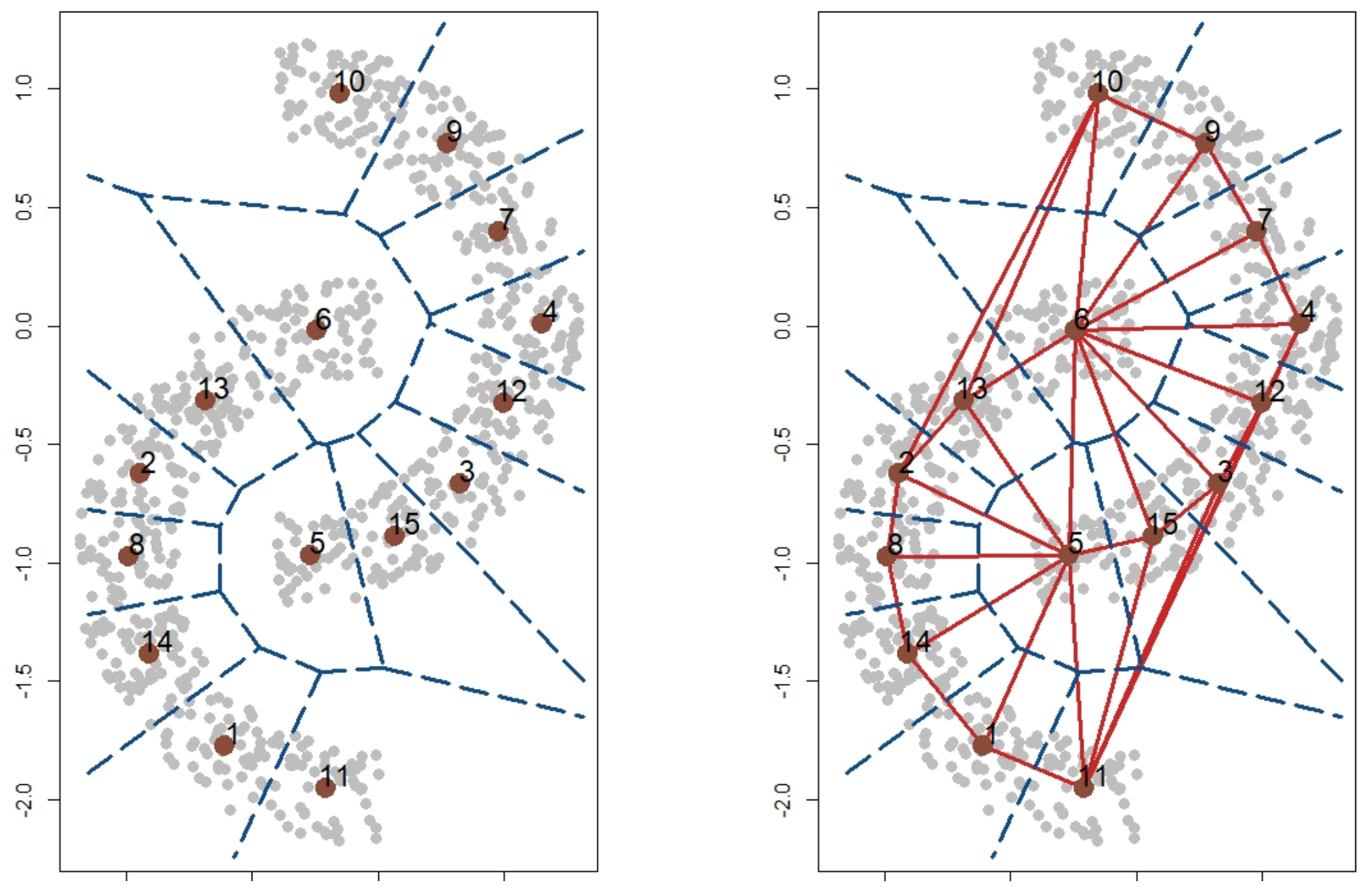
(a) Data (b) Knots

Edge Construction

The Voronoi cell (Voronoi, 1908), $\mathbb{C}_j$, associated with knot c_j is the set of all points in $\mathcal{X}$ whose distance to c_j is the smallest compared to other knots. That is,

$$\mathbb{C}_j = \{x \in \mathcal{X} : d(x, c_j) \leq d(x, c_\ell) \ \ \forall l \neq j\},$$

where $d(x, y)$ is the usual Euclidean distance.

