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

Nicolas Garcia Trillos, Pengfei He, Chenghui Li*

Abstract

In this work[1] we study statistical properties of graph-based algorithms for multi-manifold clustering (MMC). In MMC the goal is to retrieve the multi-manifold structure underlying a given Euclidean data set when this one is assumed to be obtained by sampling a distribution on a union of manifolds $\mathcal{M} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_N$ that may intersect with each other and that may have different dimensions. We investigate sufficient conditions that similarity graphs on data sets must satisfy in order for their corresponding graph Laplacians to capture the right geometric information to solve the MMC problem. Precisely, we provide high probability error bounds for the spectral approximation of a tensorized Laplacian on $\mathcal{M}$ with a suitable graph Laplacian built from the observations; the recovered tensorized Laplacian contains all geometric information of all the individual underlying manifolds. We provide an example of a family of similarity graphs, which we call annular proximity graphs with angle constraints, satisfying these sufficient conditions. We contrast our family of graphs with other constructions in the literature based on the alignment of tangent planes. Extensive numerical experiments expand the insights that our theory provides on the MMC problem.

Keywords: multi-manifold clustering, graph Laplacian, spectral convergence, manifold learning, discrete to continuum limit.

1. Introduction

In this work we study the problem of *multi-manifold clustering* (MMC) from the perspective of spectral geometry. Multi-manifold clustering is the task of identifying the structure of multiple manifolds that underlie an observed data set $X = \{x_1, \dots, x_n\}$, its main challenge being that in general the underlying manifolds may be non-linear, may intersect with each other, and may have different dimensions (see Figures 1-3 for some illustrations). While spectral methods for learning have been analyzed by several authors throughout the past two decades in settings as varied as unsupervised, semi-supervised, and supervised learning, less is known about their theoretical guarantees for the specific multi-manifold clustering problem. We analyze MMC algorithms that are based on the construction of suitable similarity graph representations for the data and in turn on the spectra of their associated graph Laplacians. We provide statistical error guarantees for the identification of the underlying manifolds as well as for the recovery of their individual geometry.

As for most spectral approaches to clustering, we are interested in studying spectral properties of graph Laplacian operators of the form

$$\Delta_n u(x_i) := \sum \omega_{ij}(u(x_i) - u(x_j)), \quad x_i \in X. \tag{1.1}$$

Here, the ω_{ij} are appropriately defined symmetric weights that in general depend on the proximity of points x_i, x_j, and importantly, on a mechanism that detects when points

*. Corrsponding Author:
Email Address: cli539@wisc.edu(Chenghui Li)

1. Submitted to JMLR, preprint see https://arxiv.org/abs/2107.13610

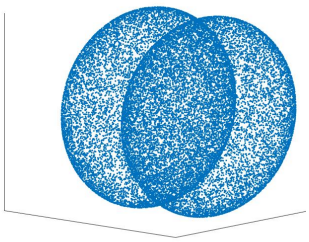

Figure 1

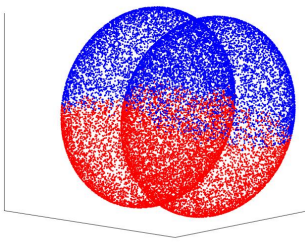

Figure 2

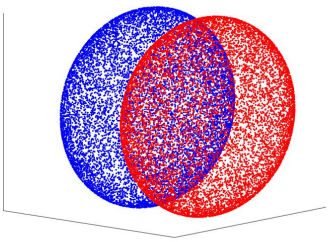

Figure 3

Figure 1 illustrates two intersecting ellipsoids (two dimensional). A *good* multi-manifold clustering algorithm must identify the two underlying ellipsoids. Figure 2 and Figure 3 show the spectral clustering with standard ε-proximity graph and annular proximity graph with angle constraint, respectively; see following discussion.

