Journal of Machine Learning Research 21 (2020) 1-35 Submitted 12/18; Revised 10/19; Published 1/20

Connecting Spectral Clustering to Maximum Margins and Level Sets

David P. Hofmeyr DHOFMEYR@SUN.AC.ZA
Department of Statistics and Actuarial Science
Stellenbosch University
Stellenbosch, South Africa

Editor: Mehryar Mohri

Abstract

We study the connections between spectral clustering and the problems of maximum margin clustering, and estimation of the components of level sets of a density function. Specifically, we obtain bounds on the eigenvectors of graph Laplacian matrices in terms of the between cluster separation, and within cluster connectivity. These bounds ensure that the spectral clustering solution converges to the maximum margin clustering solution as the scaling parameter is reduced towards zero. The sensitivity of maximum margin clustering solutions to outlying points is well known, but can be mitigated by first removing such outliers, and applying maximum margin clustering to the remaining points. If outliers are identified using an estimate of the underlying probability density, then the remaining points may be seen as an estimate of a level set of this density function. We show that such an approach can be used to consistently estimate the components of the level sets of a density function under very mild assumptions.

Keywords: spectral clustering, maximum margin clustering, density clustering, level sets, convergence, asymptotics, consistency

1. Introduction

In maximum margin clustering, the objective is to obtain cluster separators for which the distance to the nearest data points is maximised. If no constraints are placed on the formulation of the cluster separators, then the maximum margin solution partitions data so that the between cluster distance is maximised. Such solutions are intuitively attractive, since we naturally associate similarities between data with how close they are in some metric space, most frequently Euclidean space. Maximising the between cluster Euclidean separation therefore seems like a sensible approach. However, such solutions are extremely sensitive to noise, and in many cases the clustering solution which maximises between cluster distance will only separate isolated points arising in the outer regions of a collection of data.

In the statistical approach to clustering, we imagine that our data arise from some probability distribution, and it is convenient to assume that this distribution comprises a mixture of simple components, each one of which representing a cluster. The most popular parametric model in this approach is the Gaussian mixture model (GMM). In this case, the maximum margin clustering solution will separate points in the tails of the mixture with extremely high probability as the sample size increases. This is because the density

between mixture components is higher than it is in the tails. In fact, unless the clusters (mixture components) are supported on disjoint, compact sets, it is generally very unlikely that unconstrained maximum margin solutions will be relevant for clustering. A simple but effective approach to mitigating the effect of noise, or isolated tail observations, is to manually remove points which are believed to be in the tails of the underlying distribution, and only apply a large margin clustering method to the remaining points. If these tail points are identified using an empirical estimate of the underlying density, then we arrive at the very well known problem of level set estimation.

Consider a probability density function, $p : \mathbb{R}^d \to \mathbb{R}^+$. Then the level set of p, at level λ, which we denote $\mathcal{L}(\lambda)$, is given by,

$$\mathcal{L}(\lambda) = \{\mathbf{x} \in \mathbb{R}^d | p(\mathbf{x}) \geq \lambda\}. \tag{1}$$

We note that some authors refer to the above as the upper- or super-level set, at level λ. Now, if p is multimodal, then as λ increases, the level set splits into multiple connected *components* which surround the modes of p. Each such component may then be associated with a cluster. This cluster definition has been widely adopted since the introduction of this formulation given by Hartigan (1975). Implicitly then, clusters are associated with high density regions around each of the modes of the probability density. This is consistent with the intuition underlying the mixture model formulation, assuming the mixture components are prominent enough that they result in modes in the density. However, the level set formulation is not constrained by any parametric assumptions which arise in the explicit mixture model approaches, such as GMMs. It also allows the clusters to take on arbitrary shapes, where most practically adopted parametric mixture models result in convex, or nearly convex clusters. Notice also that that the truncation of the random variable X, with density p, within $\mathcal{L}(\lambda)$ may be seen as having a mixture density whose mixture components are the truncations of X within the different components of $\mathcal{L}(\lambda)$. For $\lambda > 0$, except for pathological cases, these mixture components are supported on disjoint compact sets, and so any method which performs maximum margin clustering may be reasonably expected to be able to estimate the different components of $\mathcal{L}(\lambda)$. One of the theoretical benefits of the level set approach to clustering is that, provided simple assumptions on the density function, p, it leads to a well posed statistical estimation problem. Indeed, numerous consistent procedures for the estimation of level set components have been proposed (Walther, 1997; Cuevas et al., 2000; Rinaldo et al., 2010; Pelletier and Pudlo, 2011).

In this paper we study the consistency of estimating level set components, using spectral clustering applied to a truncated sample based on an empirical estimate of p. Spectral clustering is a relatively recent approach to clustering which has become extremely popular for its flexibility and its comparative algorithmic simplicity. Spectral clustering obtains a relaxed solution of the normalised graph cut problem via the eigenvectors of the corresponding graph Laplacian matrix. We study specifically the spectral clustering solutions for similarity graphs of points in Euclidean $\mathbb{R}^d$. We begin our analysis by deriving bounds on the eigenvectors of the Laplacian matrices. These bounds are used to show that the maximum margin clustering solution arises trivially from the spectral clustering solution, as the scaling parameter is reduced towards zero. We go on to obtain sufficient conditions on the convergence rate of the scaling parameter to consistently estimate $\mathcal{L}(\lambda)$, and ensure that, almost surely as $n \to \infty$, the components correspond to the maximum margin clustering

solution. It is found that these rates are also sufficient to ensure that the spectral clustering solution recovers the maximum margin solution, and thus the components of $\mathcal{L}(\lambda)$. So far this assumes the number of components of $\mathcal{L}(\lambda)$ is known. We therefore also derive bounds on the eigenvalues of the same graph Laplacians, which allow us to consistently estimate the number of components of $\mathcal{L}(\lambda)$.

The remainder of the paper is organised as follows. In Section 2 we give a summary of the main results of the paper. In Section 3 we discuss related work, and how our results extend on this body of literature. In Section 4 we briefly discuss spectral clustering, and the formulation of graph Laplacian matrices from points in $\mathbb{R}^d$. The main results of the paper are given in Section 5. Sections 5.1 and 5.2 present derivations of bounds on the eigenvectors and eigenvalues of graph Laplacians respectively. In Section 5.3 these bounds are shown to result in the convergence of spectral clustering to the maximum margin clustering solution, as the scaling parameter is reduced towards zero. Then in Section 5.4 these results are placed in the context of level sets, and the consistency of the estimation procedure of applying spectral clustering to the truncated sample is shown. Finally we conclude with a discussion of the results in Section 6.

2. Summary of Main Results

The main contributions of this work are in theoretically connecting spectral clustering with the problems of maximum margin clustering and of level set estimation. As we discussed in the previous section, there is an intuitive connection between these problems when maximum margin clustering is applied to a truncated sample, in which observations with low empirical density are removed. We state the results at this stage in a simplified form which captures the main points of the results. Technical details regarding assumptions, etc., are deferred to the relevant sections. We begin by introducing notation and terminology which will be useful here, and in the remaining paper.

2.0.1. Notation and Terminology

Most of the notation we use is fairly standard, but for completeness we list what is not universally employed, as well as terminology which we introduce for convenience.

For natural number $n \in \mathbb{N}$, we use $[n]$ for the set containing the first n natural numbers, i.e., $[n] = \{1, 2, ..., n\}$. For a set S and natural number k, we use $\Pi_k(S)$ to denote the collection of partitions of S into k non-empty subsets. Any use of the generic norm notation, $||\cdot||$, will refer to the Euclidean, or L_2 norm. Similarly any reference to a metric, $d(\cdot, \cdot)$, will thus correspond to the Euclidean metric, i.e., $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. For $S, U \subset \mathbb{R}^d$ and $\mathbf{x} \in \mathbb{R}^d$ we use $d(\mathbf{x}, S) = \inf_{\mathbf{y} \in S} d(\mathbf{x}, \mathbf{y})$ to denote the distance between $\mathbf{x}$ and the set S, and $d(S, U) = \inf_{\mathbf{x} \in S, \mathbf{y} \in U} d(\mathbf{x}, \mathbf{y})$ to denote the distance between the sets S and U. By default we set $d(\mathbf{x}, \emptyset) = \infty$ for all $\mathbf{x} \in \mathbb{R}^d$, where $\emptyset$ is the empty set. We also use $\mathrm{Diam}(S) = \sup_{\mathbf{x}, \mathbf{y} \in S} d(\mathbf{x}, \mathbf{y})$ to represent the diameter of a set $S \subset \mathbb{R}^d$. We use $\mathcal{B}_\delta(\mathbf{x}) := \{\mathbf{y} \in \mathbb{R}^d | d(\mathbf{x}, \mathbf{y}) < \delta\}$ to denote the δ-neighbourhood of $\mathbf{x} \in \mathbb{R}^d$, and we will also write $\mathcal{B}_\delta(S) := \bigcup_{\mathbf{x} \in S} \mathcal{B}_\delta(\mathbf{x})$ for the δ-neighbourhood of a set $S \subset \mathbb{R}^d$. We say that a set $S \subset \mathbb{R}^d$ is connected at distance δ if there is no binary partition of S into S_1, S_2 such that $d(S_1, S_2) > \delta$. Equivalently, S is connected at distance δ if the closure of $\mathcal{B}_{\delta/2}(S)$ is a connected set. We use $K(\cdot)$, $K : \mathbb{R}^+ \to \mathbb{R}^+$, to represent a *kernel* function, used to determine

similarities between points in $\mathbb{R}^d$. In the general setting we will simply use $K(d(\mathbf{x},\mathbf{y}))$ to capture the similarity between points $\mathbf{x}$ and $\mathbf{y}$, whereas when considering convergence we will incorporate a scaling factor, $\sigma > 0$, either explicitly using $K(d(\mathbf{x},\mathbf{y})/\sigma)$, or implicitly using $K_\sigma(d(\mathbf{x},\mathbf{y}))$, which we intend to be taken as equivalent. Finally, if $\mathbf{U} \in \mathbb{R}^{n\times m}$ is a matrix then we will write $\mathbf{U}_{a:b,c:d}$ for the sub-matrix containing rows $a, a+1, ..., b-1, b$ and columns $c, c+1..., d-1, d$. We will just write : for all rows/columns, and use just a single index as is usual for a single row/column. For example the matrix $\mathbf{U}_{1:5,:}$ contains the first five rows and all columns of $\mathbf{U}$.

2.1. Maximum Margins

We focus on the maximum margin clustering solution defined as the partition of a data set, $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, which attains the maximum between cluster distance among all partitions of a given size. That is, the maximum margin clustering solution, for k clusters, is the solution to the optimisation problem

$$\max_{\{\mathcal{C}_1,...,\mathcal{C}_k\}\in\Pi_k(\mathcal{X})} \left\{ \min_{\substack{i,j\in[k]\\ i\neq j}} d(\mathcal{C}_i,\mathcal{C}_j) \right\}. \tag{2}$$

To establish a connection between spectral clustering and maximum margin clustering, we derive bounds on the eigenvectors of graph Laplacians which are expressed only in terms of the between cluster separatedness, defined as $\min_{i,j\in[k]} d(\mathcal{C}_i,\mathcal{C}_j); i \neq j$, and the within cluster connectedness, defined as $\max_{i\in[n],j\in[k]:\mathbf{x}_i\in\mathcal{C}_j} d(\mathbf{x}_i, \mathcal{C}_j \setminus \{\mathbf{x}_i\})$. Between cluster separatedness is simply the smallest distance between any two points which belong to different clusters, whereas within cluster connectedness is the greatest distance between any point and the rest of the cluster to which it has been assigned. We consider graph Laplacians constructed from similarity graphs in which the similarity between points $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by $K(d(\mathbf{x}_i,\mathbf{x}_j)/\sigma)$, with the kernel $K(\cdot)$ satisfying very mild assumptions, and σ being a positive scalar. These eigenvector bounds allow us to establish that, if $\mathbf{U} = \{\mathbf{u}_1, ..., \mathbf{u}_n\}$ represents the transformation of $\mathcal{X}$ obtained from the first k eigenvectors of the graph Laplacian, and $\{\mathcal{C}_1, ..., \mathcal{C}_k\}$ is the maximum margin clustering solution of $\mathcal{X}$, then (Theorems 12, 14, 15)

$$\lim_{\sigma\to 0^+} \max_{\substack{i,j,l\in[k]\\ j\neq l}} \frac{\mathrm{Diam}(\tilde{\mathcal{C}}_i)}{d(\tilde{\mathcal{C}}_j,\tilde{\mathcal{C}}_l)} = 0,$$

where $\tilde{\mathcal{C}}_i = \bigcup_{j:\mathbf{x}_j\in\mathcal{C}_i}\{\mathbf{u}_j\}$ is cluster $\mathcal{C}_i$ within the eigenvector representation. That is, for small enough scaling parameter, the diameters of the clusters in the maximum margin solution, when transformed using the eigenvectors of the graph Laplacian, are much smaller than the distances between them. The maximum margin clustering solution can thus be easily obtained from these eigenvectors.

A number of different methods have been proposed for obtaining the final clustering solution from the eigenvectors, $\mathbf{U}$. Arguably the most popular is to apply k-means (Von Luxburg, 2007), and it should be expected that most sensible approaches applied in the context described above will obtain the optimal solution. For ease of analysis, we will only discuss

the k-centers solution, for which a linear time 2-approximation algorithm exists (Gonzalez, 1985). The k-centers objective for a set of points, $\mathcal{X}$, and a set of centres, $C \subset \mathcal{X}$, is given by $\max_{\mathbf{x}\in\mathcal{X}} \min_{\mathbf{c}\in C} d(\mathbf{x}, \mathbf{c})$. It should be clear that when the diameters of all clusters are less than $\frac{1}{T}$ times the smallest distance between any pair of clusters, then any T-approximation algorithm for the optimal k-centers problem will recover the optimal solution.

2.2. Level Sets

In studying the consistent estimation of the components of the level sets of the density, p, we adopt the following definition of the limit of a sequence of sets, $\{A_n\}_{n=1}^{\infty}$. To begin, define the limits supremum and infimum as,

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m, \ \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m.$$

Then $\lim_{n\to\infty} A_n = A$ if and only if $\limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = A$, otherwise the limit does not exist. Note that if A is closed and $\{A_n\}_{n=1}^{\infty}$ is such that there is a positive sequence, $\{a_n\}_{n=1}^{\infty}$, with $\lim_{n\to\infty} a_n = 0$, and

$$A \subset A_n \subset \mathcal{B}_{a_n}(A), \ \forall n \in \mathbb{N},$$

then $\lim_{n\to\infty} A_n = A$. Note also that with the above definition, $\lim_{n\to\infty} A_n = A$ implies that $\lim_{n\to\infty} \mu(A_n \Delta A) = 0$, where $\mu(\cdot)$ is the Lebesgue measure and Δ denotes the symmetric difference, and also that $\lim_{n\to\infty} d_H(A_n, A) = 0$, where $d_H(\cdot,\cdot)$ is the Hausdorff metric.