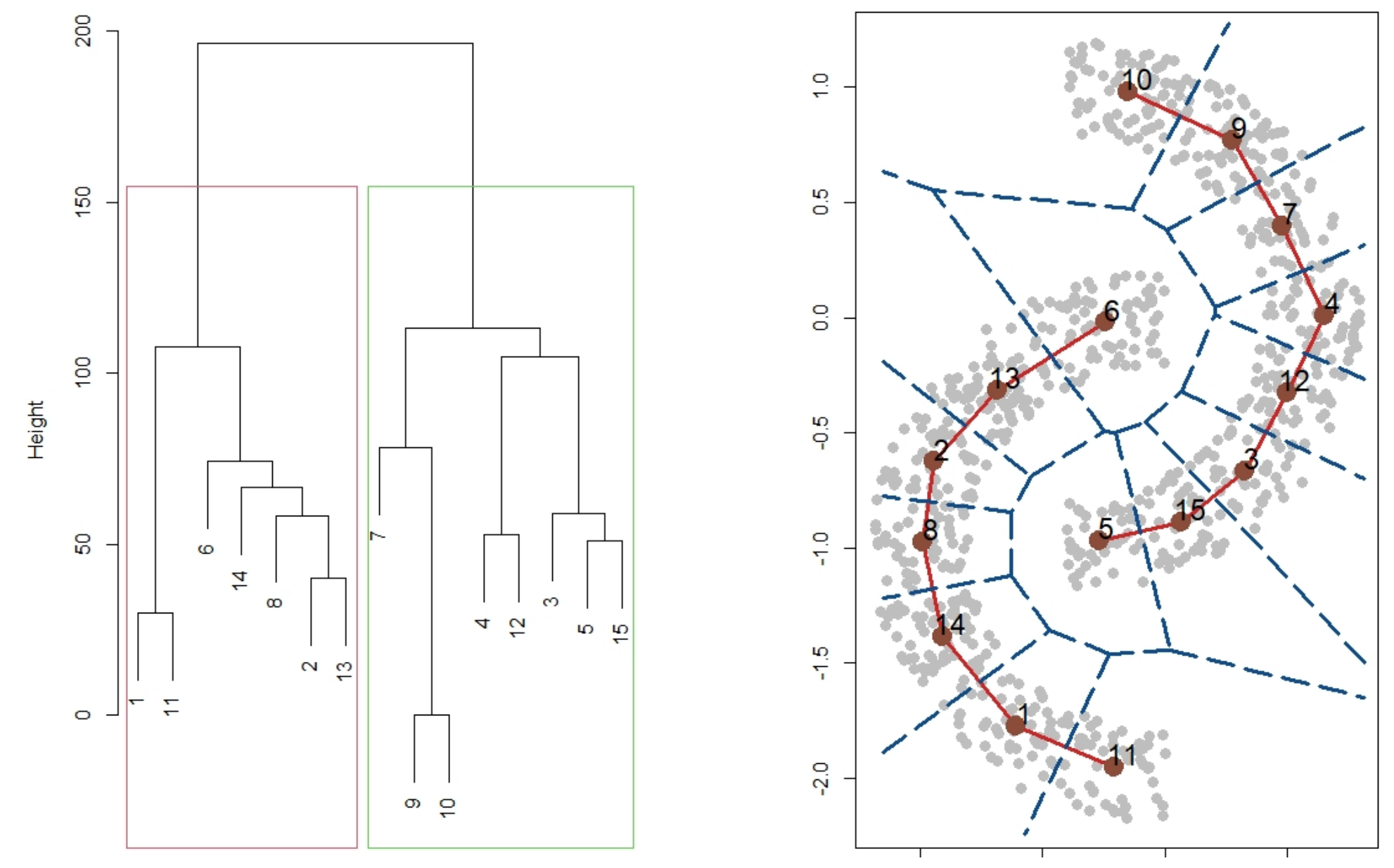


(a) Voronoi Cells (b) Delaunay Triangulation

An edge between knots (c_i, c_j) is added if $\bar{\mathbb{C}}_i \cap \bar{\mathbb{C}}_j \neq \emptyset$. Resulting graph is the Delaunay triangulation $DT(\mathcal{C})$ (Delaunay, 1934) of knots $c_1, \cdots, c_k$

