Local Spectral Clustering of Density Upper Level Sets

Anonymous Authors[1]

Abstract

1. Introduction

Given data $\mathbf{x} := \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, our statistical learning task is clustering: splitting data into groups which satisfy some notion of within-group similarity and between-group difference.

In particular, spectral clustering methods are a family of powerful non-parametric clustering algorithms. Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$ with (i,j)th entry representing the similarity between data points x_i and x_j, we form the graph Laplacian matrix [1],

$$W := D^{-1}A; \quad L = I - W \tag{1}$$

where the degree matrix D is a diagonal matrix with $D_{ii} := \sum_j A_{ij}$.

Roughly speaking, spectral clustering techniques first embed the data $\mathbf{x}$ using the spectrum of the graph Laplacian matrix and subsequently use this *spectral embedding* to find a clustering of the data. When applied to large graphs (or large point clouds) classical global spectral methods can be computationally cumbersome and can be insensitive to the local geometry of the distribution of the samples (Mahoney et al., 2012; Leskovec et al., 2010). This in turn has led to the investigation of local spectral algorithms (Spielman & Teng, 2013; Andersen et al., 2006; Leskovec et al., 2010) which leverage locally-biased spectra computed using random walks around a user-specified seed node.

A natural model to consider when analyzing point cloud data is the following:

$$x_i \sim P, \text{ independently, for } i = 1, \ldots, n, \tag{2}$$

with f the density function of P with respect to the uniform measure over $\mathbb{R}^d$. In this case, we are interested in understanding what the output of a clustering algorithm on this finite sample reveals about the unknown density f. For $\lambda > 0$ and $\{x : f(x) \geq \lambda\}$, it is intuitive (Hartigan, 1981; Chaudhuri & Dasgupta, 2010) to define clusters to be the connected components $\mathbb{C}_f(\lambda)$ of the upper level set; we call these connected regions of high density *density clusters*, and study the ability of spectral methods to identify such clusters.

Graph connectivity criteria. A somewhat more standard mode of analyzing spectral clustering methods is through approximation to a pair of graph connectivity criteria.

For A as before, let the graph $G = (V, E)$, with vertices $V = (v_1, \ldots, v_n)$ corresponding to the n rows of A, and edges $E = \{(v_i, v_j) : 1 \leq i < j \leq n : (v_i, v_j) = A_{ij}\}$ (Since A is symmetric, G is an undirected graph; also, by convention, we preclude self-loops). There are many (Yang & Leskovec, 2015; Fortunato, 2010) graph-theoretic measures which assess the cluster quality of a subset $S \subseteq V$ (or more generally the quality of a partition $S_1 \cup \ldots \cup S_m = V$, for $m \geq 2$.)

Arguably a natural way to assess cluster quality is via a pair of criteria capturing the *external* and *internal connectivity* of S, respectively. As the names suggest, external connectivity should relate to the number of edges between S and G/S (hereafter denoted S^c), while internal connectivity in turn measures the number of edges between subsets within S. The clustering task then becomes to find a subset S (or, for global algorithms, a partition $S_1 \cup \ldots \cup S_m = V$), which has both small external and large internal connectivity.

We will assess the external connectivity of a subset $S \subseteq V$ through its normalized cut. The cut of S is

$$\text{cut}(S; G) := \sum_{u \in S} \sum_{v \in S^c} \mathbf{1}\left((u,v) \in E\right)$$

– where $S^c = V/S$ is the complement of S in V – and the volume of S is

$$\text{vol}(S; G) := \sum_{u \in S} \sum_{v \in V} \mathbf{1}\left((u,v) \in E\right).$$

Then, the *normalized cut* of S is given by

$$\Phi(S; G) := \frac{\text{cut}(S; G)}{\min\left\{\text{vol}(S; G), \text{vol}(S^c; G)\right\}} \tag{3}$$

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] Often, either of the Laplacian matrices $L_{sym} := D^{\frac{1}{2}} W D^{-\frac{1}{2}}$ or $L_{unn} := D - A$ are used instead.

Intuitively, a set with low *normalized cut* has many more edges which do not cross the cut than edges which do cross the cut.