We will show herein that appropriately selected neighbourhoods of the clusters identified by spectral clustering, applied to a well-defined truncation of the sample, consistently estimates the components of a chosen level set of p. To that end, suppose that, for $\lambda > 0$, the level set $\mathcal{L}(\lambda)$ has c components, denoted by $\ell(\lambda, 1), ..., \ell(\lambda, c)$, and let $X_1, X_2, ...$ be an i.i.d. sequence of random variables with density p. Then we can find a positive sequence of scalars, $\{\sigma_n\}_{n=1}^{\infty}$, with $\lim_{n\to\infty} \sigma_n = 0$, and a sequence of thresholds, $\{\Lambda_n\}_{n=1}^{\infty}$, such that if we define for each $n \in \mathbb{N}$, the set

$$\widehat{\mathcal{L}(\lambda)}^{(n)} = \left\{ X_j \middle| j \leq n, \sum_{i=1}^{n} K(d(X_i, X_j)/\sigma_n) > \Lambda_n \right\},$$

and let $\{\mathcal{C}_1^{(n)}, ..., \mathcal{C}_c^{(n)}\}$ be the spectral clustering solution from $\widehat{\mathcal{L}(\lambda)}^{(n)}$ using similarities $K(d(X_i, X_j)/\sigma_n)$ for $X_i, X_j \in \widehat{\mathcal{L}(\lambda)}^{(n)}$, the following holds with probability one. There is a sequence of permutations of $[c]$, say $\{\omega^n\}_{n=1}^{\infty}$, such that (Theorems 18, 19, 20)

$$\lim_{n\to\infty} \mathcal{B}_{\sigma_n^{1-\epsilon}}\left(\mathcal{C}_{\omega_k^n}^{(n)}\right) = \ell(\lambda, k),$$

for all $k \in [c]$, where $\epsilon \in (0, 1)$ is any fixed constant. That is, up to a reallocation of cluster labels, well defined sequences of neighbourhoods of the clusters obtained from spectral clustering converge to the true level set components. Notice that this assumes a fixed number of clusters, c. We therefore show further that, with probability one, the first c scaled eigenvalues (with known scaling) of the graph Laplacians from the similarity graphs of $\{\widehat{\mathcal{L}(\lambda)}^{(n)}\}_{n=1}^{\infty}$

converge to zero, while all other scaled eigenvalues tend to infinity (Theorems 21, 22, 23). With probability one there is therefore an $n \in \mathbb{N}$ beyond which the estimated number of clusters, based on these eigenvalues, is equal to c.

3. Relation to Existing Work

Although multiple existing approaches for estimating components of level sets use graph theoretic partitioning algorithms, the only existing approach of which we are aware which is based directly on spectral clustering is that of Pelletier and Pudlo (2011). There the authors use an approach similar to that of Rinaldo et al. (2010), where first a consistent estimator of the underlying density is used to identify points in the estimated level set, and then a fixed bandwidth kernel is applied to these points to estimate the components/clusters. The authors show that under relatively mild assumptions their normalised spectral clustering algorithm consistently estimates the level set components. An important difference between this and our approach is that this existing work uses a fixed bandwidth parameter when applying spectral clustering. As a result, it is necessary that the distance between the components of the level set is known. Otherwise it is possible that this procedure will merge components which are close together. We consider the more natural case where the scaling parameter is reduced as the number of observations increases. In fact we find that the rate of convergence of the sequence of scaling parameters required for consistent estimation of the level set components using spectral clustering, is also sufficient for uniformly consistent estimation of the underlying density, and hence the level set itself. The same kernel computations used for estimating the level set are therefore also used in the spectral clustering step. The approach of Pelletier and Pudlo (2011) also applies only to kernels with bounded support. This makes the analysis simpler since, provided the bandwidth is smaller than half the distance between the level set components, the similarity graph of the points in the estimated level set is disconnected with high probability, and it is well known that spectral clustering recovers the components of a disconnected graph (Von Luxburg, 2007). We extend this to allow kernels with unbounded support, provided the tails do not decay too slowly, and hence include the ubiquitous Gaussian kernel. Finally, our consistency analysis extends that of Pelletier and Pudlo (2011) by considering the Laplacian matrices derived from the Ratio Cut as well as the Normalised Cut objective.

Arguably the most important existing work on the consistency of spectral clustering is the foundational work of Von Luxburg et al. (2008). There the authors investigate the consistency of spectral clustering in a general sense, rather than in relation to the estimation of a particular feature of the underlying distribution. In fact these authors also apply a fixed bandwidth kernel, and hence any asymptotic properties of the spectral clustering solution will be in relation to the convolution of the underlying distribution with the distribution whose density is given by the fixed bandwidth kernel. Other existing works which connect spectral clustering to the properties of the underlying distribution do so by studying the properties of the exact normalised cut solutions, and not the spectral clustering relaxations (Narayanan et al., 2006; Trillos et al., 2016; Hofmeyr, 2019). These approaches are therefore fundamentally different from the present work.

A distinct body of work exists which bypasses any relation to similarities between Euclidean embedded data, and instead focuses on theoretical properties of spectral clustering

applied directly to graphs satisfying certain properties. Such approaches have important applications in, e.g., network analysis. For example, Lei et al. (2015) study the consistency of spectral clustering in recovering community structure in the stochastic block model (SBM). The analysis of that problem differs considerably from that of ours in a number of ways; notably edges between vertices are binary (i.e., take only the values 0 and 1) and are independent of one another, whereas the spatial locations of collections of points in $\mathbb{R}^d$ create a strong dependence between these edges. In addition, finite sample results on the accuracy of spectral clustering have been derived for graphs whose Laplacian matrices effectively have a large gap between the k-th and $(k+1)$-th eigenvalues (Peng et al., 2015). The results of Section 5.2 can easily be used to obtain bounds on these *eigen-gaps*, and combining these approaches may lead to similar accuracy results for Euclidean embedded data satisfying certain separation assumptions.

Finally, as far as we are aware, the only existing work which explicitly connects spectral clustering to maximum margin clustering, is that of Hofmeyr et al. (2019). There the authors show that the optimal one-dimensional projection of a dataset for spectral clustering converges to the normal vector to the maximum margin hyperplane for clustering. The results in this existing work effectively ensure that the spectral clustering solution for points in $\mathbb{R}$ converges to the maximum margin solution. The large margin results presented here therefore extend these existing results to the multivariate setting.

4. Graph Cuts and Spectral Clustering

In this section we give a brief but explicit introduction to spectral clustering. For a very accessible and extended discussion on the topic, the reader is referred to Von Luxburg (2007). We begin with a description of graphs, before introducing similarity graphs for Euclidean embedded data. We then go on to discuss objectives for optimal partitioning of graphs, and their relation to clustering Euclidean data. Finally, we discuss a common re-formulation of these objectives in terms of graph Laplacian matrices, and discuss the solution to relaxed versions of these optimal partitioning problems.

A *graph* is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V}$ is a set of so-called *vertices*, and the *edges* of the graph, $\mathcal{E}$, may, without loss of generality, be seen as a map $\mathcal{E} : [|\mathcal{V}|] \times [|\mathcal{V}|] \to \mathbb{R}_+$ for which $\mathcal{E}(i,j)$ represents the weight, or strength of the connection between the pair of vertices $v_i, v_j \in \mathcal{V}$. If $\mathcal{E}(i,j) = 0$, then we say that the vertices v_i and v_j are not connected. A complete graph is one for which $\mathcal{E}(i,j) > 0$ for all $i, j \in [|\mathcal{V}|]$, and a disconnected graph is one for which there is a partition of $\mathcal{V}$ into some $c > 1$ non-empty subsets, say $\{\mathcal{V}_1, ...\mathcal{V}_c\}$, such that whenever $v_i \in \mathcal{V}_{i'}, v_j \in \mathcal{V}_{j'}$, where $i' \neq j'$, we have $\mathcal{E}(i,j) = 0$. In other words, there are no non-zero edges connecting vertices in different elements of the partition. If the sub-graphs formed by the sets of vertices $\mathcal{V}_i, i \in [c]$, are each themselves connected graphs (i.e., not disconnected), then these are referred to as the components of $\mathcal{G}$. Here, the sub-graph formed by the vertices in $\mathcal{V}_i$, some $i \in [c]$, has edges $\mathcal{E}_i$ inherited from the edges in $(\mathcal{V}, \mathcal{E})$ associated with pairs of elements in $\mathcal{V}_i$. That is, $\mathcal{E}_i(k,l) := \mathcal{E}(k', l')$ where $v_{k'}, v_{l'}$ are the k-th and l-th elements of $\mathcal{V}_i$.

Now consider a collection of points, $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, in $\mathbb{R}^d$. We study the graph with vertices given by the elements in $\mathcal{X}$, and where edges are determined by the similarities between pairs of points/vertices. That is, $\mathcal{E}(i,j) = \text{similarity}(\mathbf{x}_i, \mathbf{x}_j)$. It is common, and

intuitively appealing, to determine similarities between points based on how close they are with respect to a metric, $d(\cdot,\cdot)$, on $\mathbb{R}^d$. That is, to set similarity$(\mathbf{x}_i,\mathbf{x}_j) = K(d(\mathbf{x}_i,\mathbf{x}_j))$, where the similarity kernel, $K(\cdot)$, is non-increasing on the non-negative real numbers. In this way, pairs of points which are nearer in space are assigned higher similarity than pairs which are more distant. It is worth noting that variations on this basic formulation of the similarity graph have been studied (Von Luxburg, 2007). For example, in some cases edges below a certain threshold, or edges between pairs of points which are not in one-another's set of near neighbours, are set to zero. In the first case this can be seen equivalently as truncating the support of the kernel, $K(\cdot)$, even if the function itself is strictly positive over its entire domain. The latter has a similar effect, but the truncation varies depending on the location of the points. As mentioned previously, we study the case of kernels with unbounded support, i.e., where no such truncation is applied, and hence in which the similarity graph is complete.

Now, a *cut* of a graph refers to a partition of its vertices through the removal of a subset of its edges, i.e., setting those edge weights to zero. The partition corresponds with the components of the resulting disconnected graph, after those edges in the cut are removed. The value of the cut is given by the sum of the edges which were removed. There is thus an obvious bijection between the partitions/clusterings of $\mathcal{X}$ and the cuts of its similarity graph. We can therefore use the properties of graph cuts, and the optimisation problems associated with finding optimal cuts, to study the corresponding clustering solutions. Two popular graph cut objectives considered extensively in the clustering context are the *Ratio Cut* (RCut) and *Normalised Cut* (NCut). Stated explicitly in relation to the data set $\mathcal{X}$, if $\{\mathcal{C}_1,...,\mathcal{C}_k\} \in \Pi_k(\mathcal{X})$, then

$$\mathrm{RCut}(\mathcal{C}_1,...,\mathcal{C}_k) = \sum_{i=1}^{k} \frac{\mathrm{Cut}(\mathcal{C}_i, \mathcal{X}\setminus\mathcal{C}_i)}{|\mathcal{C}_i|}, \tag{3}$$

$$\mathrm{NCut}(\mathcal{C}_1,...,\mathcal{C}_k) = \sum_{i=1}^{k} \frac{\mathrm{Cut}(\mathcal{C}_i, \mathcal{X}\setminus\mathcal{C}_i)}{\mathrm{vol}(\mathcal{C}_i)}, \tag{4}$$

where

$$\mathrm{Cut}(\mathcal{C}, \mathcal{X}\setminus\mathcal{C}) = \sum_{\substack{i,j:\mathbf{x}_i\in\mathcal{C},\\ \mathbf{x}_j\notin\mathcal{C}}} K(d(\mathbf{x}_i,\mathbf{x}_j)),\ \mathrm{vol}(\mathcal{C}) = \sum_{\substack{i,j:\mathbf{x}_i\in\mathcal{C}\\ \mathbf{x}_j\in\mathcal{X}}} K(d(\mathbf{x}_i,\mathbf{x}_j)).$$

Broadly speaking, solutions which minimise either RCut or NCut tend to correspond with solutions in which the total similarity between points in different clusters is low, but solutions containing very small clusters or clusters with low internal similarity are avoided through normalisation by either the cardinatity $|\cdot|$, or volume vol$(\cdot)$, of the individual clusters. Both RCut and NCut are attractive objectives for clustering, but obtaining the globally optimal solutions is NP-hard (Wagner and Wagner, 1993). Furthermore, obtaining high quality locally optimal solutions is not straightforward. Instead a relaxation is considered, in which the data are transformed using the eigenvectors of the graph Laplacian matrices. Clustering using the spectral decomposition of graph Laplacian matrices is referred to as spectral clustering.

It is convenient algebraically to store the information in the graph, $\mathcal{G}$, using the so-called *affinity matrix*, $\mathbf{A} \in \mathbb{R}^{n \times n}$, which contains the edge weights for all pairs of points/vertices, i.e., $\mathbf{A}_{i,j} = \mathcal{E}(i,j) = K(d(\mathbf{x}_i, \mathbf{x}_j))$. In addition, let $\mathbf{D} \in \mathbb{R}^{n \times n}$ be the *degree matrix* of the graph, which is the diagonal matrix with i-th diagonal given by the sum of the i-th row of $\mathbf{A}$. Then the *unnormalised Laplacian* and *normalised Laplacian* of $\mathcal{G}$ are given respectively by $\mathbf{L} = \mathbf{D} - \mathbf{A}$ and $\mathbf{L}_{\mathrm{N}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. For completeness we will consider two normalised Laplacians, where the second, which we denote $\mathbf{L}_{\mathrm{N}_0}$, arises from the graph which is the same as $\mathcal{G}$ but the reflexive edges, i.e., those connecting vertices to themselves, are removed/set to zero. Algebraically we have $\mathbf{L}_{\mathrm{N}_0} = \mathbf{I} - \mathbf{D}_0^{-1/2}\mathbf{A}_0\mathbf{D}_0^{-1/2}$, where $\mathbf{A}_0 = \mathbf{A} - K(0)\mathbf{I}$, and hence similarly $\mathbf{D}_0 = \mathbf{D} - K(0)\mathbf{I}$. As it turns out, in the context we consider, the differences between analysing $\mathbf{L}_{\mathrm{N}}$ and $\mathbf{L}_{\mathrm{N}_0}$ are far greater than those between $\mathbf{L}$ and $\mathbf{L}_{\mathrm{N}}$. This arises from the fact that the diagonal elements of $\mathbf{D}_0$, unlike those of $\mathbf{D}$, are not bounded away from zero.

