Skeleton Clustering: Dimension-Free Density-based Clustering

Jerry Wei
Department of Statistics, University of Washington
and
Yen-Chi Chen
Department of Statistics, University of Washington

Density-based Clustering

Problem: Cluster high-dimensional data with unbalanced groups and complex cluster shapes.

Idea: a cluster in a data space is a contiguous region of high point density

Examples: Mode Clustering, Level-Set Clustering, DBSCAN, Cluster Tree

Advantages:

- capable of finding clusters with irregular shapes
- nice interpretation based on the underlying PDF
- can view the clustering problem as an estimation problem

Limitation: the curse of dimensionality for density estimation step, and hence not suitable for high-dimensional data.

Clustering High-dimensional Data

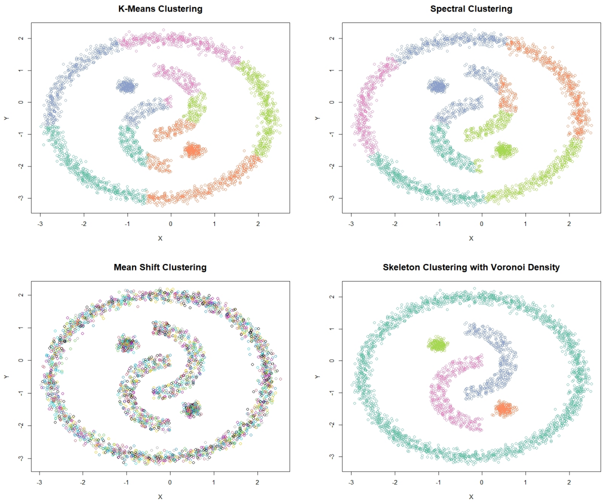


Figure: Yinyang Data with dimension 200.

Main Intuitions

- Borrow the idea of merging a large number of clusters from (Peterson et al., 2018; Fred and Jain, 2005; Maitra, 2009; Baudry et al., 2010).
- Propose density-based similarity measures similar to that in (Nugent and Stuetzle, 2010) but are suited for high-dimensional settings.

Main Contributions

- We introduce a skeleton clustering framework that combines various clustering approaches.
- We propose multiple density-based similarity measures scale well with dimensions.
- We use simulation to show the reliability of our method in agnostic scenarios.
- We show that our method can lead to meaningful clusters in real data.

Skeleton Clustering Framework

Let our training data $\mathbb{X} = \{X_1, \ldots, X_n\}$ be an IID sample from an unknown distribution with density p supported on a compact set $\mathcal{X} \in \mathbb{R}^d$. The goal of clustering is to partition $\mathbb{X}$ into clusters $\mathbb{X}_1, \ldots \mathbb{X}_S$, where S is the number of clusters.

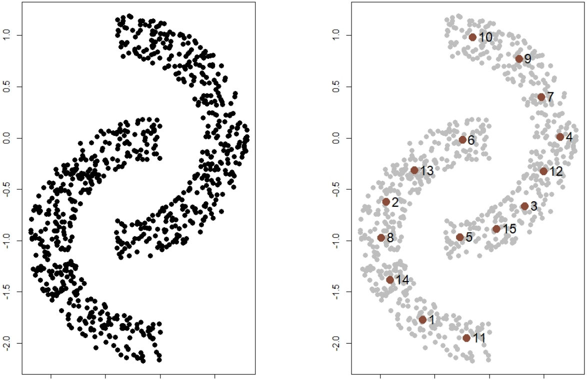
Algorithm 1 Skeleton Clustering

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. Generally, we choose $k = [\sqrt{n}]$.
2. **Edge construction.** Apply the Delaunay triangulation to the knots.
3. **Edge weights construction.** Add weights to each edge using either 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

- Some 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.
- Empirically robustness performance with sufficient number of knots.