belong to different manifolds even if lying close to each other. Once the graph Laplacian is constructed we follow the spectral clustering algorithm: the first p eigenvectors of Δ_n (denoted $\psi_1, \ldots, \psi_N$) are used to build an embedding of the data set X into $\mathbb{R}^p$:

$$x_i \in X \longmapsto \begin{pmatrix} \psi_1(x_i) \\ \vdots \\ \psi_N(x_i) \end{pmatrix} \in \mathbb{R}^p.$$

In turn, with the aid of a simple clustering algorithm such as k-means the embedded data set is clustered. A successful algorithm will produce clusters that are in agreement with the different manifolds underlying the data set. It is essential to select the weight ω_{ij} to make spectral clustering algorithm work well as standard proximity graphs do not work well in MMC problem. See Figure 1-3 as an illustration.

2. Setup

To start making our results more precise, let us suppose that the data set X is obtained by sampling a distribution μ supported on a set $\mathcal{M}$ of the form $\mathcal{M} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_N$ where the $\mathcal{M}_l$ are smooth compact connected manifolds with no boundary that for the moment are assumed to have the same dimension m; the manifolds $\mathcal{M}_l$ may have nonempty pairwise intersections, but these are assumed to have measure zero relative to the volume forms of each of the manifolds involved. The distribution μ is assumed to be a mixture model taking the form

$$\mu = w_1\rho_1 d\mathrm{vol}_{\mathcal{M}_1} + \cdots + w_N\rho_N d\mathrm{vol}_{\mathcal{M}_N} \tag{2.1}$$

for smooth density functions $\rho_l : \mathcal{M}_l \to \mathbb{R}$ and positive weights w_i that add to one; henceforth we use $d\mathrm{vol}_{\mathcal{M}_l}$ to denote the integration with respect to the Riemannian volume form associated to $\mathcal{M}_l$. A *tensorized Laplacian* $\Delta_{\mathcal{M}}$ acting on functions f on $\mathcal{M}$ (which will

be written as $f = (f_1, \ldots, f_N)$, where $f_l : \mathcal{M}_l \to \mathbb{R}$) can be defined according to

$$\Delta_{\mathcal{M}} f := (w_1 \Delta_{\mathcal{M}_1} f_1, \ldots, w_N \Delta_{\mathcal{M}_N} f_N), \tag{2.2}$$

where $\Delta_{\mathcal{M}_l}$ is a Laplacian operator mapping regular enough functions $f_l : \mathcal{M}_l \to \mathbb{R}$ into functions $\Delta_{\mathcal{M}} f_l : \mathcal{M}_l \to \mathbb{R}$ according to

$$\Delta_{\mathcal{M}_l} f_l = -\frac{1}{\rho_l} \mathrm{div}_{\mathcal{M}_l} \left(\rho_l^2 \nabla_{\mathcal{M}_l} f_l \right).$$

In other words, the operator $\Delta_{\mathcal{M}}$ acts in a coordinatewise fashion effectively treating each manifold $\mathcal{M}_i$ independently. It is then straightforward to show that eigenfunctions of $\Delta_{\mathcal{M}}$ are spanned by functions of the form $(0, \ldots, f_l, \ldots, 0)$ for some l, where f_l is an eigenfunction of $\Delta_{\mathcal{M}_l}$. This means that the spectrum of $\Delta_{\mathcal{M}}$ splits the geometries of the $\mathcal{M}_l$, and in particular, the different $\mathcal{M}_l$ can be detected by retrieving the eigenfunctions with zero eigenvalue.

3. Result

Let's begin with two sufficient conditions that the weighted graph must satisfy to solve MMC problem.