Now, it has been shown that the solution to the optimisation problem,

$$\min_{\mathbf{U} \in \mathbb{R}^{n \times k}} \operatorname{tr}(\mathbf{U}^\top \mathbf{L} \mathbf{U}), \quad \text{such that } \mathbf{U}^\top \mathbf{U} = \mathbf{I}, \tag{5}$$

can be seen as a continuous relaxation of the cluster indicator vectors for the optimal RCut solution, scaled so that the columns form an orthonormal system (Hagen and Kahng, 1992). The solution to (5) is given by the eigenvectors associated with the smallest k eigenvalues of $\mathbf{L}$. Similarly, the solution to the problem

$$\min_{\mathbf{U} \in \mathbb{R}^{n \times k}} \operatorname{tr}(\mathbf{U}^\top \mathbf{L} \mathbf{U}), \text{ such that } \mathbf{U}^\top \mathbf{D}^{-1} \mathbf{U} = \mathbf{I}, \tag{6}$$

has as columns relaxations of scaled cluster indicator vectors for the optimal NCut solution (Shi and Malik, 2000). In this case the solution can be shown to be given by $\mathbf{D}^{-1/2}\mathbf{U}$, where the columns of $\mathbf{U}$ are the first k eigenvectors of $\mathbf{L}_{\mathrm{N}}$. Importantly, since these eigenvector problems are relaxed versions of the minimum RCut and NCut problems, we have that

$$\min_{\mathbf{U} \in \mathbb{R}^{n \times k},\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \operatorname{tr}(\mathbf{U}^\top \mathbf{L} \mathbf{U}) \leq \min_{\{\mathcal{C}_1, \ldots, \mathcal{C}_k\} \in \Pi_k(\mathcal{X})} \mathrm{RCut}(\mathcal{C}_1, \ldots, \mathcal{C}_k), \tag{7}$$

$$\min_{\mathbf{U} \in \mathbb{R}^{n \times k},\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \operatorname{tr}(\mathbf{U}^\top \mathbf{L}_{\mathrm{N}} \mathbf{U}) \leq \min_{\{\mathcal{C}_1, \ldots, \mathcal{C}_k\} \in \Pi_k(\mathcal{X})} \mathrm{NCut}(\mathcal{C}_1, \ldots, \mathcal{C}_k). \tag{8}$$

Upper bounds on the optimal normalised and ratio cuts, in terms of the eigenvalues of the graph Laplacians, have also been studied, by, e.g., Lee et al. (2014). However, the lower bounds above are sufficient for our analysis.

In the remainder we study the properties of the graph Laplacian matrices in terms of their eigenvectors and eigenvalues, and show how these can be used to obtain maximum margin clustering solutions and to consistently estimate the components of the level sets of a density function on $\mathbb{R}^d$. Specifically, we will show that the matrix whose columns are the first k eigenvectors converges to one which trivially exposes the maximum margin clustering solution, as the similarities become more and more locally concentrated. We use the same supporting results to show further that by applying spectral clustering to truncations of an increasing sample from a continuous probability distribution on $\mathbb{R}^d$, we can consistently estimate the components of the level sets of its density. The eigenvalues of the corresponding

graph Laplacians are used to consistently estimate the number of components, while the eigenvectors are shown to trivially recover the partition of the points in the level set into its different components.

It is important to note that if a matrix has multiple equal eigenvalues, then the corresponding eigenvectors are not unique. The results presented in this paper do not require uniqueness of eigenvectors, and only use the above inequalities, the orthogonality of the eigenvectors, and properties of the Laplacian matrices. For simplicity, most results refer to *the eigenvectors* of the Laplacian matrices, however these can be stated to refer to any set of orthogonal eigenvectors. Similarly, when referring to the eigenvectors corresponding to the k smallest eigenvalues, if the k-th eigenvalue is repeated, then we only consider a total of k eigenvectors.

5. Connecting Spectral Clustering to Maximum Margins and Level Sets

In this section we present complete derivations of the theoretical contributions of this paper. We first derive bounds on the eigenvectors and eigenvalues of graph Laplacian matrices, in terms of the within cluster connectedness and between cluster separation. We go on to show that, given mild assumptions on the similarity function, as the scaling parameter is reduced to zero the spectral clustering solution converges to the maximum margin clustering solution, in the sense that within cluster distances (within the eigenvector representation) converge to zero, while between cluster distances are bounded below. For both the unnormalised Laplacian, $\mathbf{L}$, and the normalised Laplacian, $\mathbf{L}_{\mathrm{N}}$, these bounds arise fairly straightforwardly. However, in the case of the normalised Laplacian, $\mathbf{L}_{\mathrm{N}_0}$, derived from the graph with reflexive edges removed, no such lower bound on the between cluster distances is immediately forthcoming. Instead, in this case, we show that within cluster distances converge to zero at a much faster rate than between cluster distances, therefore having the same practical relevance of exposing the maximum margin clustering solution clearly. Finally, we go on to establish conditions on the rate of convergence of the scaling parameter, in the context of an increasing sample arising from a continuous probability distribution on $\mathbb{R}^d$, in order to simultaneously and consistently estimate the level set; the number of components of the level set; as well as ensure that spectral clustering recovers the partition of points in the level set according to the components in which they lie.

5.0.1. Assumptions on the kernel function, K

As mentioned previously, we present our analysis for kernels with unbounded support. It is worth noting that the results can be shown to hold for kernels with bounded support, after suitable changes to the presentation herein. Once again, in the case where the support of the kernels is bounded, the similarity graph becomes disconnected as the scaling parameter is reduced, and hence the recovery of the solution by spectral clustering is immediate (Von Luxburg, 2007). It is therefore the unbounded support case which we find far more interesting. In particular, we present results for kernels satisfying the following,

AK1: $K(\cdot)$ is non-increasing and strictly positive on $[0, \infty)$.

AK2: $K(0) = 1$, $c_K \int_{\mathbb{R}^d} K(||\mathbf{x}||) d\mathbf{x} = 1$.

AK3: $\exists A, \alpha > 0$ such that $K(x)/K(y) \leq A \exp(-(x-y)^{\alpha})$ for all $0 \leq y \leq x$.

Assumption AK1 is very standard, and intuitively desireable for determining similarity, since it ensures that pairs which are closer are assigned higher similarity than pairs which are further apart. Assumption AK2 can always be achieved by scaling all similarities, provided the integral $\int_{\mathbb{R}^d} K(||\mathbf{x}||)d\mathbf{x}$ is finite. The normalisation constant c_K will be relevant when considering the estimation of the density using K. Assumption AK3 places an upper bound on the tail decay of the kernel, and excludes polynomially decaying tails, but includes, for example, the ubiquitous Gaussian kernel.

5.0.2. Assumptions on the density, p, and level set $\mathcal{L}(\lambda)$

We also make a few simplifying assumptions on the density p, and the level set of interest. It is certainly possible to relax these assumptions in favour of weaker ones, however we prefer to make assumptions which are stated as simply as possible. Furthermore, any distribution can be approximated arbitrarily well by one with a density which obeys the following conditions, for all levels, $\lambda > 0$. The reason for this is that these conditions are satisfied by finite mixtures of Gaussian densities, and the class of finite Gaussian mixtures can be used to approximate any distribution arbitrarily well (Alspach and Sorenson, 1972). In particular,

A1: We assume that p has bounded first derivative, so that $||\nabla p(\mathbf{x})|| < \kappa$ for all $\mathbf{x} \in \mathbb{R}^d$.

A2: We assume that $\exists C, \gamma > 0$ s.t. $\forall 0 < g < \gamma$ we have

$$\sup_{\mathbf{x} \in \mathcal{L}(\lambda - g) \setminus \mathcal{L}(\lambda)} d(\mathbf{x}, \mathcal{L}(\lambda)) \leq gC.$$

Assumption A1 allows us to use the uniform consistency of kernel density estimators, and also ensures there are finitely many components of the level set $\mathcal{L}(\lambda)$. Assumption A2 is a convenient way expressing a degree of regularity of the density around the level of interest. In particular, decreasing the level, λ, by a very small amount cannot lead to the inclusion, into the level set, of points which are substantially distant from their nearest point in $\mathcal{L}(\lambda)$.

5.1. Eigenvector Bounds for Graph Laplacians

In this section we derive bounds on the distances between points in the same clusters, when mapped into the Laplacian eigenvector representation through spectral clustering. These bounds are expressed in terms of the within cluster connectedness, and the between cluster separation only, and so can be used directly to relate the spectral clustering solution to the maximum margin clustering solution. These results only place upper bounds on the pairwise distances between points from the same clusters, and do not directly ensure that points in different clusters are distinguishable. To achieve this we present general results which can be seen as providing lower bounds on the between cluster separation for any data set with full column rank, in terms of the within cluster distortion. We later combine these results to show that the spectral clustering solution converges to the maximum margin solution, as the scaling parameter is reduced towards zero.

The following three results respectively provide the upper bounds on the within cluster distances in the eigenvectors of the unnormalised Laplacian, $\mathbf{L}$, normalised Laplacian, $\mathbf{L}_{\mathrm{N}}$, and normalised Laplacian from the graph with reflexive edges removed, $\mathbf{L}_{\mathrm{N}_0}$. It is worth noting that these bounds are very loose in the general setting, and that for some graphs these bounds can be trivially improved. However, our intention is to obtain bounds which depend only on the between cluster separatedness and within cluster connectedness, as these quantities are relevant for our asymptotic analyses.

Lemma 1 *Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, and let $\mathcal{C}_1, ..., \mathcal{C}_k$ be a partition of $\mathcal{X}$. For each $l \in [k]$, suppose $\mathcal{C}_l$ is connected at distance δ_l. Let $\mathbf{U} \in \mathbb{R}^{n\times n}$ have as columns the eigenvectors of the unnormalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i, \mathbf{x}_j)), i, j \in [n]$. Then for each $i, j \in [n], l \in [k]$ s.t. $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l$, we have*

$$||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| \leq \max_{m\in[k]} n^{1.5}k^{0.5}\sqrt{\frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X}\setminus\mathcal{C}_m))}{K_\sigma(\delta_l)}}.$$

Proof Since spectral clustering is a relaxation of the Ratio Cut problem, we have

$$\begin{aligned}
\sum_{i=1}^k \mathbf{U}_{:,i}^\top \mathbf{L}\mathbf{U}_{:,i} &\leq \min_{\{C_1,...,C_k\}\in\Pi_k(\mathcal{X})} \sum_{i=1}^k \sum_{j,l:\mathbf{x}_j\in C_i, \mathbf{x}_l\notin C_i} \frac{K_\sigma(||\mathbf{x}_j - \mathbf{x}_l||)}{|C_i|} \\
&\leq \sum_{i=1}^k \sum_{j,l:\mathbf{x}_j\in \mathcal{C}_i, \mathbf{x}_l\notin \mathcal{C}_i} \frac{K_\sigma(||\mathbf{x}_j - \mathbf{x}_l||)}{|\mathcal{C}_i|} \\
&\leq \sum_{i=1}^k \sum_{j,l:\mathbf{x}_j\in \mathcal{C}_i, \mathbf{x}_l\notin \mathcal{C}_i} \frac{K_\sigma(d(\mathcal{C}_i, \mathcal{X}\setminus\mathcal{C}_i))}{|\mathcal{C}_i|} \\
&\leq \sum_{i=1}^k |\mathcal{X}\setminus\mathcal{C}_i| K_\sigma(d(\mathcal{C}_i, \mathcal{X}\setminus\mathcal{C}_i)) \\
&\leq nk \max_{m\in[k]} K_\sigma(d(\mathcal{C}_m, \mathcal{X}\setminus\mathcal{C}_m))
\end{aligned}$$

Now take any $l \in [k]$. Since $\mathcal{C}_l$ is connected at distance δ_l, there exist $|\mathcal{C}_l| - 1$ pairs of points in $\mathcal{C}_l$ with indices $(i_1, j_1), ..., (i_{|\mathcal{C}_l|-1}, j_{|\mathcal{C}_l|-1})$ s.t. $||\mathbf{x}_{i_m} - \mathbf{x}_{j_m}|| \leq \delta_l$ for each $m \in [|\mathcal{C}_l| - 1]$ and the union of all such $\{\mathbf{x}_{i_m}, \mathbf{x}_{j_m}\}$ is equal to $\mathcal{C}_l$. To see this, notice that for any subset $C \subsetneq \mathcal{C}_l$, there is a pair $\mathbf{x} \in C, \mathbf{y} \in \mathcal{C}_l \setminus C$ with $||\mathbf{x} - \mathbf{y}|| \leq \delta_l$. Starting with $C = \{\mathbf{x}_{i_1}\}$ for any i_1, it is therefore possible to iteratively add points to C until $C = \mathcal{C}_l$, in such a way that a point is added to C only at a time at which it forms a pair with an element already in C which is at a distance less than or equal to δ_l.

By (Von Luxburg, 2007, Proposition 1) we know that for any $\mathbf{u} \in \mathbb{R}^n$ we have

$$\mathbf{u}^\top L\mathbf{u} = \frac{1}{2}\sum_{i,j} K_\sigma(||\mathbf{x}_i - \mathbf{x}_j||)(\mathbf{u}_i - \mathbf{u}_j)^2.$$

We therefore have

$$\sum_{i=1}^{k} \mathbf{U}_{:,i}^{\top} \mathbf{L} \mathbf{U}_{:,i} = \frac{1}{2} \sum_{i,j} K_\sigma(||\mathbf{x}_i - \mathbf{x}_j||)||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}||^2$$

$$\Rightarrow K_\sigma(||\mathbf{x}_{i_m} - \mathbf{x}_{j_m}||)||\mathbf{U}_{i_m,1:k} - \mathbf{U}_{j_m,1:k}||^2 \leq \sum_{i=1}^{k} \mathbf{U}_{:,i}^{\top} \mathbf{L} \mathbf{U}_{:,i} \text{ for each } m \in [|\mathcal{C}_l| - 1]$$

$$\Rightarrow K_\sigma(\delta_l)||\mathbf{U}_{i_m,1:k} - \mathbf{U}_{j_m,1:k}||^2 \leq \sum_{i=1}^{k} \mathbf{U}_{:,i}^{\top} \mathbf{L} \mathbf{U}_{:,i} \text{ for each } m \in [|\mathcal{C}_l| - 1]$$

$$\Rightarrow ||\mathbf{U}_{i_m,1:k} - \mathbf{U}_{j_m,1:k}|| \leq \sqrt{\frac{\sum_{i=1}^{k} \mathbf{U}_{:,i}^{\top} \mathbf{L} \mathbf{U}_{:,i}}{K_\sigma(\delta_l)}} \text{ for each } m \in [|\mathcal{C}_l| - 1]$$

$$\Rightarrow ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| \leq |\mathcal{C}_l| \sqrt{\frac{\sum_{i=1}^{k} \mathbf{U}_{:,i}^{\top} \mathbf{L} \mathbf{U}_{:,i}}{K_\sigma(\delta_l)}} \text{ for any } i, j \text{ s.t. } \mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l,$$

where the final step comes from the triangle inequality, since all points in $\mathcal{C}_l$ are connected by the pairs $\mathbf{x}_{i_m}, \mathbf{x}_{j_m}, m \in [|\mathcal{C}_l| - 1]$. Putting these together, we have for any $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l$, that

$$\begin{aligned} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| &\leq \max_{m \in [k]} |\mathcal{C}_l| \sqrt{\frac{nk K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m))}{K_\sigma(\delta_l)}} \\ &\leq \max_{m \in [k]} n^{1.5} k^{0.5} \sqrt{\frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m))}{K_\sigma(\delta_l)}}, \end{aligned}$$

as required. ■

What we obtain from the above is that if clusters are internally connected at smaller distances than the distances between clusters, then because of assumption AK3 we know that as $\sigma \to 0$, the ratio $K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{X}_m))/K_\sigma(\delta_l)$ converges to zero. The result for the normalised Laplacian is extremely similar, with the main difference coming from the fact that the approximate normalised cut solution is given by $\mathbf{D}^{-1/2}\mathbf{U}$, and not the eigenvectors alone.