Skeleton Segmentation

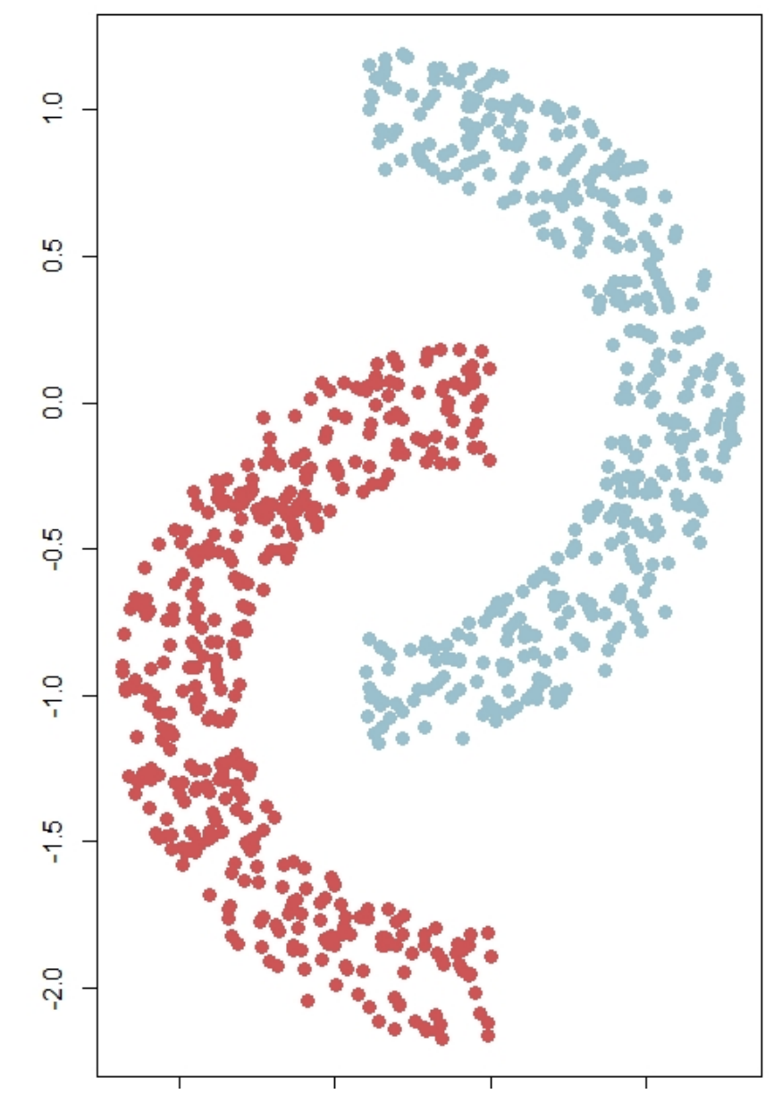
Density-based weights are assigned to the edges. We then use traditional clustering/segmentation methods such as the hierarchical clustering to segment the learnt skeleton structure.