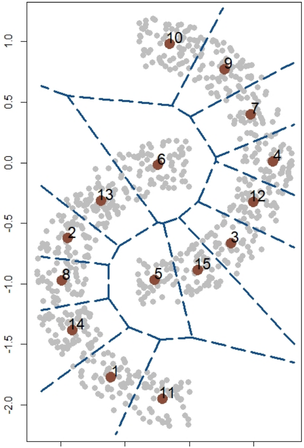
(a) Data (b) Knots

Edge Construction, Voronoi Cells

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,

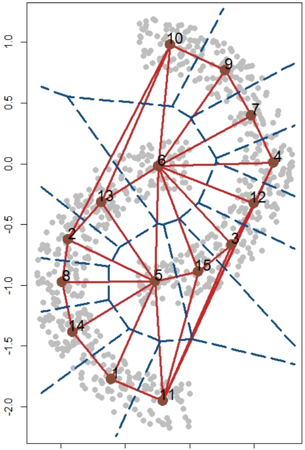
$$\mathbb{C}_j = \{x \in \mathcal{X} : d(x, c_j) \leq d(x, c_\ell) \quad \forall l \neq j\},$$

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



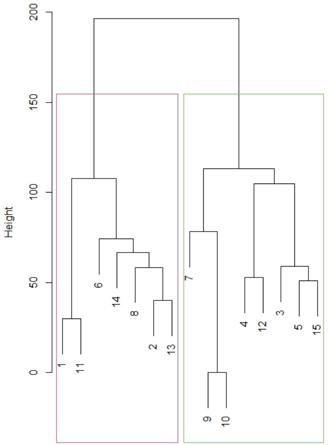
Edge Construction, Delaunay Triangulation

- Add an edge to a pair of knots if they are neighboring with each other. In other words, an edge between (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 $c_1, \cdots, c_k$



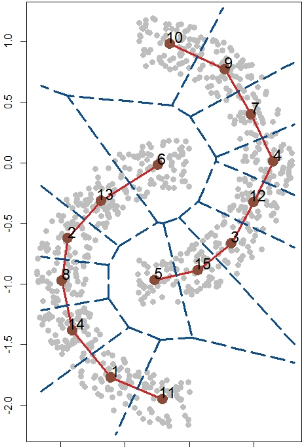
Skeleton Segmentation

- Density-based weights are assigned to the edges (discussed later).
- Use traditional clustering/segmentation methods such as the hierarchical clustering to segment the learnt skeleton structure.



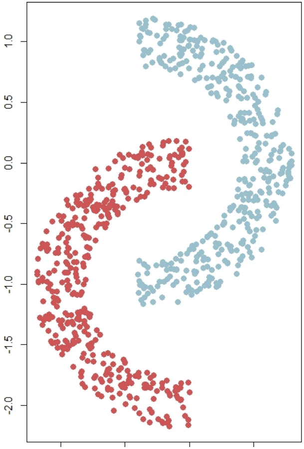
Skeleton Segmentation

The segmented skeleton is:



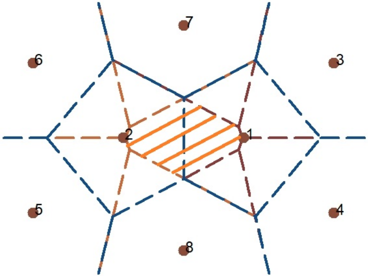
Label Assignment

- 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

- Measures the similarity between knots (c_j, c_ℓ) based on the number of observations whose 2-nearest knots are c_j and c_ℓ.
- 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 $S_{j\ell}^{VD} = \frac{\mathbb{P}(A_{j\ell})}{\|c_j - c_\ell\|}$.

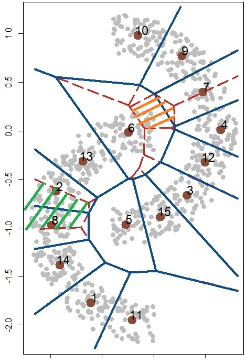


Edge Weight: Voronoi Density Estimation

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

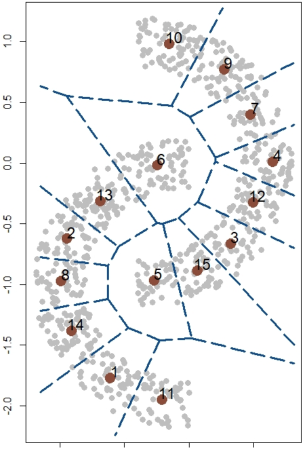
$$\hat{S}_{j\ell}^{VD} = \frac{\hat{P}_n(A_{j\ell})}{\|c_j - c_\ell\|}. \quad (1)$$