Lemma 2 *Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and let $\mathcal{C}_1, ..., \mathcal{C}_k$ be a partition of $\mathcal{X}$. For each $l \in [k]$, suppose $\mathcal{C}_l$ is connected at distance δ_l. Let $\mathbf{U} \in \mathbb{R}^{n \times n}$ have as columns the eigenvectors of the normalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i, \mathbf{x}_j)), i, j \in [n]$, and let $\mathbf{D}$ be the corresponding degree matrix. Then for each $i, j \in [n], l \in [k]$ s.t. $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l$, we have*

$$||\mathbf{D}_{ii}^{-1/2}\mathbf{U}_{i,1:k} - \mathbf{D}_{jj}^{-1/2}\mathbf{U}_{j,1:k}|| \leq \max_{m \in [k]} n^{1.5} k^{0.5} \sqrt{\frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m))}{K_\sigma(\delta_l)}}.$$

Proof The proof is very similar to before. The fact that since spectral clustering is a relaxation of the normalised graph cut problem now gives us,

$$\begin{aligned}\sum_{i=1}^{k}\mathbf{U}_{:,i}^{\top}\mathbf{L}_{\mathrm{N}}\mathbf{U}_{:,i} = \sum_{i=1}^{k}\mathbf{U}_{:,i}^{\top}\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}\mathbf{U}_{:,i} &\leq \min_{\{C_1,...,C_k\}\in\Pi_k(\mathcal{X})}\sum_{i=1}^{k}\sum_{j,l:\mathbf{x}_j\in C_i,\mathbf{x}_l\notin C_i}\frac{K_\sigma(||\mathbf{x}_j-\mathbf{x}_l||)}{\mathrm{vol}(C_i)}\\ &\leq \sum_{i=1}^{k}\sum_{j,l:\mathbf{x}_j\in\mathcal{C}_i,\mathbf{x}_l\notin\mathcal{C}_i}\frac{K_\sigma(||\mathbf{x}_j-\mathbf{x}_l||)}{\mathrm{vol}(\mathcal{C}_i)} \leq \sum_{i=1}^{k}|\mathcal{X}\setminus\mathcal{C}_i|K_\sigma(d(\mathcal{C}_i,\mathcal{X}\setminus\mathcal{C}_i))\\ &\leq nk\max_{m\in[k]}K_\sigma(d(\mathcal{C}_m,\mathcal{X}\setminus\mathcal{C}_m)),\end{aligned}$$

where since $\mathbf{D}_{jj}\geq 1$ for each $j\in[n]$ we get $|\mathcal{C}_i|\leq\mathrm{vol}(\mathcal{C}_i)$ for all $i\in[k]$. As before we have for any $\mathbf{u}\in\mathbb{R}^n$ that

$$\mathbf{u}^{\top}\mathbf{L}\mathbf{u} = \frac{1}{2}\sum_{i,j}K_\sigma(||\mathbf{x}_i-\mathbf{x}_j||)\,(\mathbf{u}_i-\mathbf{u}_j)^2$$

$$\Rightarrow \sum_{i=1}^{k}\mathbf{U}_{:,i}^{\top}\mathbf{L}_{\mathrm{N}}\mathbf{U}_{:,i} = \sum_{i=1}^{k}\mathbf{U}_{:,i}^{\top}\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}\mathbf{U}_{:,i} = \frac{1}{2}\sum_{i,j}K_\sigma(||\mathbf{x}_i-\mathbf{x}_j||)\left\|\frac{\mathbf{U}_{i,1:k}}{\mathbf{D}_{ii}^{1/2}}-\frac{\mathbf{U}_{j,1:k}}{\mathbf{D}_{jj}^{1/2}}\right\|^2$$

$$\Rightarrow K_\sigma(||\mathbf{x}_{i_m}-\mathbf{x}_{j_m}||)\left\|\frac{\mathbf{U}_{i_m,1:k}}{\mathbf{D}_{i_mi_m}^{1/2}}-\frac{\mathbf{U}_{j_m,1:k}}{\mathbf{D}_{j_mj_m}^{1/2}}\right\|^2 \leq \sum_{i=1}^{k}\mathbf{U}_{:,i}^{\top}\mathbf{L}_{\mathrm{N}}\mathbf{U}_{:,i} \text{ for each } m\in[|\mathcal{C}_l|-1]$$

$$\Rightarrow \left\|\frac{\mathbf{U}_{i_m,1:k}}{\mathbf{D}_{i_mi_m}^{1/2}}-\frac{\mathbf{U}_{j_m,1:k}}{\mathbf{D}_{j_mj_m}^{1/2}}\right\|^2 \leq \sqrt{\frac{\sum_{i=1}^{k}\mathbf{U}_{:,i}^{\top}\mathbf{L}_{\mathrm{N}}\mathbf{U}_{:,i}}{K_\sigma(\delta_l)}} \text{ for each } m\in[|\mathcal{C}_l|-1]$$

$$\Rightarrow \left\|\frac{\mathbf{U}_{i,1:k}}{\mathbf{D}_{ii}^{1/2}}-\frac{\mathbf{U}_{j,1:k}}{\mathbf{D}_{jj}^{1/2}}\right\|^2 \leq \sqrt{\frac{\sum_{i=1}^{k}\mathbf{U}_{:,i}^{\top}\mathbf{L}_{\mathrm{N}}\mathbf{U}_{:,i}}{K_\sigma(\delta_l)}} \text{ for any } i,j \text{ s.t. } \mathbf{x}_i,\mathbf{x}_j\in\mathcal{C}_l,$$

where we have used the same pairs $(\mathbf{x}_{i_1},\mathbf{x}_{j_1}),...,(\mathbf{x}_{i_{|\mathcal{C}_l|-1}},\mathbf{x}_{j_{|C_l|-1}})$ as in the previous proof. Putting these together as before gives the result. ■

Crucial in the proof of the above result is the fact that the diagonals of $\mathbf{D}$ are bounded below by 1, since each point is linked to itself in the similarity graph. Without a fixed lower bound on the diagonal elements of $\mathbf{D}$, the bounds become weaker, as seen in the following Lemma. Additional requirements will be needed to ensure that spectral clustering recovers the maximum margin clustering solution in this case. These will be discussed explicitly in the relevant section to follow.

Lemma 3 *Let $\mathcal{X}=\{\mathbf{x}_1,...,\mathbf{x}_n\}$ and let $\mathcal{C}_1,...,\mathcal{C}_k$ be a partition of $\mathcal{X}$. For each $l\in[k]$, suppose that $\mathcal{C}_l$ is connected at distance δ_l. Let $\mathbf{U}\in\mathbb{R}^{n\times n}$ have as columns the eigenvectors of the normalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i,\mathbf{x}_j)), i,j\in[n]$, but with reflexive edges removed, and let $\mathbf{D}$ be the corresponding degree matrix. Then for each $i,j\in[n], l\in[k]$ s.t. $\mathbf{x}_i,\mathbf{x}_j\in\mathcal{C}_l$, we have*

$$||\mathbf{D}_{ii}^{-1/2}\mathbf{U}_{i,1:k}-\mathbf{D}_{jj}^{-1/2}\mathbf{U}_{j,1:k}|| \leq \max_{m\in[k]}n^{1.5}k^{0.5}\sqrt{\frac{K_\sigma(d(\mathcal{C}_m,\mathcal{X}\setminus\mathcal{C}_m))}{K_\sigma(\delta_m)K_\sigma(\delta_l)}}.$$

Proof The proof is exactly as in the previous lemma, except that now we have $\mathbf{D}_{jj} \geq K_\sigma(\delta_l)$ for all $j \in C_l$, and hence $\mathrm{vol}(\mathcal{C}_l) \geq |\mathcal{C}_l| K_\sigma(\delta_l)$ instead of $\mathrm{vol}(\mathcal{C}_l) \geq |\mathcal{C}_l|$, and hence

$$\sum_{i=1}^{k} \mathbf{U}_{:,i}^\top \mathbf{L}_{\mathrm{N}} \mathbf{U}_{:,i} \leq nk \max_{m \in [k]} \frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m))}{K_\sigma(\delta_m)}.$$

rather than

$$\sum_{i=1}^{k} \mathbf{U}_{:,i}^\top \mathbf{L}_{\mathrm{N}} \mathbf{U}_{:,i} \leq nk \max_{m \in [k]} K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m)).$$

■

Remark 4 *We note that some authors recommend placing a lower bound on the diagonals of the degree matrix to enhance the stability of the eigenvector solver being used. From a practical point of view, therefore, allowing the elements of* $\mathbf{D}$ *to approach zero may be undesirable. Note that any fixed lower bound would ensure convergence of the spectral clustering solution to the maximum margin clustering. It is still interesting, however, to investigate theoretically the requirements needed in the event that no such lower bound is in place.*

The above results place upper bounds on the within cluster distances in the eigenvalue representation, in terms of the connectedness and separation of clusters in the input space. The following general results allow us to place lower bounds on the between cluster distances within the eigenvector representation. Although not explicitly related to eigenvectors, the following proposition may be easily placed in relation to the unnormalised Laplacian, since the eigenvectors used in clustering are orthogonal. On the other hand, in the normalised solution we use the matrix $\mathbf{D}^{-1/2}\mathbf{U}$, which does not have orthogonal columns. In the first corollary to the following result we provide a more general result which admits such matrices.

Proposition 5 *Let* $\mathbf{V} \in \mathbb{R}^{n \times k}$ *have orthonormal columns, and for* $\mathbf{W} \in \mathbb{R}^{k \times k}$ *let*

$$\epsilon(\mathbf{V}, \mathbf{W}) = \max_{i \in [n]} \left\{ \min_{l \in [k]} ||\mathbf{V}_{i,:} - \mathbf{W}_{l,:}|| \right\}.$$

Then, provided $\epsilon(\mathbf{V}, \mathbf{W}) < (3nk^2)^{-1}$*, we have*

$$\min_{i,j \in [k], i \neq j} ||\mathbf{W}_{i,:} - \mathbf{W}_{j,:}|| \geq \sqrt{\frac{2}{n}} - \sqrt{12}k \left(3n\epsilon(\mathbf{V}, \mathbf{W})\right)^{1/4}.$$

Proof Take $\mathbf{W} \in \mathbb{R}^{k \times k}$ with $\epsilon = \epsilon(\mathbf{V}, \mathbf{W}) < (3nk^2)^{-1}$. For each $i \in [n]$, let $c(i) \in [k]$ be such that $||\mathbf{V}_{i,:} - \mathbf{W}_{c(i),:}|| \leq \epsilon$. Then, for each $l \in [k]$, let $n(l) = \sum_{i=1}^{n} \mathbf{1}_{[c(i)=l]}$, where $\mathbf{1}_{[A]}$ is the indicator function for A. Note that we lose no generality by assuming that $n_l \geq 1$

for each $l \in [k]$, since $\mathbf{V}$ has rank k and so contains at least k unique rows. Then define $\tilde{\mathbf{W}} = \text{diag}(\sqrt{n_1}, ..., \sqrt{n_k})\mathbf{W}$. Now consider that, for any $i \in [k]$,

$$\begin{aligned}
\sum_{l=1}^{n} \mathbf{V}_{l,i}^2 &= 1 \\
\Rightarrow \sum_{l=1}^{n} (\mathbf{V}_{l,i} - \mathbf{W}_{c(l),i} + \mathbf{W}_{c(l),i})^2 &= 1 \\
\Rightarrow \sum_{l=1}^{n} \mathbf{W}_{c(l),i}^2 + \sum_{l=1}^{n} (\mathbf{V}_{l,i} - \mathbf{W}_{c(l),i})^2 + 2\sum_{l=1}^{n} (\mathbf{V}_{l,i} - \mathbf{W}_{c(l),i})\mathbf{W}_{c(l),i} &= 1 \\
\Rightarrow \left| ||\tilde{\mathbf{W}}_{,:i}||^2 - 1 \right| = \left| \sum_{l=1}^{k} n_l \mathbf{W}_{l,i}^2 - 1 \right| = \left| \sum_{l=1}^{n} \mathbf{W}_{c(l),i}^2 - 1 \right| &\leq n\epsilon^2 + 2n\epsilon(1+\epsilon) \leq 3n\epsilon,
\end{aligned}$$

where we have used the fact that the elements of $\mathbf{V}$ are bounded between -1 and 1, and hence the elements of $\mathbf{W}$ are bounded between $-(1+\epsilon)$ and $1+\epsilon$. Similar to above, for any $i, j \in [k]$ we have,

$$\begin{aligned}
\sum_{l=1}^{n} \mathbf{V}_{l,i}\mathbf{V}_{l,j} &= 0 \\
\Rightarrow \left| \tilde{\mathbf{W}}_{:,i}^{\top} \tilde{\mathbf{W}}_{:,j} \right| = \left| \sum_{l=1}^{k} n_l \mathbf{W}_{l,i}\mathbf{W}_{l,j} \right| = \left| \sum_{l=1}^{n} \mathbf{W}_{c(l),i}\mathbf{W}_{c(l),j} \right| &\leq 3n\epsilon.
\end{aligned}$$

We therefore have $||\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}} - \mathbf{I}||_{\infty} \leq 3n\epsilon$. Weyl's inequality (Weyl, 1912) ensures that the eigenvalues of $\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}$ lie in $\left[1 - k\sqrt{3n\epsilon}, 1 + k\sqrt{3n\epsilon}\right]$. Note that $\epsilon < (3nk^2)^{-1} \Rightarrow k\sqrt{3n\epsilon} < 1$ and hence $\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}$ is non-singular. So consider the matrix $\mathbf{W}^{\star} := \tilde{\mathbf{W}}(\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1/2}$. It is simple to check that $\mathbf{W}^{\star}$ is orthogonal. Now let $||\cdot||_F$ be the Frobenius norm, and consider

$$\begin{aligned}
||\tilde{\mathbf{W}} - \tilde{\mathbf{W}}(\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1/2}||_F^2 &= ||\tilde{\mathbf{W}}(\mathbf{I} - (\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1/2})||_F^2 \\
&= \text{tr}\left(\tilde{\mathbf{W}}(\mathbf{I} - (\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1/2})(\mathbf{I} - (\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1/2})\tilde{\mathbf{W}}^{\top}\right) \\
&= \text{tr}\left(\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}(\mathbf{I} - 2(\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1/2} + (\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1})\right) \\
&= \text{tr}\left(\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}(\mathbf{I} - 2(\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1/2} + (\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{-1})\right) \\
&= \text{tr}\left(\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}}\right) - 2\text{tr}\left((\tilde{\mathbf{W}}^{\top}\tilde{\mathbf{W}})^{1/2}\right) + k.
\end{aligned}$$

