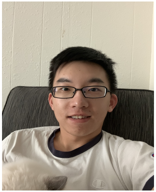
Large sample spectral analysis of graph-based multi-manifold clustering

Nicolas Garcia Trillos[1]

Pengfei He[2]



Chenghui Li[1]

[1]University of Wisconsin-Madison

[2]Michigan State University

Names are ordered alphabetically

October 29, 2023

Multi-manifold clustering

Setup

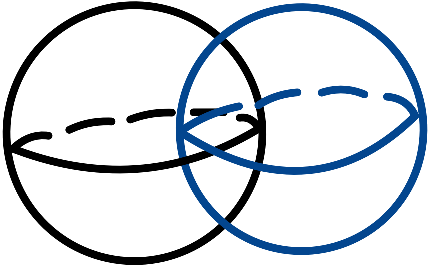
1. $\{\mathcal{M}_l\}_{l=1}^N$: a collection of manifolds that cannot be tangential to each other.
2. m_l: dimension of manifold $\mathcal{M}_l$.
3. $m = \max_{l=1,\dots,N}\{m_l\}$.
4. $\mathcal{M} := \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_N$.

Let $X = \{x_1, \dots, x_n\}$ be i.i.d. samples from a distribution μ on $\mathcal{M}$ of the form:

$$d\mu = \sum_{l=1}^{N} w_l \rho_l(x) d\mathrm{vol}_{\mathcal{M}_l}(x), \quad \text{where } w_l > 0, \quad \sum_{l=1}^{N} w_l = 1. \quad (1)$$

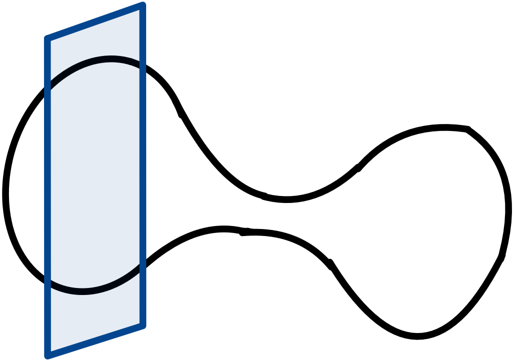
Multi-manifold clustering (MMC)

Some Examples



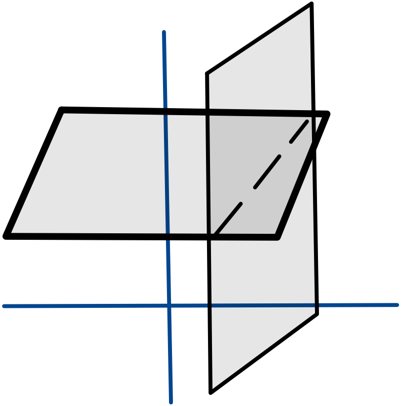
Multi-manifold clustering (MMC)

Some Examples



Multi-manifold clustering (MMC)

Some Examples



Spectral embedding and spectral clustering

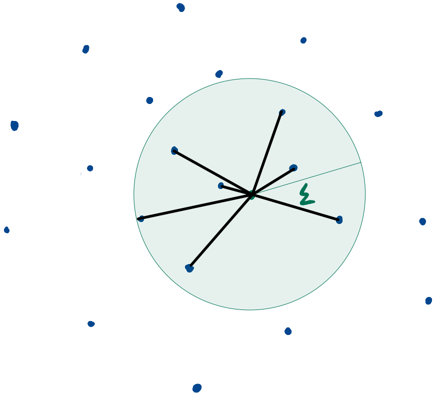
Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.
Output: Spectral Embedding $u_1, \ldots, u_k$ of S; Clusters $A_1, \ldots, A_k$ with $A_i = \{j \mid y_j \in C_i\}$

- Compute degree diagonal matrix $D = diag(\sum_j S_{ij})$ where $\sum_j S_{ij}$ is the degree of i-th node.
- Compute the Laplacian $L = D - S$.
- Compute the first k eigenvectors $u_1, \ldots, u_k$ of the eigenproblem $Lu = \lambda u$.

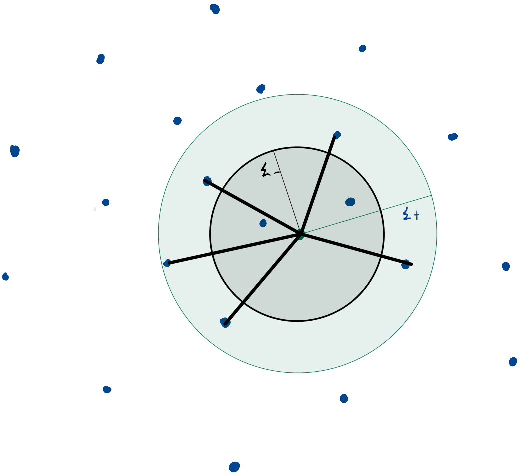
- Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, \ldots, u_k$ as columns.
- For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i -th row of U.
- Cluster the points $(y_i)_{i=1,\ldots,n}$ in $\mathbb{R}^k$ with the k -means algorithm into clusters $C_1, \ldots, C_k$.