Given $S \subseteq V$, the subgraph induced by S is given by $G[S] = (S, E_S)$, where $(u,v) \in E_S$ if both u and v are in S and $(u,v) \in E_S$. Introduce and define the **inverse mixing time**.

$$\Psi(S;G) = \tag{4}$$

If S has normalized cut no greater than Φ, and inverse mixing time no less than Ψ, we will refer to it as a (Φ,Ψ)-cluster. Both local (Zhu et al., 2013) and global (Kannan et al., 2004) spectral algorithms have been shown to output clusters (or partitions) which provably satisfy approximations to the optimal (Φ,Ψ)-cluster (or partition), where the optimization is carried out over the graph G. [2]

Personalized PageRank. As mentioned previously, global algorithms which find spectral cuts may be computationally infeasible for large graphs; in this setting, local algorithms may be preferred or even required. We will restrict our attention in particular to one such popular algorithm: *personalized PageRank* (PPR). The personalized PageRank algorithm was first introduced by (Haveliwala, 2003) and variants of this algorithm have been studied further in several recent works (Spielman & Teng, 2011; 2014; Zhu et al., 2013; Andersen et al., 2006; Mahoney et al., 2012).

The random walk matrix is given by W defined as in (1), and vertices V and edges E are as above. The PPR vector is defined with respect to the following inputs: a user-specified seed node $v_i \in V$, and $\alpha \in [0,1]$ a teleportation parameter. Letting $v = v_i$ for notational simplicity, and e_v be the indicator vector for the vth node (meaning e_v has a 1 in the ith location and 0 everywhere else), the *PPR vector* is given by the recursive formulation

$$p(v,\alpha;G) := \alpha e_v + (1-\alpha) p(v,\alpha;G) W \tag{5}$$

We note in passing that, for $\alpha > 0$ the vector $p(v,\alpha;G)$ can be well-approximated by a simple local computation (of a random walk with restarts at the node v.) We also point out that, from a density clustering standpoint, since density clusters are inherently local, using the PPR algorithm eases the analysis, and as we will observe in the sequel our analysis requires fewer global regularity conditions relative to more classical global spectral algorithms.

To compute a cluster $\widehat{C}_n \subset V$ using the PPR vector, we will take sweep cuts of $p(r,\alpha;G)$. For a vector p, $p[j]$ be the jth entry of p, and $p_{(k)}$ be the kth *largest* entry of p. Then the sweep cut S_k is

$$S_k = \{u_j : p(r,\alpha;G)[j] > p(r,\alpha;G)_{(k)}\} \tag{6}$$

We delay formal introduction of the local clustering algorithm we analyze until we have given defined a method for forming a graph over the data $\mathbf{x}$.

Large sample behavior. Given a kernel function $\mathbf{k} : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ of the form $\mathbf{k}(x,x') = k(\|x - x'\| / r_n)$ with k non-increasing and some $r_n > 0$, and data $\mathbf{x} = \{x_1, \ldots, x_n\}$ sampled from P as before, form the (weighted, complete) similarity graph $G_n = (\mathbf{x}, E_n)$ with $E_n = \{k(x_i, x_j) : 1 \leq i < j \leq n\}$. (Here, $\|\cdot\|$ is used to denote Euclidean norm).

It is worth pointing out that in this context, some theory has been developed regarding how graph theoretic quantities such as the normalized cut Φ (and others) relate to properties of the underlying distribution f as well as the kernel function $\mathbf{k}$. Such analyses typically proceed by defining a continuous analogue to the measure of cluster quality under consideration. Then, under appropriate specification of $\mathbf{k}$ and a proper schedule of $\{r_n\}_{n \in N}$, convergence of clusters output by spectral (and other) algorithms to the corresponding minima of these continuous analogues has been shown (von Luxburg et al., 2008; Trillos & Slepcev, 2018).

These continuous analogues, and their corresponding minimizers, are not always easily identifiable for the particular density function under consideration. Relating these partitions to the arguably more simply defined high density clusters can be also challenging in general. Intuitively, however, under the right conditions such high-density clusters should have more edges within themselves than to the remainder of the graph. We formalize this intuition next.