- Essentially counting points in the 2-NN region, which can be computed fast by k-d tree algorithm (Bentley, 1975)
- Dimension independent



Edge Weight: Face Density (FD)

- For connected components we expect to see many observations around their mutual boundary.
- The *Face Density (FD)* as the PDF integrated over the face region.
- let the face region between two knots c_j, c_ℓ be $F_{j\ell} \equiv \mathbb{C}_j \cap \mathbb{C}_\ell$. Then $S^{FD}_{j\ell} = \int_{F_{j\ell}} p(x)dx = \int_{F_{j\ell}} d\mathbb{P}(x)$.

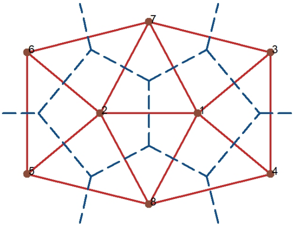


Edge Weight: Face Density Estimation

- 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 $\hat{S}^{FD}_{j\ell}$ is defined as

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

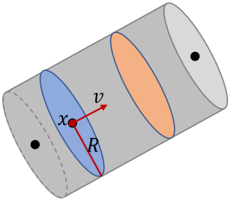
- This is 1-D KDE.



Edge Weight: Tube Density (TD)

- Similar to face density but has a predefined regular shape.
- Define a disk area centered at x with radius R and normal direction ν as

$$\text{Disk}(x, R, \nu) = \{y : ||x - y||_2 \leq R, (x - y)^T \nu = 0\}$$



- Parameterize the central line through c_j, c_ℓ as $\{c_j + t(c_\ell - c_j) : t \in [0, 1]\}$.
- Examine the integrated density within the disks along the central line.

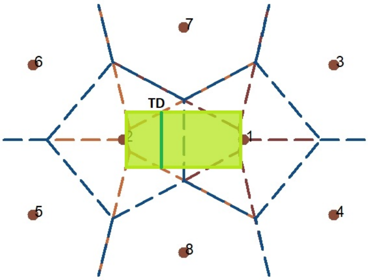
Edge Weight: Tube Density (TD)

Define the integrated density (called disk density) in the disk region as

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

Tube density (TD) is the minimal disk density along the central line, i.e.,

$$S_{j\ell}^{TD} = \inf_{t\in[0,1]} \mathrm{pDisk}_{j\ell,R}(t). \quad (2)$$



Edge Weight: Tube Density Estimation

- Similar to the FD, estimate the TD by projected KDE.
- $\Pi_{j\ell}(x)$ be the projection of a point x on the line through c_j, c_ℓ.$\Pi_{j\ell}(x)$ be the projection of a point x on the line through c_j, c_ℓ.
- Estimate the pDisk via

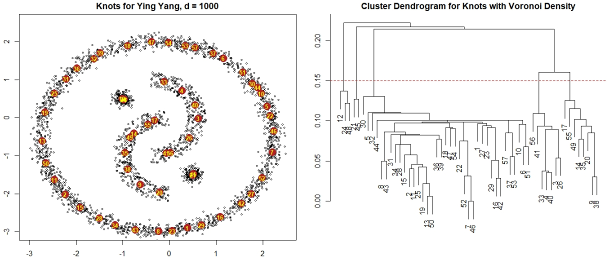
$$\widehat{\text{pDisk}}_{j\ell,R}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\Pi_{j\ell}(X_i) - c_j - t(c_\ell - c_j)}{h}\right) I(||X_i - \Pi_{j\ell}(X_i)|| \le R)$$