Traditional way to pick S when data is in Euclidean space.

- Usual assumption: data clusters are somewhat separable.
- Often, use $\varepsilon - graph$ or kNN to build the similarity graph.



$(\varepsilon_+, \varepsilon_-)$-graph



How spectral clustering with ε-graph works

Figure: ε-graph

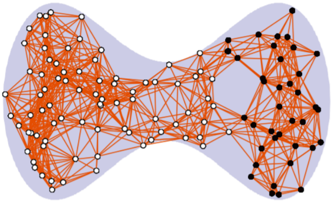


Figure: Run Spectral Clustering for 2 clusters

Multi-manifold clustering

For the following data set, a good **multi-manifold clustering(MMC)** algorithm must identify the two underlying overlapping spheres.

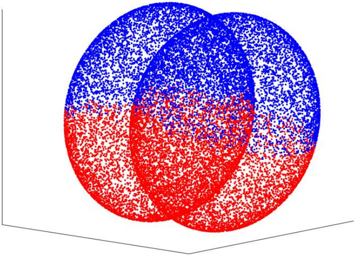


Figure: Spectral clustering with ε-graph

Multi-manifold clustering

A general framework

As soon as **full inner connectivity** and **sparse outer connectivity** are satisfied for S, spectral clustering solves MMC when $n \to \infty$.

A general framework

Points on the same manifold *should* connect.

Definition (Full Inner Connectivity)

With probability $1 - C_1(n)$, where $C_1(n) \to 0$ as $n \to \infty$, for any pair of points x_i, x_j belonging to the same manifold $\mathcal{M}_k$ we have $\omega_{x_i,x_j} = \omega_{x_i,x_j}^{\varepsilon_+,\varepsilon_-}$.

A general framework

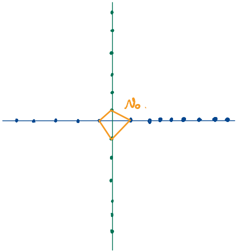
The number of between-different-manifolds connections cannot grow too quickly.

Definition (Sparse Outer Connectivity)

Let N_{sl} be the number of $x_i \in \mathcal{M}_s$ and $x_j \in \mathcal{M}_l$ such that $\omega_{ij} > 0$, and let

$$N_0 := \max_{l \neq s}\{N_{ls}\}.$$

Then, with probability one, $\frac{N_0}{n^2(\varepsilon_+^{m+2} - \varepsilon_-^{m+2})} \to 0$ as $n \to \infty$.



$$m_1 = m_2 = \cdots = m$$

Main Result

Given a graph construction satisfies full inner connectivity and sparse outer connectivity. Under some mild assumptions, with high probability, the eigenvalues and eigenvectors of the graph Laplacian converge to the eigenvalues and eigenfunctions of weighted Laplace Beltrami operator on $\mathcal{M}$.

Remark

The first N eigenfunctions of the weighted Laplace Beltrami operator is the linear combination of the indicator functions of individual manifolds.

When dimensions are different

Unfortune of Unnormalized Graph Laplacian

When $k = N$, different manifolds can still be recovered. If $k > N$, only the highest dimensional manifolds will be separated.

When dimensions are different

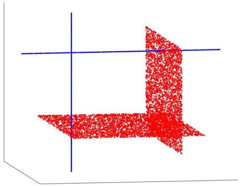


Figure: 2 Clusters

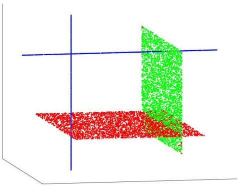


Figure: 3 Clusters

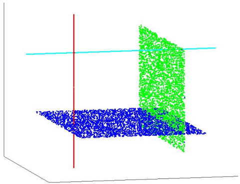


Figure: 4 Clusters

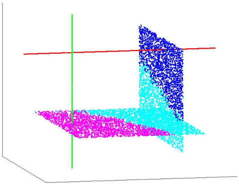
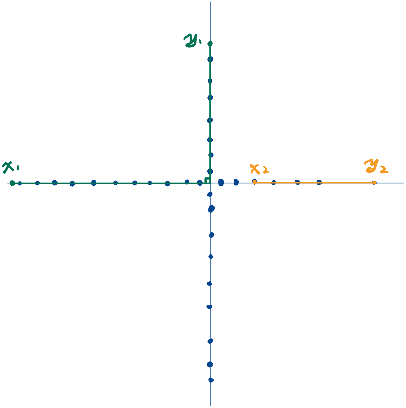


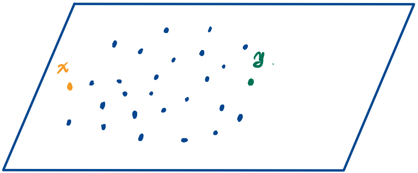
Figure: 5 Clusters

Path-based Algorithm

Intuition: there is a smooth path between points on the same manifold, but hard to have one when points are on different manifolds. Curvature information of the path matters.