(a) Dendrogram (b) Segmented Skeleton

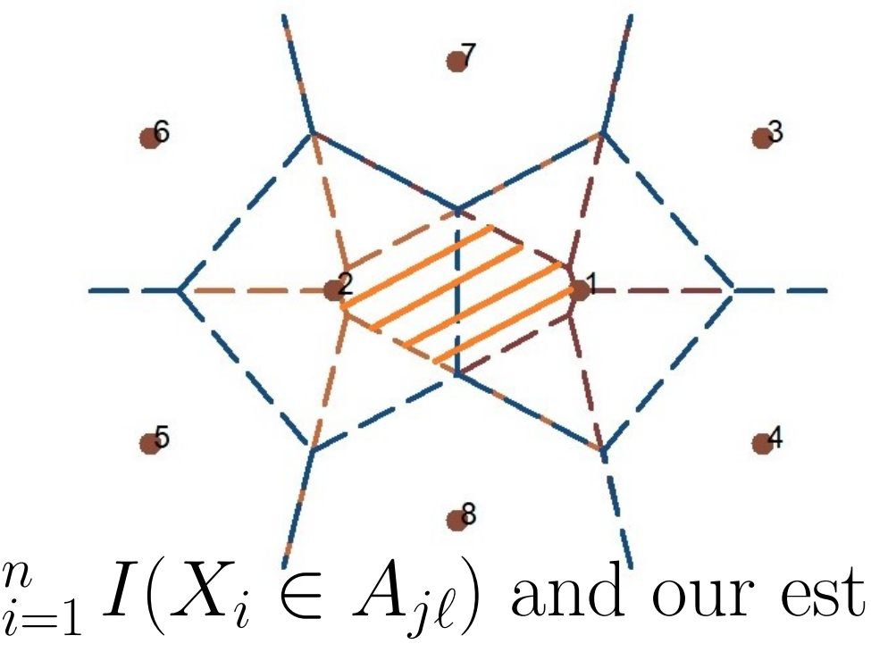
Label Assignment

Assign the individual labels according to the segmented skeleton. In practice we assign the labels the same as the nearest knot.



Edge Weight: Voronoi Density (VD)

Define the 2-NN region as $A_{j\ell} \equiv \{x \in \mathcal{X} : d(x, c_i) > max\{d(x, c_j), d(x, c_\ell)\}, \forall i \neq j, \ell\}$. The *Voronoi density (VD)*is defined as $S_{j\ell}^{VD} = \frac{\mathbb{P}(A_{j\ell})}{\|c_j - c_\ell\|}$.



Let $\hat{P}_n(A_{j\ell}) = \frac{1}{n}\Sigma_{i=1}^{n} I(X_i \in A_{j\ell})$ and our estimator is

$$\hat{S}_{j\ell}^{VD} = \frac{\hat{P}_n(A_{j\ell})}{\|c_j - c_\ell\|}.$$

which is dimension independent

Edge Weight: Face Density (FD)

For connected components we expect to see many observations around their mutual boundary. Let the face region between two knots c_j, c_ℓ be $F_{j\ell} \equiv \mathbb{C}_j \cap \mathbb{C}_\ell$. Then the *Face Density (FD)* is defined as the PDF integrated over the face region:

$$S_{j\ell}^{FD} = \textstyle\int_{F_{j\ell}} p(x)dx = \int_{F_{j\ell}} d\mathbb{P}(x).$$

For estimation, note that the boundary of two Voronoi regions is orthogonal to the line passing through the two corresponding knots and is at the middle point. Let $\Pi_{j\ell}(x)$ be the projection of $x \in \mathcal{X}$ onto the line passing through c_j and c_ℓ. The estimator $\hat{S}_{j\ell}^{FD}$ is defined as

$$\hat{S}_{j\ell}^{FD} = \frac{1}{nh}\Sigma_{X_i \in \mathbb{C}_j \cup \mathbb{C}_\ell} K\left(\frac{\Pi_{j\ell}(X_i) - (c_\ell + c_j)/2}{h}\right)$$

which is 1-D KDE.