- Estimate the TD as

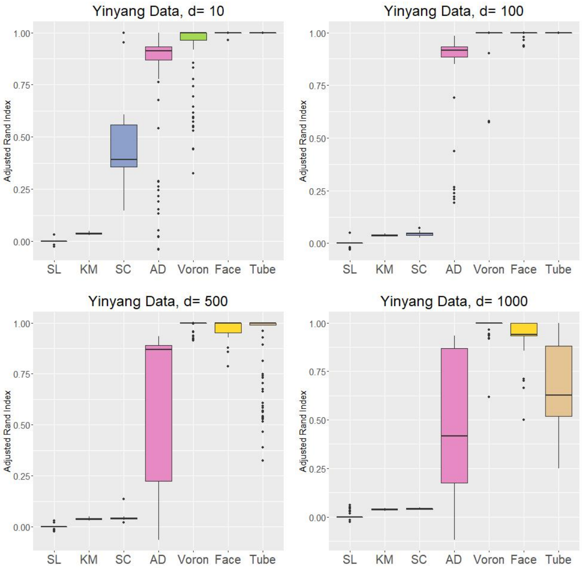
$$\hat{S}_{j\ell}^{TD} = \min_{t\in[0,1]} \widehat{\text{pDisk}}_{j\ell,R}(t). \tag{3}$$

Simulation: Yinyang Data

- Sample size $n = 3200$ ($k = 57 \approx \sqrt{3200}$)
- Increase the dimension of noise variables to make dimensions $d = 10, 100, 500, 1000$.

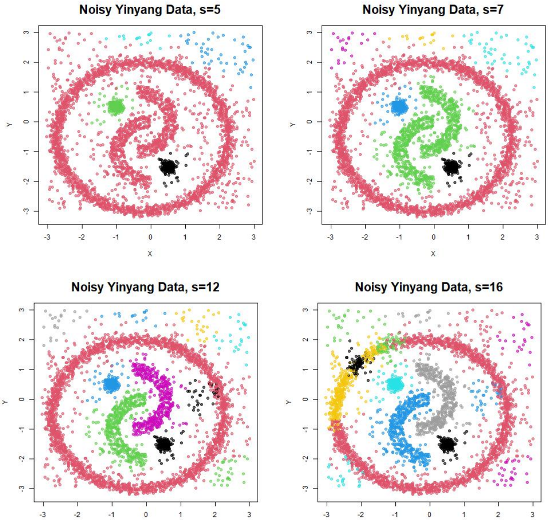


Yinyang Data Clustering Performance

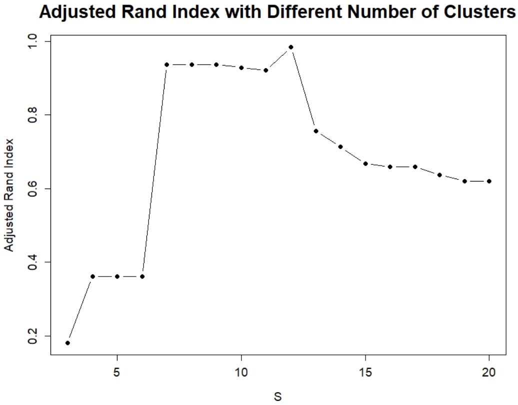


Data with Noise

- Added 640 (20% of the true signals) noisy points to the Yinyang dataset ($d = 1000$)
- Use Voronoi density and apply single linkage for knot segmentation.

