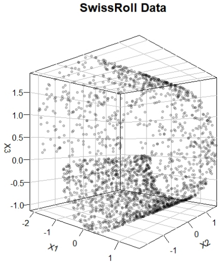
Skeleton Framework for Manifold Learning Tasks

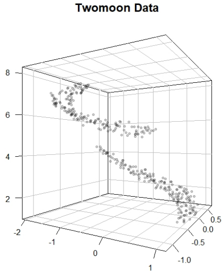
Jerry Wei
Department of Statistics, University of Washington
and
Yen-Chi Chen
Department of Statistics, University of Washington

Outline

Background

Many data nowadays have a geometric structure that the input data lies on a low dimensional manifold embedded inside the large-dimensional vector space.

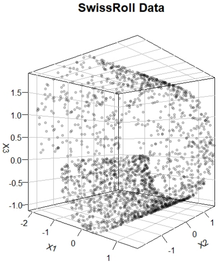


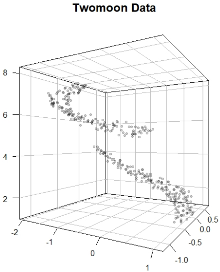


For various data analysis tasks to perform well, we need to understand such manifold structures of the data.

Background

Many data nowadays have a geometric structure that the input data lies on a low dimensional manifold embedded inside the large-dimensional vector space.

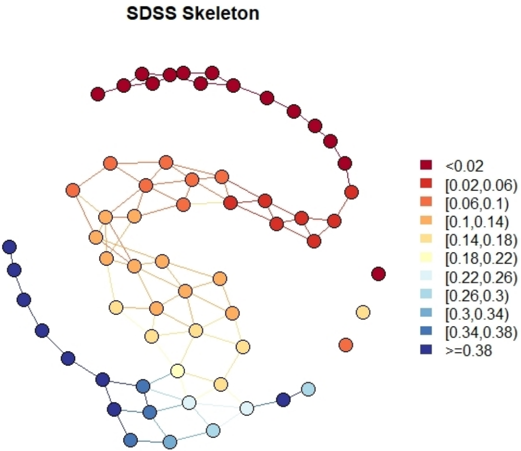




For various data analysis tasks to perform well, we need to understand such manifold structures of the data.

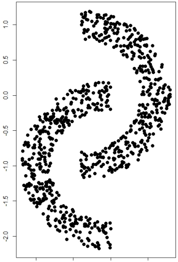
Our line of work propose to use a graph, called *Skeleton*, to summarize the manifold structure and assist various manifold learning tasks.

Example of Skeleton Representation

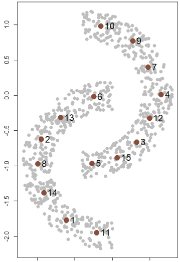


Sloan Digital Sky Survey (SDSS) data with 5 covariates measuring apparent magnitude of stars from images taken using 5 photometric filters. Response is the true redshift.

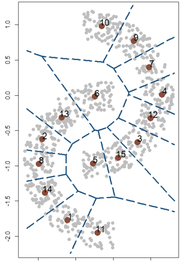
Skeleton Clustering



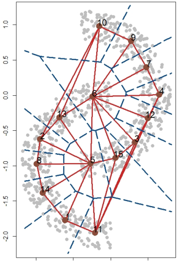
(a) Data



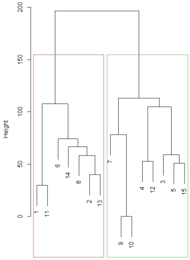
(b) Knots



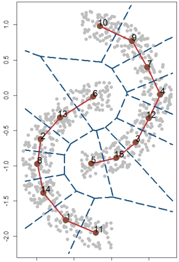
(c) Voronoi Cells



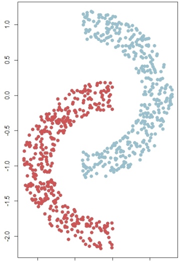
(d) Skeleton



(e) Dendrogram



(f) Segmentation



(g) Clustering

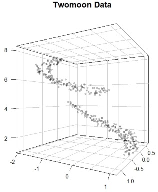
Skeleton Clustering

Algorithm Skeleton Clustering

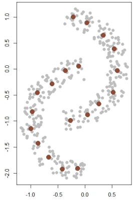
Input: Observations $X_1, \cdots, X_n$, number of knots k

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 density-based similarity weights to each edge using Voronoi density (also Face density, 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.

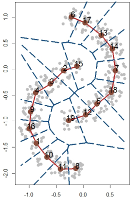
Skeleton Regression Framework



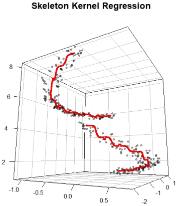
(a) Data



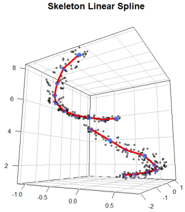
(b) Knots



(c) Skeleton



(d) S-Kernel Regression



(e) Linear Interpolation

Figure: Skeleton Regression illustrated by Two Moon Data (d=2).