We thus find

$$||\tilde{\mathbf{W}} - \mathbf{W}^{\star}||_F^2 \leq 3k^2\sqrt{3n\epsilon},$$

since the eigenvalues of $\tilde{\mathbf{W}}^\top \tilde{\mathbf{W}}$ lie in $\left[1 - k\sqrt{3n\epsilon}, 1 + k\sqrt{3n\epsilon}\right]$, and hence the eigenvalues of $(\tilde{\mathbf{W}}^\top \tilde{\mathbf{W}})^{1/2}$ also lie in $\left[1 - k\sqrt{3n\epsilon}, 1 + k\sqrt{3n\epsilon}\right]$. Finally, we have,

$$\begin{aligned}
\left\| \frac{1}{\sqrt{n_i}} \mathbf{W}^\star_{i,:} - \frac{1}{\sqrt{n_j}} \mathbf{W}^\star_{j,:} \right\| &= \left\| \frac{1}{\sqrt{n_i}} \mathbf{W}^\star_{i,:} - \mathbf{W}_{i,:} + \mathbf{W}_{i,:} - \mathbf{W}_{j,:} + \mathbf{W}_{j,:} - \frac{1}{\sqrt{n_j}} \mathbf{W}^\star_{j,:} \right\| \\
&\leq \left\| \frac{1}{\sqrt{n_i}} \mathbf{W}^\star_{i,:} - \mathbf{W}_{i,:} \right\| + \|\mathbf{W}_{i,:} - \mathbf{W}_{j,:}\| + \left\| \mathbf{W}_{j,:} - \frac{1}{\sqrt{n_j}} \mathbf{W}^\star_{j,:} \right\| \\
&= \|\mathbf{W}_{i,:} - \mathbf{W}_{j,:}\| + \frac{1}{\sqrt{n_i}} \left\| \mathbf{W}^\star_{i,:} - \tilde{\mathbf{W}}_{i,:} \right\| + \frac{1}{\sqrt{n_j}} \left\| \tilde{\mathbf{W}}_{j,:} - \mathbf{W}^\star_{j,:} \right\|
\end{aligned}$$

Therefore, since $\mathbf{W}^\star$ is orthogonal, we have

$$\begin{aligned}
||\mathbf{W}_{i,:} - \mathbf{W}_{j,:}|| &\geq \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} - \sqrt{\frac{3k^2\sqrt{3n\epsilon}}{n_i}} - \sqrt{\frac{3k^2\sqrt{3n\epsilon}}{n_j}} \\
&\geq \sqrt{\frac{2}{n}} - \sqrt{12}k(3n\epsilon)^{1/4}.
\end{aligned}$$

This proves the result.

■

Corollary 6 *Take $\mathbf{V} \in \mathbb{R}^{n\times k}$ with full column rank, and let $e_1, e_k > 0$ be respectively the smallest and largest eigenvalues of $\mathbf{V}^\top\mathbf{V}$. For each $\mathbf{W} \in \mathbb{R}^{k\times k}$, define $\epsilon(\mathbf{V}, \mathbf{W})$ as in Lemma 5. Then, provided $\epsilon(\mathbf{V}, \mathbf{W}) < \sqrt{e_1}(3nk^2)^{-1}$, we have*

$$\min_{i,j\in[k], i\neq j} ||\mathbf{W}_{i,:} - \mathbf{W}_{j,:}|| \geq e_k^{-1/2} \left(\sqrt{\frac{2}{n}} - \sqrt{12}k(3ne_1^{-1/2}\epsilon(\mathbf{V}, \mathbf{W}))^{1/4} \right).$$

Proof Take $\mathbf{W} \in \mathbb{R}^{k\times k}$ with $\epsilon = \epsilon(\mathbf{V}, \mathbf{W}) < \sqrt{e_1}(3nk^2)^{-1}$. Then let $\Sigma = (\mathbf{V}^\top\mathbf{V})^{-1}$, so that $\mathbf{V}\Sigma^{1/2}$ is orthonormal. Now take any $i \in [n], l \in [k]$, then

$$||\mathbf{V}_{i,:}\Sigma^{1/2} - \mathbf{W}_{l,:}\Sigma^{1/2}||^2 \leq \frac{1}{e_1}||\mathbf{V}_{i,:} - \mathbf{W}_{l,:}||^2.$$

Therefore $\mathbf{W}\Sigma^{1/2}$ satisfies,

$$\max_{i\in[n]} \left\{ \min_{l\in[k]} ||\mathbf{V}_{i,:}\Sigma^{1/2} - \mathbf{W}_{l,:}\Sigma^{1/2}|| \right\} \leq e_1^{-1/2}\epsilon.$$

Thus we can apply Proposition 5 to see that for any $i, j \in [k], i \neq j$, we have

$$\begin{aligned}
&||\mathbf{W}_{i,:}\Sigma^{1/2} - \mathbf{W}_{j,:}\Sigma^{1/2}|| \geq \sqrt{\frac{2}{n}} - \sqrt{12}k(3ne_1^{-1/2}\epsilon)^{1/4} \\
\Rightarrow &||\mathbf{W}_{i,:} - \mathbf{W}_{j,:}|| \geq e_k^{-1/2} \left(\sqrt{\frac{2}{n}} - \sqrt{12}k(3ne_1^{-1/2}\epsilon)^{1/4} \right).
\end{aligned}$$

■

By viewing the matrix $\mathbf{W}$ in the above two results as the centroids from a clustering of the rows of $\mathbf{V}$, these effectively place bounds on the distances between cluster centroids. In particular, if the rows of $\mathbf{V}$ are well clusterable in the sense that there is a set of k centroids for which each row of $\mathbf{V}$ lies very close to its nearest centroid, then these centroids cannot be too close to one another. These results hold because the number of dimensions (number of columns of $\mathbf{V}$) is equal to or fewer than the number of centroids. If we try to extend this to the case where the number of dimensions is greater than the number of centroids, then the matrix of centroids cannot have full column rank. This means that if $\mathbf{V} \in \mathbb{R}^{n\times(k+1)}$ has full column rank, then it cannot be arbitrarily well clusterable by the rows of any $\mathbf{W} \in \mathbb{R}^{k\times(k+1)}$. We first formalise the above for the case where $\mathbf{V}$ has orthonormal columns, and then extend it to the general case.

Corollary 7 *Let $\mathbf{V} \in \mathbb{R}^{n\times(k+1)}$ have orthonormal columns. Then for all $\mathbf{W} \in \mathbb{R}^{k\times(k+1)}$ we have*

$$\max_{i\in[n]}\left\{\min_{l\in[k]}||\mathbf{V}_{i,:}-\mathbf{W}_{l,:}||\right\} \geq (3n(k+1)^2)^{-1}.$$

Proof Suppose that the result does not hold, i.e., that there is a $\mathbf{W} \in \mathbb{R}^{k\times(k+1)}$ s.t. $\epsilon(\mathbf{V},\mathbf{W}) < (3n(k+1)^2)^{-1}$, for $\epsilon(\mathbf{V},\mathbf{W})$ as in the previous two lemmas. Then let $\tilde{\mathbf{W}} \in \mathbb{R}^{(k+1)\times(k+1)}$ have as its first k rows the rows of $\mathbf{W}$, and as its last row $\mathbf{W}_{l,:}$ where $\mathbf{W}_{l,:}$ is within $\epsilon(\mathbf{V},\mathbf{W})$ distance of at least two rows of $\mathbf{V}$. Such a $\ell \in [k]$ must exist since $\mathbf{V}$ has at least $k+1$ unique rows. Then $\tilde{\mathbf{W}}$ satisfies the conditions of Proposition 5. This would imply that $||\tilde{\mathbf{W}}_{l,:}-\tilde{\mathbf{W}}_{k+1,:}|| > 0$, a contradiction. Therefore $\epsilon(\mathbf{V},\mathbf{W}) \geq (3n(k+1)^2)^{-1}$. ■

As before, the result can be extended to the case where the columns of $\mathbf{V}$ need not be orthogonal.

Corollary 8 *Let $\mathbf{V} \in \mathbb{R}^{n\times(k+1)}$ have full column rank and let $e_1 > 0$ be the smallest eigenvalue of $\mathbf{V}^\top\mathbf{V}$. Then for all $\mathbf{W} \in \mathbb{R}^{k\times(k+1)}$ we have*

$$\max_{i\in[n]}\left\{\min_{l\in[k]}||\mathbf{V}_{i,:}-\mathbf{W}_{l,:}||\right\} \geq \sqrt{e_1}(3n(k+1)^2)^{-1}.$$

Proof The proof is exactly analogous to the above proof, where the contradiction now arises by applying Corollary 6 to the extended matrix $\tilde{\mathbf{W}}$. ■

These final two results will be useful for deriving lower bounds on the eigenvalues of graph Laplacians, which are presented in the following subsection.

5.2. Eigenvalue Bounds for Graph Laplacians

The bounds presented in the previous subsection deal with the structure of the data within the Laplacian eigenvector representation. These can be used to ensure recovery of the

maximum margin clustering solution, as discussed in the next subsection. However, no attention has yet been given to selecting the number of clusters. It is generally understood that the relative values of the eigenvalues of graph Laplacians might be useful in determining the number of clusters in a data set (Von Luxburg, 2007). In this section we derive upper and lower bounds on these eigenvalues. These bounds will be used later to show that the number of components of $\mathcal{L}(\lambda)$ can be consistently estimated using the eigenvalues of the graph Laplacians computed from truncations of an increasing sample arising from an assumed underlying probability distribution. The bounds are simply derived, and similar for the different Laplacian matrices. For completeness, we state them all explicitly.

Lemma 9 *Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and let $\mathcal{C}_1, ..., \mathcal{C}_k$ be a partition of $\mathcal{X}$. For each $l \in [k]$ suppose that $\mathcal{C}_l$ is connected at distance δ_l. For each $l \in [n]$ let e_l be the l-th eigenvalue of the unnormalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i, \mathbf{x}_j)), i, j \in [n]$. Then,*

$$\sum_{l=1}^{k} e_l \leq nk \max_{m \in [k]} K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m)),$$

$$e_{k+1} \geq \min_{l,m \in [k]} \frac{K_\sigma(\delta_l)}{9n^2(k+1)^4} - n^3 k K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{X}_m)).$$

Proof Let $\mathbf{U}$ be the eigenvectors of the unnormalised Laplacian of the similarity graph. The upper bound on the sum of the first k eigenvalues follows from the beginning of the proof of Lemma 1, since $\sum_{l=1}^{k} e_l = \sum_{l=1}^{k} \mathbf{U}_{:,l}^\top \mathbf{L} \mathbf{U}_{:,l}$. Now, by Corollary 7 we know that $\exists i, j \in [n], l \in [k]$ with $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l$, such that

$$||\mathbf{U}_{i,:1:(k+1)} - \mathbf{U}_{j,:1:(k+1)}||^2 \geq (3n(k+1)^2)^{-2}.$$

But from Lemma 1 we know that

$$||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}||^2 \leq \max_{m \in [k]} n^3 k \frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m))}{K_\sigma(\delta_l)},$$

and thus,

$$(\mathbf{U}_{i,k+1} - \mathbf{U}_{j,k+1})^2 \geq (3n(k+1)^2)^{-2} - \max_{m \in [k]} n^3 k \frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m))}{K_\sigma(\delta_l)}.$$

Again using (Von Luxburg, 2007, Proposition 1), we have

$$\begin{aligned} e_{k+1} = \mathbf{U}_{:,k+1}^\top \mathbf{L} \mathbf{U}_{:,k+1} &= \frac{1}{2} \sum_{g,h} K_\sigma(||\mathbf{x}_g - \mathbf{x}_h||)(\mathbf{U}_{g,k+1} - \mathbf{U}_{h,k+1})^2 \\ &\geq K_\sigma(\delta_l)(\mathbf{U}_{i,k+1} - \mathbf{U}_{j,k+1})^2 \\ &\geq \min_{l,m \in [k]} \frac{K_\sigma(\delta_l)}{9n^2(k+1)^4} - n^3 k K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{X}_m)). \end{aligned}$$

■

Lemma 10 *Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and let $\mathcal{C}_1, ..., \mathcal{C}_k$ be a partition of $\mathcal{X}$. For each $l \in [k]$ suppose that $\mathcal{C}_l$ is connected at distance δ_l. For each $l \in [n]$ let e_l be the l-th eigenvalue of the normalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i, \mathbf{x}_j)), i, j \in [n]$. Then,*

$$\sum_{l=1}^{k} e_l \leq nk \max_{m \in [k]} K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m)),$$
$$e_{k+1} \geq \min_{l,m \in [k]} \frac{K_\sigma(\delta_l)}{9n^3(k+1)^4} - n^3 k K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{X}_m)).$$

Proof Let $\mathbf{U}$ be the eigenvectors of the normalised Laplacian of the similarity graph. The upper bound on the sum of the first k eigenvalues now follows immediately from the first part of the proof of Lemma 2. Unlike in the previous proof, we cannot use Corollary 7 since $\mathbf{D}^{-1/2}\mathbf{U}_{:,1:(k+1)}$ is not orthogonal. However, observe that,

$$\begin{aligned}
\min_{\mathbf{v} \in \mathbb{R}^d} \frac{\mathbf{v}^\top \mathbf{U}_{:,1:(k+1)}^\top \mathbf{D}^{-1} \mathbf{U}_{:,1:(k+1)} \mathbf{v}}{\mathbf{v}^\top \mathbf{v}} &= \min_{\mathbf{v} \in \mathbb{R}^d} \frac{\mathbf{v}^\top \mathbf{U}_{:,1:(k+1)}^\top \mathbf{D}^{-1} \mathbf{U}_{:,1:(k+1)} \mathbf{v}}{\mathbf{v}^\top \mathbf{U}_{:,1:(k+1)}^\top \mathbf{U}_{:,1:(k+1)} \mathbf{v}} \\
&\geq \min_{\mathbf{u} \in \mathbb{R}^n} \frac{\mathbf{u}^\top \mathbf{D}^{-1} \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} \\
&= \min_{i \in [n]} \mathbf{D}_{i,i}^{-1} \geq \frac{1}{n}.
\end{aligned}$$