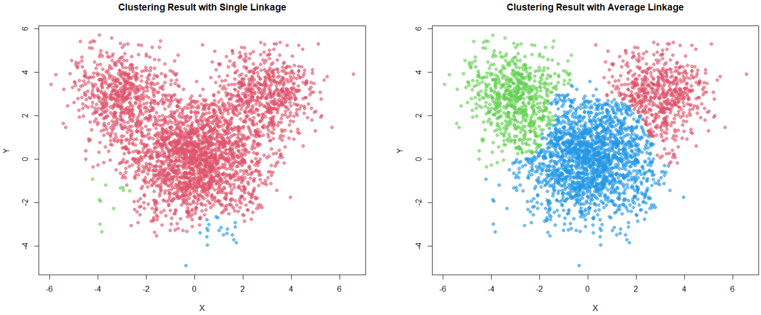


Data with Noise



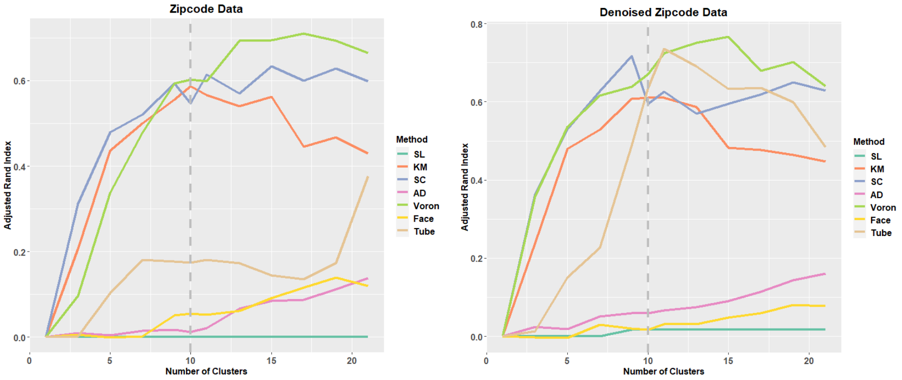
Overlapping Clusters

- Add additional noises to make the three structures overlap
- Using Single linkage for knots segmentation fails to discover the true structure.
- Using average linkage recovers the underlying three components.



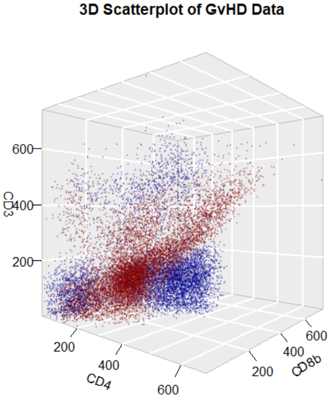
Zipcode Data

- 2000 16×16 images of handwritten Hindu-Arabic numerals from (Stuetzle and Nugent, 2010).
- 'denoised' data: Estimate the density of each observation by $\sqrt{n}$-NN density estimator and remove 10% observation with the lowest density.

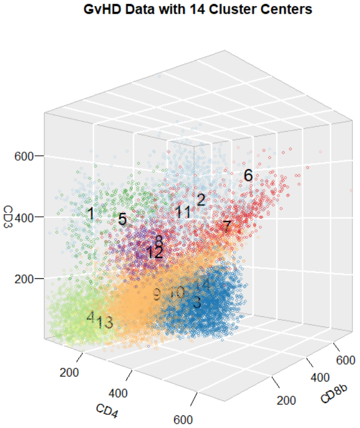


GvHD Data

- Flow cytometry data from (Brinkman et al., 2007)
- 9083 observations from a patient with graft-versus-host disease (GvHD) and 6809 observations from a control patient.
- 4 biomarker variables, CD4, CD8β, CD3, and CD8.
- Previous studies (Brinkman et al., 2007; Baudry et al., 2010) identified high values of CD3, CD4, CD8β cell sub-populations in the GvHD positive sample.



GvHD Data



GvHD Data with 14 Cluster Centers
CD3
600
400
200
CD4
200
400
600
CD8b
200
400
600
1
2
3
4
5
6
7
8
9
10
11
12
13
14

GvHD Data

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Size	202	948	3881	1859	338	17	812	468	6191	251	37	478	402	8
Prop	.458	.343	.008	.296	.341	.000	.934	.690	.888	.673	.669	.794	.841	.310
p-value	.32	1e-19	0	8e-63	6e-08	1e-04	3e-102	3e-13	0	1e-06	.11	2e-29	8e-33	.52

Conclusion

- **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**

Some possible future directions:

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

Reference

J.-P. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2):332–353, 2010. doi: 10.1198/jcgs.2010.08111.

J. L. Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, Sept. 1975. ISSN 0001-0782. doi: 10.1145/361002.361007.

R. R. Brinkman, M. Gasparetto, S. J. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith. High-Content Flow Cytometry and Temporal Data Analysis for Defining a Cellular Signature of Graft-Versus-Host Disease. *Biology of Blood and Marrow Transplantation*, 13(6):691–700, jun 2007. ISSN 10838791. doi: 10.1016/j.bbmt.2007.02.002.

B. Delaunay. Sur la sphère vide. a la mémoire de georges voronoï. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, 6: 793–800, 1934.

A. L. N. Fred and A. K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005. doi: 10.1109/TPAMI.2005.113.

R. Maitra. Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(1):144–157, 2009. doi: 10.1109/TCBB.2007.70244.

Thanks for listening!

Robustness to Number of Knots