Definition 3.1 (Fully inner Connected graphs)**.** *Let $X = x_1, \ldots, x_n$ be samples from μ as defined in* (2.1)*. A weighted graph (X, ω) is said to be fully inner connected relative to ε_+ and ε_- which converge to zeros as $n \to \infty$ if 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 $\omega_{x_i,x_j} = \omega^{\varepsilon_+,\varepsilon_-}_{x_i,x_j}$.*

Definition 3.2 (Sparsely Outer Connected graphs)**.** *Let $X = x_1, \ldots, x_n$ be samples from μ as defined in* (2.1)*, and let (X, ω) be a weighted graph. Let N_{sl} be the number of connections between $x_i \in \mathcal{M}_s$ and $x_j \in \mathcal{M}_l$ such that $\omega_{ij} > 0$, and let*

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

The graph is said to be sparsely outer connected relative to ε_+ and ε_- converging to zero as $n \to \infty$ if with probability one, $\frac{N_0}{n^2(\varepsilon_+^{m+2} - \varepsilon_-^{m+2})} \to 0$ as $n \to \infty$. We recall that $m = \max_{l=1,\ldots,N} m_l$.

Full inner connectivity condition guarantees that points within one manifold connect to each other with high probability, and sparse outer connectivity condition guarantees that the number of connections between points from different manifolds cannot be too large.

Our first main results (**Theorem 2.5 and Theorem 2.7**) say that provided that the weights ω_{ij} defining the graph Laplacian operator Δ_n in (1.1) satisfy full inner connectivity and sparse outer connectivity, then the eigenvalues (appropriately scaled) and eigenvectors of Δ_n approximate the eigenvalues and eigenfunctions of the tensorized Laplacian $\Delta_{\mathcal{M}}$; we obtain high probability quantitative bounds for the error of this approximation. The bottom line is that our results imply that the spectral methods studied here are guaranteed, at least for large enough n, to recover the underlying multi-manifold structure of the data;

see Figure 3 for an illustration. Our work extends the growing literature of works that study the connection between graph Laplacians on data sets and their continuum analogues. This literature has mostly focused on the smooth setting where multiple intersecting manifolds are not allowed.

In our second main result (**Theorem 2.8**) we present some results for the case when the dimensions of the manifolds $\mathcal{M}_i$ do not agree. In this more general setting, the spectrum of the graph Laplacian Δ_n does not recover the tensorized geometry captured by $\Delta_{\mathcal{M}}$ as introduced earlier, but rather, only the tensorized geometry of the manifolds with the *largest* dimension, effectively quotienting out the geometric information of manifolds with dimension strictly smaller than the maximum dimension. Detailed discussions and proofs can be seen in Trillos et al. (2021).

The theory above showed that if two sufficient conditions are satisfied for the weighted graph, then spectral clustering is guaranteed to be consistent for MMC. In the following, We also present a graph construction that we refer to as annular proximity graph with angle constraint that satisfies full inner connectivity and sparse outer connectivity conditions. Numerical experiments support and expand our insights on the algorithm's behavior.

4. Contribution and Discussion

- We analyze graph Laplacians on families of proximity graphs when the nodes of the graphs are random data points that are supported on a union of unknown *intersecting* manifolds. The manifolds may all have *different* dimensions.

- We introduce two sufficient conditions that similarity graphs must satisfy in order to recover, from a graph Laplacian operator, the geometric information of the individual smooth manifolds underlying the data set. These conditions are referred to as *full inner connectivity* and *sparse outer connectivity*.

- We introduce and analyze *annular* proximity graphs and their effect on multi-manifold clustering. These are simple extensions of ε-proximity graphs that nonetheless can be shown to be, theoretically and numerically, better than the vanilla ε-graphs for multi-manifold clustering.

- We analyze a family of *annular proximity graphs with angle constraints*. This family is shown to satisfy the full inner connectivity and sparse outer connectivity conditions when their parameters are tuned appropriately. We contrast this construction with other constructions such as those based on local PCA which in general do not satisfy the full inner connectivity condition.

- Through numerical examples and some heuristic computations we provide further insights into the use of spectral methods for multi-manifold clustering.

References

Nicolas Garcia Trillos, Pengfei He, and Chenghui Li. Large sample spectral analysis of graph-based multi-manifold clustering, 2021.