1.1. Summary of results

Hereafter, we consider the *uniform kernel function* for a fixed $r > 0$,

$$\mathbf{k}(x,x') = \mathbf{1}(\|x - x'\| \leq r) \tag{7}$$

and the associated *neighborhood graph*

$$G_{n,r} = (\mathbf{x}, E_{n,r}),\ (x_i, x_j) \in E_{n,r} \text{ if } \mathbf{k}(x_i, x_j) = 1 \tag{8}$$

For a given high density cluster $C \subseteq \mathbb{C}_f(\lambda)$, we call $C[\mathbf{x}] = C \cap \mathbf{x}$ the *empirical density cluster*. We now introduce a notion of consistency for the task of density cluster estimation:

Definition 1 (Consistent density cluster estimation)**.** *For an estimator $\widehat{C}_n$, and any $C, C' \in C_f(\lambda)$, we say $\widehat{C}_n$ is a*

[2] In the case of (Kannan et al., 2004), the internal connectivity parameter ϕ is actually the conductance, i.e. the minimum normalized cut within the subgraph $G[S]$. See Theorem 3.1 for details; however, note that $\phi^2 / \log(\text{vol}(S)) \leq O(\Psi)$, and so the lower bound on ϕ translates to a lower bound on Ψ.

consistent estimator of C if the following statement holds: as the sample size $n \to \infty$,

$$C[\mathbf{x}] \subseteq \widehat{C}_n, \text{ and } \widehat{C}_n \cap C'[\mathbf{x}] = \emptyset \tag{9}$$

occurs with probability tending to 1.

Our results can now be summarized by the following two points:

1. Under a natural set of geometric conditions[3], Theorems 1 and Theorem 2 upper and lower bound, respectively, the normalized cut and inverse mixing time of an empirical density cluster $C[\mathbf{x}]$.

2. We show these bounds on the graph connectivity criteria have algorithmic consequences for the PPR algorithm. An immediate consequence of Theorems 1 and Theorem 2, along with the previous work of (Zhu et al., 2013), is to yield an upper bound on the normalized cut of the set $\widehat{C}_n$ output by Algorithm **??**, as well as upper bounding the symmetric set difference between $\widehat{C}_n$ and $C[\mathbf{x}]$. Further, in Section 4 we show that a careful analysis of the form typical to local clustering algorithms yields Theorem 4, which states that Algorithm **??**, properly initialized, performs consistent density cluster estimation.

Organization. In Section 4, we provide some example density functions, to clarify the relevance of our results. In Section 5, we show empirical performance of the PPR algorithm, which demonstrates that violations of the geometric conditions we set out in Section 2 manifestly impact density cluster recovery (i.e. the conditions are not superfluous), before concluding in Section 6. First, however, we summarize some related work.

1.2. Related Work

Graph notation. Let $G = (V, E)$ be an undirected, unweighted graph, with $S, S' \subseteq V$. The volume of S is given by $\mathrm{vol}(S; G) = \sum_{v \in S} \deg(v; G)$ where $\deg(v; G) = \sum_{u \in V} 1\big((u, v) \in E\big)$ is the degree of v in G. For the random walk over G, denote the stationary distribution by π, where $\pi(S; G) = \frac{\mathrm{vol}(S;G)}{\mathrm{vol}(V;G)}$.

For the sample $\mathbf{X} := \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and a set $A \subset \mathbb{R}^d$, define $A[\mathbf{X}] = A \cap \mathbf{X}$. We denote the uniform measure over $\mathbb{R}^d$ by $\nu(\cdot)$.

2. Measures of cluster quality

Let P be a distribution supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$, with continuous density function f (with respect to the uniform measure). The Euclidean distance is denoted by $\|\cdot\|$.

Density upper-level set. For a number $\tau \geq 0$, let $C_f(\tau)$ be the collection of collected components of the density upper-level set $\{x \in \mathcal{X} : f(x) \geq \tau\}$. Define a *$\tau$-density cluster* to be one such connected component $A \in C_f(\tau)$. We will sometimes refer to $A[\mathbf{X}]$ as an *empirical density cluster*.