That is, the smallest eigenvalue of $\mathbf{U}_{:,1:(k+1)}^\top \mathbf{D}^{-1} \mathbf{U}_{:,1:(k+1)}$ is at least n^{-1}, and so by Corollary 8 we know there exist $i, j \in [n], l \in [k]$ with $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l$ such that,

$$||\mathbf{D}_{i,i}^{-1/2} \mathbf{U}_{i,1:(k+1)} - \mathbf{D}_{j,j}^{-1/2} \mathbf{U}_{j,1:(k+1)}||^2 \geq \frac{1}{9} n^{-3} (k+1)^{-4}.$$

The rest of the proof is analogous to the previous proof. ■

Lemma 11 *Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and let $\mathcal{C}_1, ..., \mathcal{C}_k$ be a partition of $\mathcal{X}$. For each $l \in [k]$ suppose that $\mathcal{C}_l$ is connected at distance δ_l. For each $l \in [n]$ let e_l be the l-th eigenvalue of the normalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i, \mathbf{x}_j)), i, j \in [n]$, but with reflexive edges removed. Then,*

$$\sum_{l=1}^{k} e_l \leq nk \max_{m \in [k]} \frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m))}{K_\sigma(\delta_m)},$$
$$e_{k+1} \geq \min_{l,m \in [k]} \frac{K_\sigma(\delta_l)}{9n^3(k+1)^4} - n^3 k \frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{X}_m))}{K_\sigma(\delta_m)}.$$

Proof The proof is exactly as above, but using the bound from Lemma 3 instead of Lemma 2. ■

5.3. Maximum Margins from Graph Laplacians

In this section we provide a detailed derivation of the convergence of spectral clustering to the maximum margin clustering solution. The results for the unnormalised Laplacian, $\mathbf{L}$, and normalised Laplacian, $\mathbf{L}_{\mathrm{N}}$, follow from assumptions AK1–AK3 on the kernel, K, and the results from Section 5.1. Convergence of the spectral clustering solution arising from the normalised Laplacian of the similarity graph without reflexive edges, $\mathbf{L}_{\mathrm{N}_0}$, requires stronger assumptions. To ease the analysis, we assume in this case that the kernel function is given simply by $K(x) = \exp(-x^\alpha)$ for some $\alpha > 0$. We also require that the within cluster connectedness is sufficiently below the between cluster separatedness, in terms of the parameter α. While in the scenario of level set estimation from an increasing sample, this is not a problem since the within cluster connectivity converges to zero as the sample size increases. In the finite sample setting, however, without this additional requirement, convergence instead occurs as $\sigma \to 0$ and $\alpha \to \infty$, rather than only requiring that $\sigma \to 0$ as in the cases of $\mathbf{L}$ and $\mathbf{L}_{\mathrm{N}}$.

We state the results of this section in a form which is convenient for the consistency analysis given in the following section, where convergence occurs as the sample size, n, increases. Broadly speaking, the results show that for $\sigma < A\log(Bn^z)^{-C}$, where $A, B, C > 0$ are independent of n, and any z sufficiently large, we have that the ratio of within cluster distances to between cluster distances, in the eigenvector representation, are $\mathcal{O}(n^{D-Ez})$ for constants $D, E > 0$. This means that in the finite sample setting, where n is fixed, as σ approaches zero (and thus z increases towards ∞), the maximum margin clustering solution becomes trivially attainable from the eigenvectors. Once again we investigate each Laplacian matrix separately.

Theorem 12 *Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and assume that there is a partition of $\mathcal{X}$ into $\mathcal{C}_1, ..., \mathcal{C}_k$ such that $\min_{m\in[k]} d(\mathcal{C}_m, \mathcal{X}\setminus\mathcal{C}_m) - \max_{l\in[k]}\delta_l = \delta > 0$, where $\mathcal{C}_l$ is connected at distance δ_l for each $l \in [k]$. For $\sigma > 0$ let $\mathbf{L}$ be the unnormalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i, \mathbf{x}_j)), i, j \in [n]$, where K satisfies assumptions AK1–AK3. Let $\mathbf{U}$ have as columns the eigenvectors of $\mathbf{L}$. Then, provided $0 < \sigma < \delta\log(An^{z/3})^{-1/\alpha}$, where A and α are as in assumption AK3, and z satisfies $n^{z-15} \geq 81k^{15}$, we have*

$$\max_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i,\mathbf{x}_j\in\mathcal{C}_l}} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| \leq \left(\frac{k^3}{n^{z-9}}\right)^{1/6},$$

$$\min_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i\in\mathcal{C}_l,\mathbf{x}_j\notin\mathcal{C}_l}} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| \geq \sqrt{\frac{2}{n}} - 6\left(\frac{k^{27}}{n^{z-15}}\right)^{1/24}.$$

Proof First, combining Lemma 1 with assumption AK3, we get

$$\begin{aligned}\max_{i,j\in[n],l\in[k]:\mathbf{x}_i,\mathbf{x}_j\in\mathcal{C}_l} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}||^2 &\leq \max_{m\in[k]} n^3 k\frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X}\setminus\mathcal{C}_m))}{K_\sigma(\delta_l)} \leq An^3k\exp\left(-\left(\frac{\delta}{\sigma}\right)^\alpha\right)\\ &\leq An^3k\exp\left(-\log(An^{z/3})\right) = \frac{k}{n^{(z-9)/3}},\end{aligned}$$

as required. Now let $i_1, ..., i_k$ be such that $\mathbf{x}_{i_l} \in \mathcal{C}_l$ for each $l \in [k]$. Then for σ below the assumed upper bound the matrix $[\mathbf{U}_{i_1,1:k}, ..., \mathbf{U}_{i_k,1:k}]$ satisfies the conditions on the matrix $\mathbf{W}$ in Proposition 5. Therefore,

$$\begin{aligned}
&\min_{\substack{l,m\in[k],\\ l\neq m}} ||\mathbf{U}_{i_l,1:k} - \mathbf{U}_{i_m,1:k}|| = \\
&\qquad \min_{\substack{i,j\in[n],l,m\in[k],\\ l\neq m:\\ \mathbf{x}_i\in\mathcal{C}_l,\mathbf{x}_j\in\mathcal{C}_m}} ||\mathbf{U}_{i_l,1:k} - \mathbf{U}_{i,1:k} + \mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k} + \mathbf{U}_{j,1:k} - \mathbf{U}_{i_m,1:k}|| \\
&\qquad \leq \min_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i\in\mathcal{C}_l,\mathbf{x}_j\notin\mathcal{C}_l}} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| + 2\sqrt{An^3k}\exp\left(-\frac{1}{2}\left(\frac{\delta}{\sigma}\right)^\alpha\right) \\
\Rightarrow \min_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i\in\mathcal{C}_l,\mathbf{x}_j\notin\mathcal{C}_l}} &||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| \geq \sqrt{\frac{2}{n}} - \sqrt{12}k\left(3n\sqrt{An^3k}\exp\left(-\delta^\alpha/2\sigma^\alpha\right)\right)^{1/4} \\
&\qquad - 2\sqrt{An^3k}\exp\left(-\frac{1}{2}\left(\frac{\delta}{\sigma}\right)^\alpha\right).
\end{aligned}$$

Now, it is simple to verify that the assumption on the value of σ ensures that the second two terms on the right hand side above sum to less than $4k\left(3n\sqrt{An^3k}\exp\left(-\delta^\alpha/2\sigma^\alpha\right)\right)^{1/4}$, for any $n \geq 2$. Therefore,

$$\begin{aligned}
\min_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i\in\mathcal{C}_l,\mathbf{x}_j\notin\mathcal{C}_l}} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| &\geq \sqrt{\frac{2}{n}} - 4k\left(3n\sqrt{An^3k}\exp\left(-\delta^\alpha/2\sigma^\alpha\right)\right)^{1/4} \\
&\geq \sqrt{\frac{2}{n}} - 6k^{9/8}n^{-(z-15)/24},
\end{aligned}$$

with the last step coming from simple rearrangement after substituting in the upper bound for σ.

■

Remark 13 *The above result assumes that the within cluster connectedness is strictly less than the between cluster separatedness. This occurs with probability one if $\mathcal{X}$ is seen as a sample of realisations of a continuous random variable on $\mathbb{R}^d$.*

Stating the bounds in the above theorem in terms of n is convenient for the theory presented in the next subsection. However, it can be seen directly from the above that

$$\max_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i,\mathbf{x}_j\in\mathcal{C}_l}} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}|| \xrightarrow{as\ \sigma\to 0^+} 0,$$

while the lower bound on the term $\min_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i\in\mathcal{C}_l,\mathbf{x}_j\notin\mathcal{C}_l}} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}||$ converges to $\sqrt{2/n}$ as $\sigma \to 0^+$. The maximum margin clustering solution is therefore trivially obtained from the limit of the spectral clustering solution using the unnormalised Laplacian. The corresponding result to Lemma 12 for the normalised Laplacian requires only minor modifications.

Theorem 14 *Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and assume that there is a partition of $\mathcal{X}$ into $\mathcal{C}_1, ..., \mathcal{C}_k$ such that $\min_{m\in[k]} d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m) - \max_{l\in[k]} \delta_l = \delta > 0$, where $\mathcal{C}_l$ is connected at distance δ_l for each $l \in [k]$. For each $\sigma > 0$ let $\mathbf{L}_{\mathrm{N}}$ be the normalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i, \mathbf{x}_j)), i, j \in [n]$, where K satisfies assumptions AK1–AK3, and let $\mathbf{D}$ be the corresponding degree matrix. Let $\mathbf{U}$ be the eigenvectors of $\mathbf{L}_{\mathrm{N}}$. Then, for $0 < \sigma < \delta \log(An^{z/3})^{-1/\alpha}$, where A and α are as in assumption AK3, and z satisfies $n^{z-18} \geq 81k^{15}$, we have*

$$\max_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i,\mathbf{x}_j\in\mathcal{C}_l}} ||\mathbf{D}_{i,i}^{-1/2}\mathbf{U}_{i,1:k} - \mathbf{D}_{j,j}^{-1/2}\mathbf{U}_{j,1:k}|| \leq \left(\frac{k^3}{n^{z-9}}\right)^{1/6},$$

$$\min_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i\in\mathcal{C}_l,\mathbf{x}_j\notin\mathcal{C}_l}} ||\mathbf{D}_{i,i}^{-1/2}\mathbf{U}_{i,1:k} - \mathbf{D}_{j,j}^{-1/2}\mathbf{U}_{j,1:k}|| \geq \sqrt{\frac{2}{n}} - 6\left(\frac{k^{27}}{n^{z-18}}\right)^{1/24}.$$

Proof The proof is similar to before, except that now we cannot use Proposition 5, since $\mathbf{D}^{-1/2}\mathbf{U}_{:,1:k}$ is not orthogonal. As in the proof of Lemma 10, we know that the smallest eigenvalue of $\mathbf{U}_{:,1:k}^\top \mathbf{D}^{-1}\mathbf{U}_{:,1:k}$ is at least n^{-1}. Similarly, since all diagonal elements of $\mathbf{D}$ are at least one, the largest eigenvalue is at most 1. We now have, therefore, using Corollary 6, that

$$\min_{\substack{i,j\in[n],l\in[k]:\\ \mathbf{x}_i\in\mathcal{C}_l,\mathbf{x}_j\notin\mathcal{C}_l}} ||\mathbf{D}_{i,i}^{-1/2}\mathbf{U}_{i,1:k} - \mathbf{D}_{j,j}^{-1/2}\mathbf{U}_{j,1:k}|| \geq \sqrt{\frac{2}{n}} - 4k\left(3n^{3/2}\sqrt{An^3k}\exp\left(-\delta^\alpha/2\sigma^\alpha\right)\right)^{1/4},$$

with a slightly higher power of n in the second term on the right hand side than we had before. The rest of the proof is exactly analogous to the previous proof. ■

In the above two results, we obtained explicit lower bounds on the between cluster distances within the eigenvector representations. Combining this with the fact that the within cluster distances converge to zero, the recovery of the maximum margin clustering solution from these eigenvectors is immediate. In the case where the diagonal elements of the affinity matrix are set to zero, however, we are not able to place a lower bound on the between cluster distances within the eigenvectors. We therefore only show that in this case the within cluster distances converge to zero at a much faster rate than the between cluster distances, as σ tends to zero. This is a slightly weaker result, but still ensures the maximum margin solution arises trivially from the eigenvectors of the normalised Laplacian.

As mentioned previously, we also require additional assumptions. We study the case where the kernel takes the explicit form of $K(x) = \exp(-x^\alpha)$ for some $\alpha > 0$. This allows us to generalise assumption AK3 to allow not only ratios of individual kernel values, but also fractions involving multiple such kernel evaluations. We also require a stricter assumption on the relationship between the within cluster connectedness and between cluster separation than was used previously. This is made explicit in the statement of the result below.

Theorem 15 *Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and assume that there is a partition of $\mathcal{X}$ into $\mathcal{C}_1, ..., \mathcal{C}_k$ such that $\min_{l\in[k]} |\mathcal{C}_l| \geq 2$, $\min_{m\in[k]} d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m)^\alpha - 3\max_{l\in[k]} \delta_l^\alpha = \delta > 0$, where $\mathcal{C}_l$ is*

connected at distance δ_l for each $l \in [k]$ and α is given in the formulation of the kernel, K, which follows. For each $\sigma > 0$ let $\mathbf{L}_{\mathrm{N}_0}$ be the normalised Laplacian of the similarity graph of $\mathcal{X}$ with pairwise similarities given by $K_\sigma(d(\mathbf{x}_i, \mathbf{x}_j)), i, j \in [n]$, but with reflexive edges removed, where $K(x) = \exp(-x^\alpha)$ for some $\alpha > 0$, and let $\mathbf{D}$ be the corresponding degree matrix. Let $\mathbf{U}$ be the eigenvectors of $\mathbf{L}_{\mathrm{N}_0}$. Then, for $0 < \sigma < \delta^{1/\alpha} \log\left(13^8 k^9 n^z\right)^{-1/\alpha}$, where $z \geq 10$, we have

$$\max_{\substack{i,j,g,h \in [n], l,m \in [k]: \\ \mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l \\ \mathbf{x}_g \in \mathcal{C}_m, \mathbf{x}_h \notin \mathcal{C}_m}} \frac{||\mathbf{D}_{i,i}^{-1/2}\mathbf{U}_{i,1:k} - \mathbf{D}_{j,j}^{-1/2}\mathbf{U}_{j,1:k}||}{||\mathbf{D}_{g,g}^{-1/2}\mathbf{U}_{g,1:k} - \mathbf{D}_{h,h}^{-1/2}\mathbf{U}_{h,1:k}||} \leq n^{(4-z)/2}.$$

Proof First let $\delta_\star = \max_{l \in [k]} \delta_l$. Combining Lemma 3 with assumptions on K, we get

$$\begin{aligned}\max_{i,j \in [n], l \in [k]: \mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l} ||\mathbf{U}_{i,1:k} - \mathbf{U}_{j,1:k}||^2 &\leq \max_{m \in [k]} n^3 k \frac{K_\sigma(d(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m))}{K_\sigma(\delta_l) K_\sigma(\delta_m)} \\ &\leq n^3 k \exp\left(-\frac{\delta_\star^\alpha + \delta}{\sigma^\alpha}\right).\end{aligned}$$