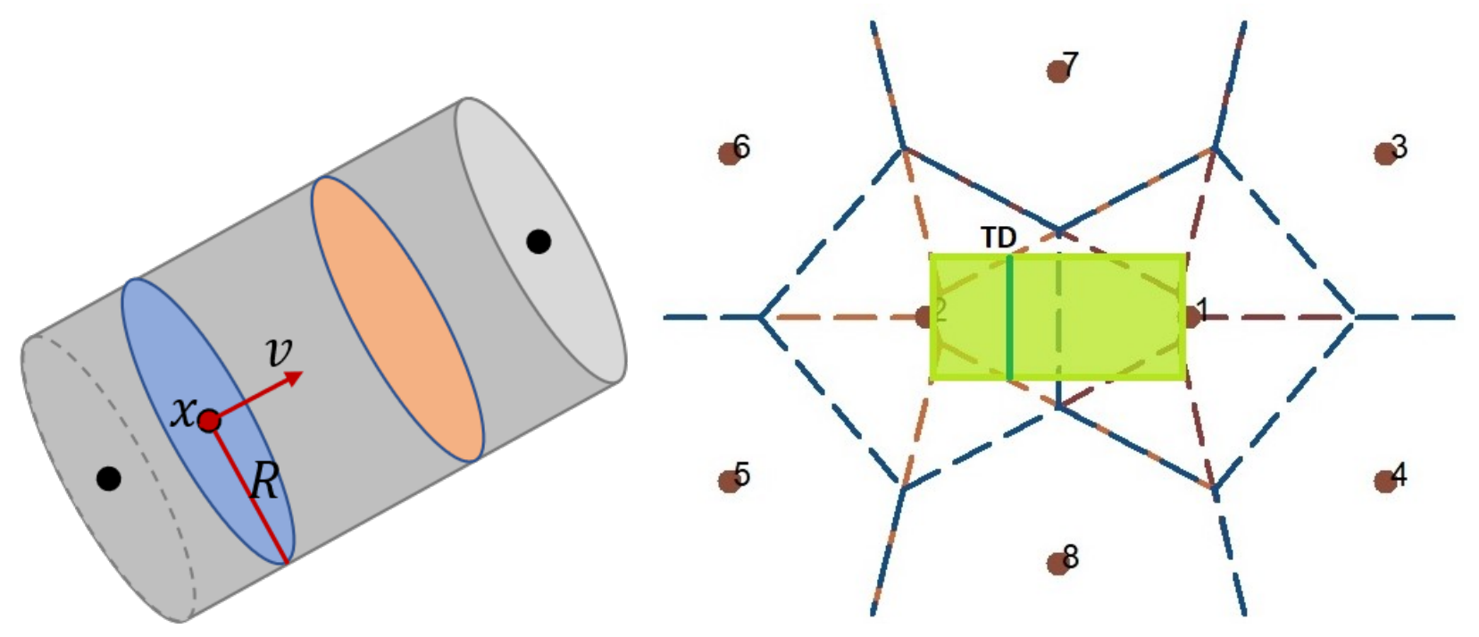
Edge Weight: Tube Density (TD)

Define a disk area centered at x with radius R and normal direction ν as $\mathsf{Disk}(x, R, \nu) = \{y : ||x - y||_2 \leq R, (x - y)^T \nu = 0\}$. Define the integrated density in the disk region as

$$\begin{aligned}\mathsf{pDisk}_{j\ell,R}(t) &= \mathbb{P}\left(\mathsf{Disk}(c_j + t(c_\ell - c_j), R, c_\ell - c_j)\right)\\ &= \textstyle\int_{\mathsf{Disk}(c_j + t(c_\ell - c_j), R, c_\ell - c_j)} p(x)dx.\end{aligned}$$

Tube density (TD) is the minimal disk density along the central line

$$S_{j\ell}^{TD} = \inf_{t \in [0,1]} \mathsf{pDisk}_{j\ell,R}(t).$$



Let $\Pi_{j\ell}(x)$ be the projection of a point x on the line through c_j, c_ℓ. Get the estimate $\widehat{\mathsf{pDisk}}_{j\ell,R}(t)$ as