Graph bicriteria. For $G = (V, E)$ an undirected, unweighted graph and $S, S' \subset V$ as before, let $|E(S, S'; G)|$ denote the cut of S and S', given by

$$|E(S, S'; G)| = \sum_{v \in S} \sum_{u \in S'} 1((v, u) \in E).$$

Define the balance $B(S; G)$ to be

$$B(S; G) = \min\{\mathrm{vol}(S; G), \mathrm{vol}(V \setminus S; G)\}.$$

We can now formally introduce our first criterion for assessing the quality of graph cuts: the **conductance**, Φ, given by

$$\Phi(S; G) \overset{\text{def}}{=} \frac{|E(S, V \setminus S; G)|}{B(S; G)}. \tag{10}$$

We typically seek a set $S^\star \subset V$ such that $\Phi(S^\star; G)$ is small.

Neighborhood graph. Given $r \geq 0$, define the neighborhood graph to be $G_{n,r} = (\mathbf{X}, E_n)$, where for $x_i, x_j \in \mathbf{X}$, $(x_i, x_j) \in E_n$ if $\|x_i - x_j\| \leq r$. (By convention, we do not allow loops, meaning $(x_i, x_i) \notin E_n$). For ease of notation, for $S \subset \mathbf{X}$, let $\Phi_{n,r}(S) := \Phi(S; G_{n,r})$.

2.1. Well-conditioned density clusters.

In order to satisfy the bicriteria, we must introduce some assumptions on the density f. Let $B(x, r)$ be a closed ball with respect to Euclidean distance) of radius r around the point x. Given a set $A \subset \mathbb{R}^d$, and a number $\sigma > 0$, define the set $A_\sigma = A + B(0, \sigma) = \big\{y \in \mathbb{R}^d : \inf_{x \in A} \|y - x\| \leq \sigma\big\}$.

(A1) *Minimum density:* For numbers $\sigma, \tau_\sigma > 0$, a set $A \subset \mathcal{X}$ is said to be (σ, τ_σ)-*regular* if

$$\inf_{x \in A_\sigma} f(x) = \tau_\sigma \tag{11}$$

(A2) *Low noise density:* For numbers $\gamma, \sigma > 0$, and a set $A \subset \mathcal{X}$, the density function f is said to be (γ, σ)- *low*

[3] We formally introduce the geometric conditions in 2. They preclude clusters which are too thin and long, or those for which the gap in density between the high density area and the outside is not sufficiently large

noise around A if there exists a constant C_1 such that for all $x \in \mathcal{X}$ with $0 < \rho(x, A_\sigma) \leq \sigma$,

$$\inf_{x' \in A_\sigma} f(x') - f(x) \geq C_1 \rho(x, A_\sigma)^\gamma.$$

where $\rho(x, A_\sigma) = \min_{x_0 \in A_\sigma} \|x - x_0\|$.

(A3) *Cluster separation:* For $\tau, \sigma > 0$, $C_f(\tau)$ the set of connected components of the density upper-level set is said to be σ*-well separated* if for all τ-density clusters $A, A' \in C_f(\tau)$,

$$\rho(A, A') > \sigma.$$

where $\rho(A, A') = \min_{x \in A, x' \in A'} \|x - x'\|$.

Assumptions are standard, provide references. We note that these assumptions are all local in nature, further motivating the study of a local algorithm. Assumptions (A1)-(A3) are used to upper bound bicriteria 1, whereas (A1) and (A4) are necessary to lower bound bicriteria 2.

Definition 2. *For $\tau > 0$ and $A \in C_f(\tau)$ a τ-density cluster, we say that A is a $(\lambda, \sigma, \gamma)$-well-conditioned cluster if A is (λ, σ)-regular, the density f is (γ, σ)-low noise around A^σ, and $C_f(\tau)$ is σ-well separated.*

We note that σ plays dual roles here, both in effect precluding arbitrarily narrow and long clusters in (A1) and arbitrarily flat densities in (A2). We give some examples of densities which satisfy these assumptions in Section 4.

3. Local Clustering on Density Level Sets

In this section we show that a well-conditioned cluster A satisfies the bicriteria criteria 1 and criteria 2, and therefore the PPR algorithm outputs a low conductance set which has small symmetric set difference with the empirical cluster.