By assumption, points within any clusters containing at least two points are within $\delta_\star$ of their nearest neighbours. The lower bound on the diagonals of the degree matrix is therefore now $K_\sigma(\delta_\star)$, instead of 1 as before. The largest eigenvalue of $\mathbf{U}_{:,1:k}^\top \mathbf{D}^{-1} \mathbf{U}_{:,1:k}^\top$ is thus at most $\max_{i \in [n]} \mathbf{D}_{i,i}^{-1} \leq K_\sigma(\delta_\star)^{-1} = \exp(\delta_\star^\alpha / \sigma^\alpha)$. The smallest eigenvalue is again at least n^{-1}. Therefore in this case we have, using Lemma 6, and after simple rearranging,

$$\begin{aligned}\min_{\substack{i,j \in [n], l \in [k]: \\ \mathbf{x}_i \in \mathcal{C}_l, \mathbf{x}_j \notin \mathcal{C}_l}} ||\mathbf{D}_{i,i}^{-1/2}\mathbf{U}_{i,1:k} - \mathbf{D}_{j,j}^{-1/2}\mathbf{U}_{j,1:k}|| &\geq \sqrt{\frac{2}{n}} \exp\left(-\frac{\delta_\star^\alpha}{2\sigma^\alpha}\right) - \sqrt{12} k^{9/8} n^{3/4} \exp\left(-\frac{5\delta_\star^\alpha + \delta}{8\sigma^\alpha}\right) \\ &\quad - 2n^{3/2} k^{1/2} \exp\left(-\frac{\delta_\star^\alpha + \delta}{2\sigma^\alpha}\right).\end{aligned}$$

Now, the upper bound on σ ensures that the above difference is at least $n^{-1/2} \exp(-\delta_\star^\alpha / 2\sigma^\alpha)$. Putting these together, we get, after simplification,

$$\begin{aligned}\max_{\substack{i,j,g,h \in [n], l,m \in [k]: \\ \mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_l \\ \mathbf{x}_g \in \mathcal{C}_m, \mathbf{x}_h \notin \mathcal{C}_m}} \frac{||\mathbf{D}_{i,i}^{-1/2}\mathbf{U}_{i,1:k} - \mathbf{D}_{j,j}^{-1/2}\mathbf{U}_{j,1:k}||}{||\mathbf{D}_{g,g}^{-1/2}\mathbf{U}_{g,1:k} - \mathbf{D}_{h,h}^{-1/2}\mathbf{U}_{h,1:k}||} &\leq \frac{n^{3/2} k^{1/2} \exp\left(-\frac{\delta_\star^\alpha + \delta}{2\sigma^\alpha}\right)}{n^{-1/2} \exp\left(-\frac{\delta_\star^\alpha}{2\sigma^\alpha}\right)} \\ &\leq n^2 \exp\left(-\frac{\delta}{2\sigma^\alpha}\right) \leq n^{(4-z)/2}.\end{aligned}$$

■

Remark 16 *The stricter assumption in the previous theorem may appear to be stated only for the convenience of making the theorem hold, rather than being practically relevant. However, consider that if $0 < a < b$ then there is a $\alpha > 0$ s.t. $b^\alpha > 3a^\alpha$. From a practical point of*

view, therefore, if we were to determine sequences of similarities using $\exp(-||\mathbf{x}_i-\mathbf{x}_j||^\alpha/\sigma^\alpha)$ *where* $\sigma \to 0$ *and* $\alpha \to \infty$*, then the convergence of the spectral clustering solution to the maximum margin solution would hold under the same assumptions as in Theorems 12 and 14. In addition, in the situation where the within cluster connectedness decreases appropriately towards zero as n increases, while the between cluster separation is bounded below, then for any fixed value of* α *the above theorem takes effect. This will be relevant in the following section, where we study the behaviour of the spectral clustering solution applied to a truncated sample, as the size of the sample increases.*

5.4. Consistently Estimating Level Set Components using Spectral Clustering

In this section we study the estimation of level sets, and their components, using spectral clustering. Unlike existing work on this problem (Pelletier and Pudlo, 2011), we study multiple versions of spectral clustering. Specifically, those arising from the relaxations of the Ratio Cut problem, and two similar versions of the Normalised Cut problem. In addition, we consider kernels with unbounded support, and we also consider what we believe is a more desirable and natural context where the scaling parameter is decreased towards zero as the sample size increases. This means that our estimation procedure requires weaker assumptions than the existing theory. Importantly, the minimum distance between components of the target level set need not be known. Furthermore, the requirements on the rate of decrease of the scaling parameter which we require admits the asymptotically optimal mean integrated squared error (MISE) rate for the related problem of kernel density estimation. This adds a superficial (and minor computational) benefit, which is that the same similarities used in the spectral clustering algorithm may be used to estimate the density, and hence level set as well. In particular, the quantities $\frac{c_K}{n\sigma_n^d}\sum_{i=1}^n K(||X_i-X_j||/\sigma_n), j \in [n]$, form standard kernel based estimates for the values $p(X_j), j \in [n]$, and so the collection of $K(||X_i - X_j||/\sigma_n), i, j \in [n]$, provides both the pairwise similarities as well as the collection of points whose estimated densities lie above any chosen threshold.

Suppose that $X_1, X_2, ...$ is a sequence of i.i.d. random variables on $\mathbb{R}^d$ with distribution admitting density p, which satisfies assumptions A1 and A2 for level $\lambda > 0$. Suppose also that the kernel, K, satisfies assumptions AK1–AK3. We begin by deriving some connectivity properties of the elements of $X_1, X_2, ..., X_n$ which lie in an estimate of $\mathcal{L}(\lambda)$, say $\widehat{\mathcal{L}(\lambda)}^{(n)}$, in relation to its components. To that end, if the level set $\mathcal{L}(\lambda)$ has c components, then we will use $\ell(\lambda, 1), ..., \ell(\lambda, c)$ to denote these components, listed according to any arbitrary ordering. What we show is that, with probability one, points arising in a shrinking sequence of neighbourhoods of one of the components are connected at distances approximately $\mathcal{O}(\sigma_n)$, where $\{\sigma_n\}_{n=1}^\infty$ is an appropriately chosen sequence of scaling parameters. Furthermore, with probability one, no points outside these neighbourhoods of the components of $\mathcal{L}(\lambda)$ are included in $\widehat{\mathcal{L}(\lambda)}^{(n)}$, for large values of n. This second point ensures that the level set itself is consistently estimated by taking shrinking neighbourhoods around the points in $\widehat{\mathcal{L}(\lambda)}^{(n)}$. We go on to show that the sequence $\{\sigma_n\}_{n=1}^\infty$ can be chosen so that the first c eigenvectors of the Laplacian matrices of the similarity graph of $\widehat{\mathcal{L}(\lambda)}^{(n)}$, using similarity kernel K_{σ_n}, trivially expose the separation of $\widehat{\mathcal{L}(\lambda)}^{(n)}$ into the subsets falling in the shrinking

neighbourhoods of the different components of $\mathcal{L}(\lambda)$, mentioned above. Finally, the same sequence of scaling parameters leads to the eigenvalues of these Laplacians allowing for consistent estimation of c.

Lemma 17 *Let $X_1, X_2, ...$ be an i.i.d. sequence of random variables on $\mathbb{R}^d$, $d \geq 2$, with density p satisfying assumptions A1–A2 for level $\lambda > 0$. Let $K : \mathbb{R}^+ \to \mathbb{R}^+$ satisfy assumptions AK1–AK3. Let $\xi > 1$ and let $\{\sigma_n\}_{n=1}^{\infty}$ and $\{S_n\}_{n=1}^{\infty}$ be positive sequences satisfying,*

- $\lim_{n\to\infty} \sigma_n = \lim_{n\to\infty} S_n = 0$.
- *For all large n we have*

$$\max\left\{ n^{-1/d}\log(n), \left(\frac{\log(n^{1/d}/\log(n))}{nS_n^{2\xi}}\right)^{1/d}\right\} < \sigma_n < \min\{\log(n)^{-\log(\log(n))}, S_n^{\xi}\}.$$

- *$\exists c > 0$ s.t. $\sigma_n \leq c\sigma_{2n}$ for all n.*

Then there exists a sequence $\{a_n\}_{n=1}^{\infty}$ with $a_n = O(S_n)$ such that if we define, for each $k \in [c]$ and $n \in \mathbb{N}$,

$$\widehat{\ell(\lambda,k)}^{(n)} = \left\{ X_j \middle| j \leq n, \frac{c_K}{n\sigma_n^d}\sum_{i=1}^{n} K(||X_i - X_j||/\sigma_n) > \lambda - S_n, X_j \in \mathcal{B}_{a_n}(\ell(\lambda,k))\right\},$$

then with probability one, for all n sufficiently large, we have,

1. *$\widehat{\ell(\lambda,k)}^{(n)}$ is connected at distance a_n, for all $k \in [c]$,*
2. *$\ell(\lambda,k) \subset \mathcal{B}_{\frac{a_n}{2}}(\widehat{\ell(\lambda,k)}^{(n)}) \subset \mathcal{B}_{a_n}(\ell(\lambda,k))$ for all $k \in [c]$,*
3. *$\max\left\{ \frac{c_K}{n\sigma_n^d}\sum_{i=1}^{n} K(||X_i - X_j||/\sigma_n) \middle| j \leq n, X_j \notin \bigcup_{k\in[c]} \widehat{\ell(\lambda,k)}^{(n)}\right\} \leq \lambda - S_n$.*

Proof The assumptions on p, K and $\{\sigma_n\}_{n=1}^{\infty}$ satisfy the conditions for the uniform convergence of the density estimator,

$$\hat{p}_n(\mathbf{x}) := \frac{c_K}{n\sigma_n^d}\sum_{i=1}^{n} K\left(\frac{||\mathbf{x} - X_i||}{\sigma_n}\right).$$

In particular, Giné and Guillou (2002) have shown that, for all sufficiently large n, we have

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \left|\hat{p}_n(\mathbf{x}) - E[\hat{p}_n(\mathbf{x})]\right| \leq B\sqrt{-\frac{\log(\sigma_n)}{n\sigma_n^d}},$$

for B not dependent on n or $\{\sigma_n\}_{n=1}^{\infty}$. The conditions on the sequence $\{\sigma_n\}_{n=1}^{\infty}$ in the statement of the lemma are largely dictated by the requirements of their result. Another requirement of the result, stated in a sufficient form which is appropriate for our context, is that $K(\cdot)$ can be expressed as $f(|\text{poly}(\cdot)|)$, where f is a function of bounded variation on

$\mathbb{R}_+$, and poly$(\cdot)$ denotes a polynomial. This condition is easily guaranteed by the fact that $K(\cdot)$ is bounded and non-increasing in the magnitude of its argument, and hence is itself of bounded variation.

In addition, note that the bias of the standard kernel density estimator, using bandwidth σ_n, for a density with bounded first derivative is $\mathcal{O}(\sigma_n)$ (Rosenblatt, 1991). Combining these, there is therefore a constant B' such that with probability one, for all n sufficiently large,

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |\hat{p}_n(\mathbf{x}) - p(\mathbf{x})| \leq B' \left(\sqrt{-\frac{\log(\sigma_n)}{n\sigma_n^d}} + \sigma_n \right) =: g_n.$$

It is easy to check that the assumptions on the sequences $\{\sigma_n\}_{n=1}^{\infty}$ and $\{S_n\}_{n=1}^{\infty}$ ensure that $g_n = o(S_n)$ as $n \to \infty$. Now take any $0 < \epsilon < 1 - \frac{1}{\xi}$. Then for all n large enough we have the above as well as the following,

N1: $g_n + \kappa\sigma_n^{1-\epsilon} < S_n$, where κ is as in Assumption A1.

N2: $S_n + g_n < \gamma$, where γ is as in assumption A2.

N3: $\frac{\sigma_n^d}{c_K}(\lambda - g_n - S_n) - A\exp(-\sigma_n^{-\alpha\epsilon}) \geq 1/n$, where A and α are as in Assumption AK3.

Note that N3 is ensured by the upper bound on σ_n. We now define the sequence $a_n = 2(C(S_n + g_n) + \sigma_n^{1-\epsilon})$, for C as in Assumption A2. Then by N1 above we have $a_n = O(S_n)$. We now go on to show that $\{a_n\}_{n=1}^{\infty}$ satisfies the three results stated in the lemma. Combining N1 and N2 above, we have,

$$\begin{aligned} &\min\left\{p(X_j)\middle| j \leq n, \hat{p}_n(X_j) > \lambda - S_n\right\} > \lambda - S_n - g_n \\ \Rightarrow &\max\left\{d(X_j, \mathcal{L}(\lambda))\middle| j \leq n, \hat{p}_n(X_j) > \lambda - S_n\right\} < C(S_n + g_n) \end{aligned}$$

As a result, every element of $\{X_1, ..., X_n\}$ whose estimated density is above $\lambda - S_n$ is within $C(S_n + g_n)$ of a component of $\mathcal{L}(\lambda)$. Result 3. in the statement of the lemma follows immediately. Furthermore, take any $\mathbf{w} \in \mathcal{L}(\lambda)$. Then, $\hat{p}_n(\mathbf{w}) \geq \lambda - g_n$, and so

$$\begin{aligned} \lambda - g_n \leq \frac{c_K}{n\sigma_n^d}\sum_{i=1}^{n} K\left(\frac{||\mathbf{w} - X_i||}{\sigma_n}\right) &\leq \frac{c_K}{n\sigma_n^d}\sum_{i:||\mathbf{w}-X_i||<\sigma_n^{1-\epsilon}} K\left(\frac{||\mathbf{w} - X_i||}{\sigma_n}\right) + \frac{c_K}{\sigma_n^d}K\left(\sigma_n^{-\epsilon}\right) \\ \Rightarrow \frac{n\sigma_n^d(\lambda - g_n)}{c_K} - nA\exp(-\sigma_n^{-\alpha\epsilon}) &\leq \sum_{i:||\mathbf{w}-X_i||<\sigma_n^{1-\epsilon}} K\left(\frac{||\mathbf{w} - X_i||}{\sigma_n}\right) \\ &\leq \left|\{X_1, ..., X_n\} \cap \mathcal{B}_{\sigma_n^{1-\epsilon}}(\mathbf{w})\right|, \end{aligned}$$

and so by N3, $\left|\{X_1, ..., X_n\} \cap \mathcal{B}_{\sigma_n^{1-\epsilon}}(\mathbf{w})\right| \geq 1$. As a result, for any $\mathbf{w} \in \ell(\lambda, k)$, for some $k \in [c]$, there exists $j \in [n]$ such that $d(\mathbf{w}, X_j) < \sigma_n^{1-\epsilon}$, and so $p(X_j) \geq p(\mathbf{w}) - \kappa\sigma_n^{1-\epsilon} \Rightarrow \hat{p}_n(X_j) > \lambda - \kappa\sigma_n^{1-\epsilon} - g_n > \lambda - S_n \Rightarrow X_j \in \widehat{\ell(\lambda, k)}^{(n)}$.