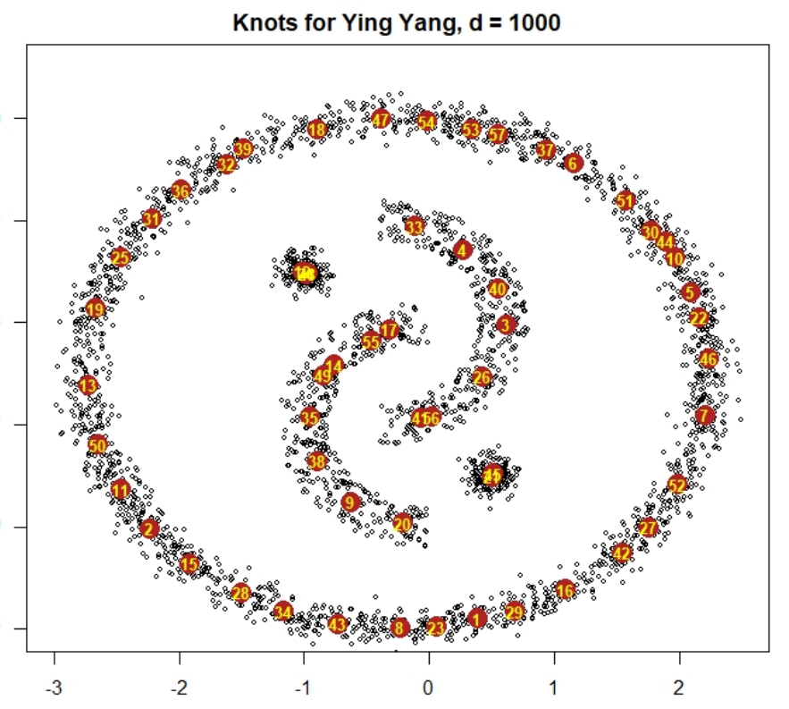
$$\frac{1}{nh}\Sigma_{i=1}^{n} K\left(\frac{\Pi_{j\ell}(X_i) - c_j - t(c_\ell - c_j)}{h}\right) I(||X_i - \Pi_{j\ell}(X_i)|| \leq R)$$