Theorem 1. *For some $\tau > 0$, let a τ-density cluster $A \in C_f(\tau)$ satisfy Assumptions (A1)-(A3) for some $\sigma, \tau_\sigma, \gamma > 0$. Then, for any $r < \sigma$ and $\delta > 0$, the following statements hold with probability at least $1 - \delta$:*

- ***Additive error bound.*** *Fix $\epsilon > 0$. Then, for*

$$n \geq \log(2/\delta) \left(\frac{1 + \epsilon/2}{2\tau_\sigma^2 \nu(A_\sigma) \nu_d (r/2)^d} \right)^2 \tag{12}$$

we have

$$\frac{\Phi_{n,r}(A_\sigma[\mathbf{X}])}{r} \leq C_\sigma \frac{\tau}{\tau_\sigma} \frac{(\tau_\sigma - \frac{r^{\gamma+1}}{\gamma+1})}{\tau_\sigma} + \epsilon \tag{13}$$

- ***Multiplicative error bound.***

where $C_0 = 2^{2d+1}d$ and $C_\sigma = C_0/\sigma$.

A few remarks are in order.

Remark 1. Note that the bound of (13) depends exponentially on d; precisely, on $C_0 = d2^{2d+1}$. It is possible to improve this dependency to the order of $(1 + \frac{r}{\sigma})^{2d}$. However, in this setting, we think of r as being a constant radius (rather than $r = r_n \to 0$ as $n \to \infty$, and therefore even this improvement results in exponential dependency on the dimension d. Make references suggesting this type of exponential dependency is not uncommon in non-parametric clustering problems.

Remark 2. Other than the looseness implied by Remark 1, the error bound of (13) is almost tight. Specifically, choosing

$$A_\sigma = B(0, \sigma),$$
$$f(x) = \begin{cases} \tau \text{ for } x \in A_\sigma, \\ f(x) = \tau - \rho(x, A_\sigma)^\gamma \text{ for } 0 < \rho(x, A_\sigma) < r \end{cases}$$

we have

$$\frac{\Phi_{n,r}(A_\sigma[\mathbf{X}])}{r} \geq C_1 \frac{(\tau - \frac{r^{\epsilon+1}}{\epsilon+1})}{\tau} \tag{14}$$

for some constant C_1, with probability at least $1 - \delta$. (Note that a factor of $1/\sigma$ is not (13) replicated in this lower bound.) Provide justification in supplement, and cite it.

4. Implications, extensions, and discussion

4.1. Examples

4.2. Experiments

References

Andersen, R., Chung, F., and Lang, K. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 475–486, 2006.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for the cluster tree. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 343–351. Curran Associates, Inc., 2010.

Fortunato, S. Community detection in graphs. *Physics Reports*, 486(3):75 – 174, 2010. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2009.11.002. URL `http://www.sciencedirect.com/science/article/pii/S0370157309002841`.

Hartigan, J. A. Consistency of single-linkage for high-density clusters. *Journal of the American Statistical Association*, 1981.

Haveliwala, T. H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796, 2003.

Kannan, R., Vempala, S., and Vetta, A. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, May 2004. ISSN 0004-5411. doi: 10.1145/990308.990313. URL `http://doi.acm.org/10.1145/990308.990313`.

Leskovec, J., Lang, K. J., and Mahoney, M. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

Mahoney, M. W., Orecchia, L., and Vishnoi, N. K. A local spectral method for graphs: with applications to improving graph partitions and exploring data graphs locally. *Journal of Machine Learning Research*, 13:2339–2365, 2012.

Spielman, D. A. and Teng, S.-H. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

Spielman, D. A. and Teng, S.-H. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26, 2013.

Spielman, D. A. and Teng, S.-H. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):835–885, 2014.

Trillos, N. G. and Slepcev, D. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239 – 281, 2018. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2016.09.003. URL `http://www.sciencedirect.com/science/article/pii/S106352031630063X`.

von Luxburg, U., Belkin, M., and Bousquet, O. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 04 2008.

Yang, J. and Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, Jan 2015. ISSN 0219-3116. doi: 10.1007/s10115-013-0693-z. URL `https://doi.org/10.1007/s10115-013-0693-z`.

Zhu, Z. A., Lattanzi, S., and Mirrokni, V. S. A local algorithm for finding well-connected clusters. In *ICML (3)*, pp. 396–404, 2013.