Our Approach: Skeleton Regression Framework

Algorithm Skeleton Regression

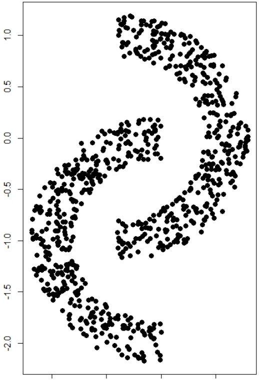
Input: Observations $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_N, Y_N)$.

1. **Skeleton Construction.** Construct a skeleton representation of the input space. Knots and edges can be tuned with subject knowledge.
2. **Data Projection.** Project the input vectors onto the skeleton structure.
3. **Skeleton Regression Function Estimation.** Fitting nonparametric regression functions on the skeleton using kernel regression, linear interpolation, or additional methods
4. **Prediction.** Project the feature vectors of new data onto the learnt skeleton structure and use the estimated regression function for prediction.

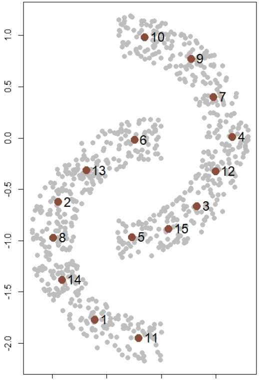
Skeleton Construction

Knots Construction

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



(a) Data



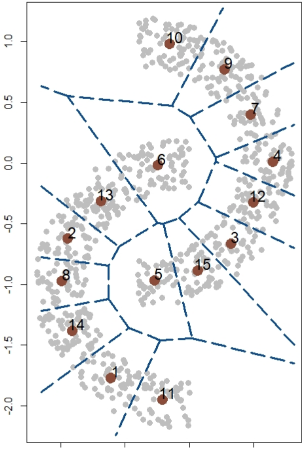
(b) Knots

Edge Construction, Voronoi Cells

The Voronoi cell (**?**), $\mathbb{C}_j$, 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,

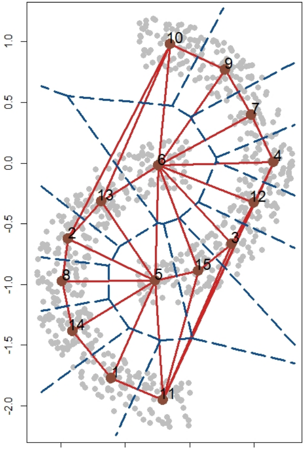
$$\mathbb{C}_j = \{x \in \mathcal{X} : d(x, c_j) \leq d(x, c_\ell) \ \ \forall l \neq j\},$$

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



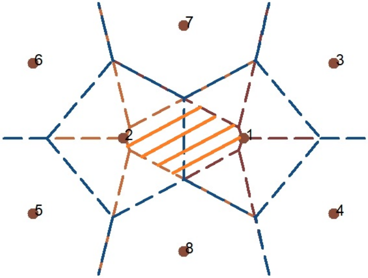
Edge Construction, Delaunay Triangulation

- 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 $DT(\mathcal{C})$ (**?**) of knots $c_1, \cdots, c_k$



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 $S_{j\ell}^{VD} = \frac{\mathbb{P}(A_{j\ell})}{\|c_j - c_\ell\|}$.

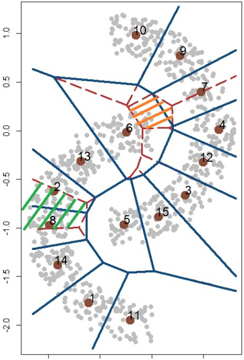


Edge Weight: Voronoi Density Estimation

- 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

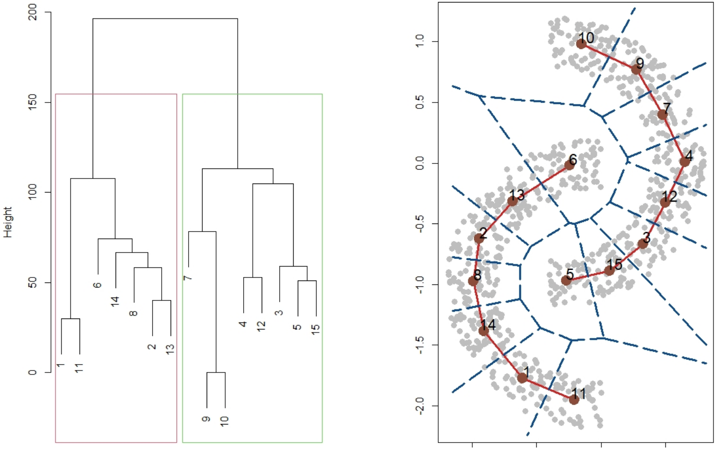
$$\hat{S}_{j\ell}^{VD} = \frac{\hat{P}_n(A_{j\ell})}{\|c_j - c_\ell\|}. \tag{1}$$

- Essentially counting points in the 2-NN region, which can be computed fast by k-d tree algorithm
- Effect of dimension small



Skeleton Segmentation

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



Tasks based on Skeleton

Clustering: Assign cluster membership according to its nearest knot.
Regression:

- Skeleton-based Kernel Regression
- Skeleton-based Linear Spline
- Higher-order splines

Thanks for listening!