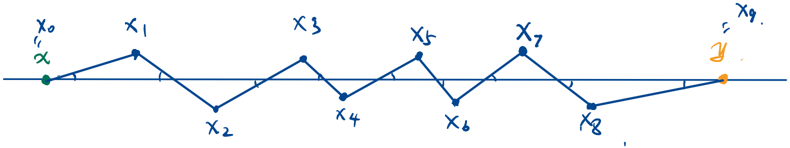


Path Algorithm



① $\varepsilon_- < |x-y| < \varepsilon_+$.

② There is a "smooth path" between x, y.



s.t. $|x_i - x_{i-1}| \le r$ and $\angle(\overrightarrow{xy}, \overrightarrow{x_i x_{i+1}}) \le \alpha$.

Multi-manifold clustering

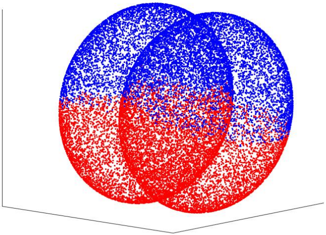


Figure: Spectral clustering with ε-graph

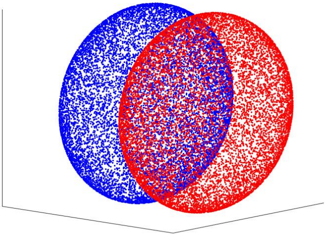


Figure: Spectral clustering with path algorithm with $\varepsilon_+, \varepsilon_-$-graph

Figure: *

For the data set illustrated above, a good multi-manifold clustering algorithm must identify the two underlying overlapping spheres.

Annular Graph Helps

It improves the rate of outer connectivity, but does not harm inner connectivity.

How $(\varepsilon_+, \varepsilon_-)$-graph helps in multi-clustering problem

When $\varepsilon_- \sim \varepsilon_+$, it will not affect full inner connectivity too much but improve sparse outer connectivity rate. Also, only calculating $(\varepsilon_+, \varepsilon_-)$-graph can help computationally.

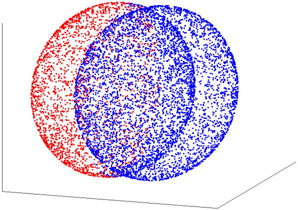


Figure: Annular proximity graph with angle constraint when $\varepsilon_- = 0$

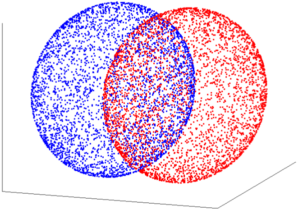


Figure: Annular proximity graph with angle constraint when $\varepsilon_- \sim \varepsilon_+$

Simulation

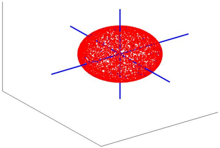


Figure: 2 clusters

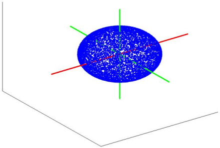


Figure: 3 clusters

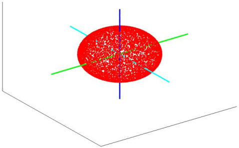


Figure: 4 clusters

MNIST

Algorithm	[0,1]	[0,2]	[0,3]	[0,4]	[0,5]	[0,6]	[0,7]	[0,8]	[0,9]
path	14.0%	5.6%	1.9%	1.8%	2.6%	7.7%	46.4%	9.7%	1.9%
local PCA	6.4%	25.9%	30.0%	45.5%	34.8%	34.5%	34.1%	26.6%	25.1%
SMCE	20.0%	25.5%	6.9%	9.2%	24.1%	12.1%	2.9%	17.8%	3.8%
SC	18.8%	12.8%	1.8%	2.2%	2.6%	10.0%	46.4%	11.8%	2.3%

Table: Misclustering rates for some subsets of MNIST

Takeaway

- Two sufficient conditions on graph that guarantee consistency for MMC.
- There is a specific graph construction satisfying sufficient conditions.
- $(\varepsilon_+, \varepsilon_-)$-graph improves the convergence rate.

Thank You

Project Page: https://github.com/chl781/manifold-clustering

Neurips Link: https://neurips.cc/virtual/2023/poster/73915

Question? `cli539@wisc.edu`