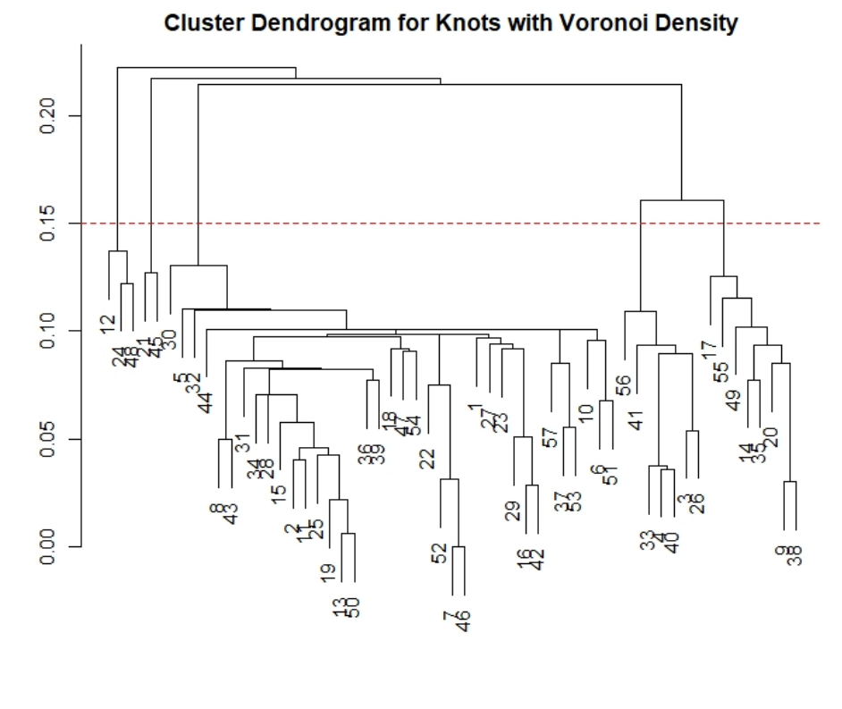
Estimate the TD as

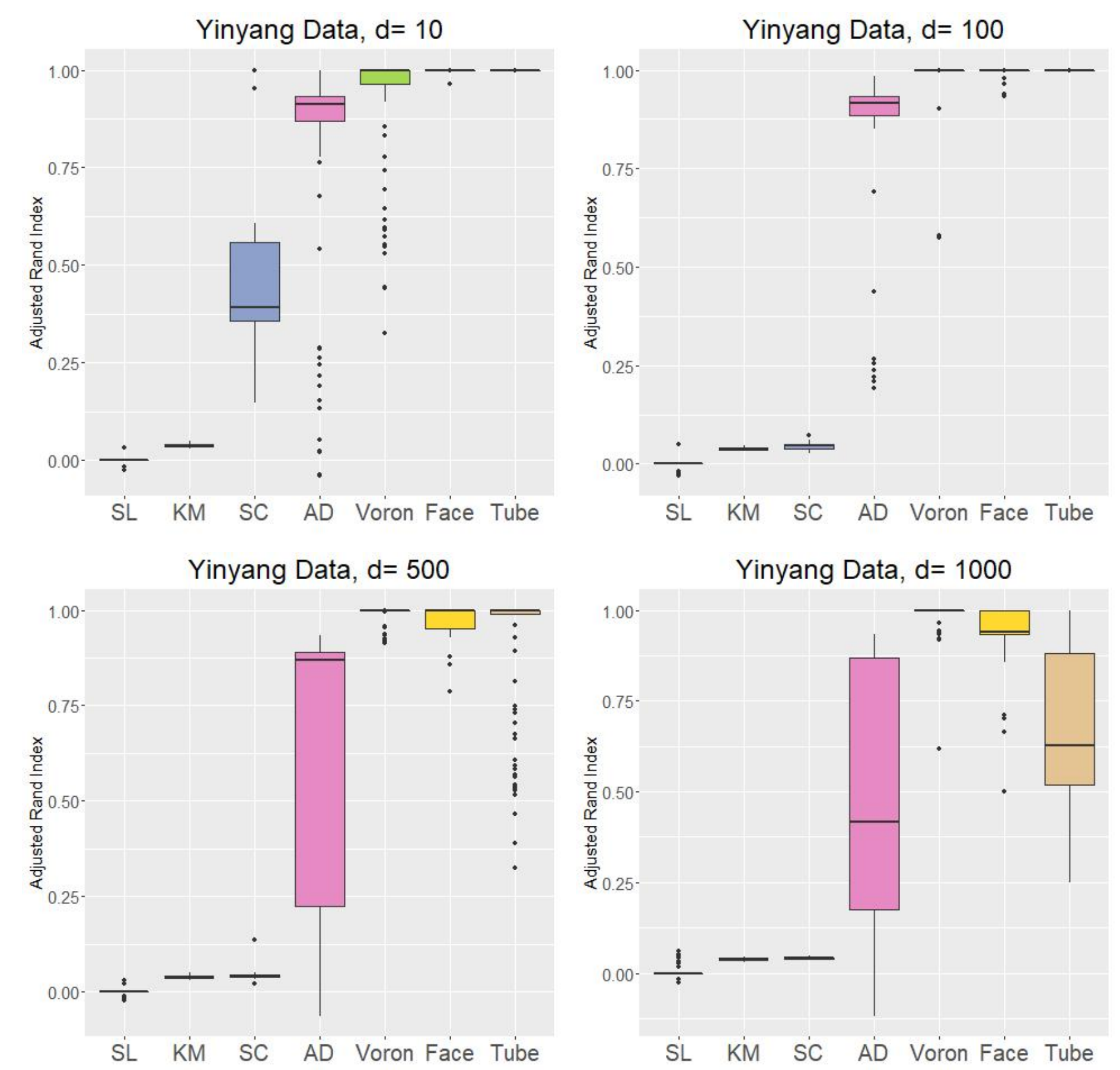
$$\hat{S}_{j\ell}^{TD} = \min_{t \in [0,1]} \widehat{\mathsf{pDisk}}_{j\ell,R}(t).$$

Simulation: Yinyang Data

- Intrinsically 2-dimensional data containing 5 components with different shapes. ($n = 3200$, $k = 57$)
- Additional variables from Gaussian noise $N(0, 0.1)$. Increase the dimension of noise variables so that the total dimensions are $d = 10, 100, 500, 1000$.
- Empirically compare: direct single-linkage hierarchical clustering (SL), direct k-means clustering (KM), spectral clustering (SC), skeleton clustering with average distance density (AD), skeleton clustering with Voronoi density (Voron), skeleton clustering with Face density (Face), and skeleton clustering with Tube density (Tube).







Conclusion & Future Work

- Clustering high-dimensional data with complex cluster shapes.
- Bypass the curse of dimensionality by using surrogate density such as Voronoi density, Face density, and Tube density. Showed the consistency of estimated similarity measures.

Some possible future directions:

- Skeleton clustering with similarity matrix.
- Theory accounting for the randomness of knots.
- Detection boundary points between clusters.
- Clustering after dimension reduction.

Additional Info

Code: https://github.com/JerryBubble/skeletonClus