Notice also, from above, that, since there is no $j \in [n]$ s.t. $\hat{p}(X_j) > \lambda - S_n$ and $d(X_j, \mathcal{L}(\lambda)) \geq C(S_n + g_n)$, we have

$$\widehat{\ell(\lambda,k)}^{(n)} \subset \mathcal{B}_{C(S_n+g_n)}(\ell(\lambda,k)).$$

Now take any $\mathbf{w} \in \mathcal{B}_{C(S_n+g_n)}(\ell(\lambda,k))$. Then $d(\mathbf{w}, \ell(\lambda,k)) < C(S_n+g_n) \Rightarrow d(\mathbf{w}, \widehat{\ell(\lambda,k)}^{(n)}) < C(S_n+g_n) + \sigma_n^{1-\epsilon}$, since every point in $\ell(\lambda,k)$ is within $\sigma_n^{1-\epsilon}$ of some $X_j, j \in [n]$ with $\hat{p}_n(X_j) > \lambda - S_n$. Since $\mathbf{w}$ was arbitrary, we thus have that $\widehat{\ell(\lambda,k)}^{(n)}$ is connected at distance $2(C(S_n+g_n)+\sigma_n^{1-\epsilon}) = a_n$. This proves result 1. in the lemma.

Result 2. also follows immediately from above, since we have established that every $\mathbf{w} \in \ell(\lambda,k)$ lies in $\mathcal{B}_{\sigma_n^{1-\epsilon}}(\widehat{\ell(\lambda,k)}^{(n)}) \subset \mathcal{B}_{\frac{a_n}{2}}(\widehat{\ell(\lambda,k)}^{(n)})$, and also that $\widehat{\ell(\lambda,k)}^{(n)} \subset \mathcal{B}_{C(S_n+g_n)}(\ell(\lambda,k)) \subset \mathcal{B}_{\frac{a_n}{2}}(\ell(\lambda,k))$.

■

The first and second results in the above lemma ensure that points in the sequence $X_1, X_2, ...$ which fall in the same level set component are connected at small distances for large values of n, whereas the second and third results ensure that, with probability one, if $X_i \in \ell(\lambda,k)$ and $X_j \notin \ell(\lambda,k)$, for some k, then for all large n, either $X_j \notin \widehat{\mathcal{L}(\lambda)}^{(n)}$ or there is no subset of $\widehat{\mathcal{L}(\lambda)}^{(n)}$ containing both X_i and X_j which is connected at as small a distance.

Next we show that the degree of connectedness and separation of points falling in each of the level set components is sufficient for spectral clustering to allow trivial recovery of the desired partition, almost surely, as $n \to \infty$. As always, we cover the different Laplacians separately for completeness. We have simplified the conditions surrounding the sequence of scale parameters, $\{\sigma_n\}_{n=1}^{\infty}$, and the related sequence, $\{S_n\}_{n=1}^{\infty}$, in the remaining results. In particular, the sequence $\{S_n\}_{n=1}^{\infty}$ only arises implicitly as $S_n = D\sigma_n^{1-\epsilon}$, for any $D > 0$ and an appropriately chosen ϵ. We have retained considerable generality in the following, however, by not suppressing the third condition above, i.e., that there is a $c > 0$ such that $\sigma_n \leq c\sigma_{2n}$ for all n. The conditions in the remaining results could be simplified so that the only requirement is that $\{\sigma_n\}_{n=1}^{\infty}$ decreases in the limit as $Nn^{-\delta}$ for any $N > 0$ and any $0 < \delta < \frac{1}{(d+2)}$. Importantly this includes the rate $n^{-\frac{1}{d+4}}$ which is the asymptotic mean integrated squared error optimal rate for kernel density estimation.

Theorem 18 *Let $X_1, X_2, ...$ be an i.i.d. sequence of random variables on $\mathbb{R}^d$, $d \geq 2$, with density p satisfying assumptions A1–A2 for level $\lambda > 0$, and suppose that $\mathcal{L}(\lambda)$ has components $\ell(\lambda,1), ..., \ell(\lambda,c)$. Let $K : \mathbb{R}^+ \to \mathbb{R}^+$ satisfy assumptions AK1–AK3. Let $0 < \epsilon < 0.5$ be fixed and let $\{\sigma_n\}_{n=1}^{\infty}$ be a sequence of positive scalars for which $\exists c > 0$ with $\sigma_n \leq c\sigma_{2n}$ for all n, and satisfying, for all large n,*

$$\log(n)n^{-1/(d+2)} < \sigma_n < \log(n)^{-\log(\log(n))}.$$

For each $n \in \mathbb{N}$ let

$$\widehat{\mathcal{L}(\lambda)}^{(n)} = \left\{ X_j \middle| j \leq n, \frac{c_K}{n\sigma_n^d} \sum_{i=1}^{n} K(||X_i - X_j||/\sigma_n) > \lambda - D\sigma_n^{1-\epsilon} \right\},$$

for any fixed $D > 0$. Let $\mathbf{L}^{(n)}$ be the unnormalised Laplacian of the graph with vertices $\widehat{\mathcal{L}(\lambda)}^{(n)}$ and edge weights determined using K_{σ_n}, and let $\mathbf{U}^{(n)}$ be its eigenvectors. Finally, let $\{\mathcal{C}_1^{(n)}, ..., \mathcal{C}_c^{(n)}\}$ be any constant approximation to the optimal c-centers clustering solution obtained from $\mathbf{U}^{(n)}_{:,1:c}$. Then with probability one, there is a sequence of permutations of $[c]$, say $\{\omega^n\}_{n=1}^{\infty}$, such that

$$\lim_{n\to\infty} \mathcal{B}_{\sigma_n^{1-2\epsilon}}\left(\mathcal{C}^{(n)}_{\omega_k^n}\right) = \ell(\lambda, k),$$

for all $k \in [c]$.

Proof First note that if we set, for each $n \in \mathbb{N}$, $S_n = D\sigma_n^{1-\epsilon}$, then $\{\sigma_n\}_{n=1}^{\infty}$ and $\{S_n\}_{n=1}^{\infty}$ satisfy the requirements in Lemma 17, if we choose $1 < \xi < \frac{1}{1-\epsilon}$. Now, since $\mathcal{L}(\lambda)$ is closed and has finitely many components, we know that there exists a $\Delta > 0$ such that for all $k, l \in [c], k \neq l$, we have $\mathcal{B}_\Delta \ell(\lambda, k) \cap \mathcal{B}_\Delta \ell(\lambda, l) = \emptyset$. Combining the results of Lemma 17, it is straightforward to verify that there exists a $M > 0$ independent of n such that, with probability one, for all n sufficiently large, we have

1. $\widehat{\ell(\lambda, k)}^{(n)} := \widehat{\mathcal{L}(\lambda)}^{(n)} \cap \mathcal{B}_{M\sigma_n^{1-\epsilon}}(\ell(\lambda, k))$ is connected at distance $M\sigma_n^{1-\epsilon}$ for all $k \in [c]$.
2. $\ell(\lambda, k) \subset \mathcal{B}_{M\sigma_n^{1-\epsilon}}(\widehat{\ell(\lambda, k)}^{(n)}) \subset \mathcal{B}_{2M\sigma_n^{1-\epsilon}}(\ell(\lambda, k))$.
3. For all $k, l \in [c]$, $k \neq l$, we have $d(\widehat{\ell(\lambda, k)}^{(n)}, \widehat{\ell(\lambda, l)}^{(n)}) \geq \Delta$.
4. $\{\widehat{\ell(\lambda, 1)}^{(n)}, ..., \widehat{\ell(\lambda, c)}^{(n)}\}$ is a partition of $\widehat{\mathcal{L}(\lambda)}^{(n)}$.

Now, it is clear that for large n we have $M\sigma_n^{1-\epsilon} < \sigma_n^{1-2\epsilon}$, and so point 2 above ensures that

$$\lim_{n\to\infty} \mathcal{B}_{\sigma_n^{1-2\epsilon}}\left(\widehat{\ell(\lambda, k)}^{(n)}\right) = \ell(\lambda, k),$$

for all $k \in [c]$. It is therefore sufficient to show that the eigenvectors of $\mathbf{L}^{(n)}$ allow us to recover the partition in point 4 above. To that end, consider that for n large enough we have $\sigma_n < (\Delta - M\sigma_n^{1-\epsilon}) \log(An^{13})^{-1/\alpha}$, since $\sigma_n < \log(n)^{-\log(\log(n))} < N \log(n)^{-1/\alpha}$ for any $N > 0$ as n is large. Combining this with points 1, 3 and 4 above we can apply the results of Theorem 12 to see that for these n and any $l \in [c]$, and $i, j, k \in [n]$ with $X_i, X_j \in \widehat{\ell(\lambda, l)}^{(n)}, X_k \in \widehat{\mathcal{L}(\lambda)}^{(n)} \setminus \widehat{\ell(\lambda, l)}^{(n)}$ we have,

$$||\mathbf{U}^{(n)}_{(i),1:c} - \mathbf{U}^{(n)}_{(j),1:c}|| \leq c^{\frac{1}{2}} n^{-5},$$
$$||\mathbf{U}^{(n)}_{(i),1:c} - \mathbf{U}^{(n)}_{(k),1:c}|| \geq \sqrt{\frac{2}{n}} - 6c^{\frac{9}{8}} n^{-1},$$

where we assume that X_i, X_j, X_k are the (i)-th, (j)-th and (k)-th elements of $\widehat{\mathcal{L}(\lambda)}^{(n)}$, respectively. Since this holds simultaneously for all such i, j, k, l, it follows that any T-approximation for the optimal c-center clustering solution of $\mathbf{U}^{(n)}_{:,1:c}$ will recover the partition

in point 4, provided n is large enough that $Tc^{\frac{1}{2}}n^{-5} < \frac{1}{2}\left(\sqrt{\frac{2}{n}} - 6c^{\frac{9}{8}}n^{-1}\right)$. ■

The case of the normalised Laplacian of the graph where reflexive edges are not removed follows exactly analogously.

Theorem 19 *Let the conditions of Theorem 18 hold. For each $n \in \mathbb{N}$ let $\mathbf{L}_{\mathrm{N}}^{(n)}$ be the normalised Laplacian of the graph with vertices $\widehat{\mathcal{L}(\lambda)}^{(n)}$ and edge weights determined using K_{σ_n}, and let $\mathbf{U}^{(n)}$ be its eigenvectors and $\mathbf{D}^{(n)}$ the corresponding degree matrix. Let $\{\mathcal{C}_1^{(n)}, ..., \mathcal{C}_c^{(n)}\}$ be any constant approximation to the optimal c-centers clustering solution obtained from $(\mathbf{D}^{(n)})^{-1/2}\mathbf{U}_{:,1:c}^{(n)}$. Then with probability one, there is a sequence of permutations of $[c]$, say $\{\omega^n\}_{n=1}^{\infty}$, such that*

$$\lim_{n\to\infty} \mathcal{B}_{\sigma_n^{1-2\epsilon}}\left(\mathcal{C}_{\omega_k^n}^{(n)}\right) = \ell(\lambda, k),$$

for all $k \in [c]$.

Proof The proof is exactly analogous to the previous proof. ■

When the reflexive edges in the graph are removed, then, as in previous cases, some modifications are needed. These are given explicitly in the proof of the following.

Theorem 20 *Let the conditions of Theorem 18 hold, and let $K(x) = \exp(-x^{\alpha})$ for some $\alpha > 0$. For each $n \in \mathbb{N}$ let $\mathbf{L}_{\mathrm{N}_0}^{(n)}$ be the normalised Laplacian of the graph with vertices $\widehat{\mathcal{L}(\lambda)}^{(n)}$ and edge weights determined using K_{σ_n}, but with reflexive edges removed, and let $\mathbf{U}^{(n)}$ be its eigenvectors and $\mathbf{D}^{(n)}$ the corresponding degree matrix. Let $\{\mathcal{C}_1^{(n)}, ..., \mathcal{C}_c^{(n)}\}$ be any constant approximation to the optimal c-centers clustering solution obtained from $(\mathbf{D}^{(n)})^{-1/2}\mathbf{U}_{:,1:c}^{(n)}$. Then with probability one, there is a sequence of permutations of $[c]$, say $\{\omega^n\}_{n=1}^{\infty}$, such that*

$$\lim_{n\to\infty} \mathcal{B}_{\sigma_n^{1-2\epsilon}}\left(\mathcal{C}_{\omega_k^n}^{(n)}\right) = \ell(\lambda, k),$$

for all $k \in [c]$.

Proof The proof is similar, but in this case we state the results from Lemma 17 slightly differently. Specifically, we replace point 3 in the proof of Theorem 18 with

3. With probability one, for all n large enough and for all $k, l \in [c]$, $k \neq l$, we have $d(\widehat{\ell(\lambda, k)}^{(n)}, \widehat{\ell(\lambda, l)}^{(n)})^{\alpha} - 3M^{\alpha}\sigma_n^{\alpha(1-\epsilon)} \geq \Delta^{\alpha}$.

Now, for large n we have $\sigma_n < \Delta \log(13^8 c^9 n^{10})^{-1/\alpha}$. Using Theorem 15, we thus find that if $X_i, X_j \in \widehat{\ell(\lambda, l)}^{(n)}, X_k \in \widehat{\mathcal{L}(\lambda)}^{(n)} \setminus \widehat{\ell(\lambda, l)}^{(n)}$ are the (i)-th, (j)-th and (k)-th elements of $\widehat{\mathcal{L}(\lambda)}^{(n)}$ respectively, then, letting $\mathbf{V}^{(n)} = (\mathbf{D}^{(n)})^{-1/2}\mathbf{U}^{(n)}$, we have

$$\frac{||\mathbf{V}_{(i),1:c}^{(n)} - \mathbf{V}_{(j),1:c}^{(n)}||}{||\mathbf{V}_{(i),1:c}^{(n)} - \mathbf{V}_{(k),1:c}^{(n)}||} \leq n^{-3}.$$

Thus for $n^3 > T$, a T-approximation to the optimal c-centers solution obtained from $\mathbf{V}_{:,1:c}$ will recover the partition of $\widehat{\mathcal{L}(\lambda)}^{(n)}$ into $\{\widehat{\ell(\lambda,1)}^{(n)}, ..., \widehat{\ell(\lambda,c)}^{(n)}\}$, as required. The convergence of the $\sigma_n^{1-2\epsilon}$ neighbourhoods of these sets to the components of $\mathcal{L}(\lambda)$ was discussed in the proof of Theorem 18. ■

The above three results show that the level set components are consistently estimated by spectral clustering applied to the estimated level set, $\widehat{\mathcal{L}(\lambda)}^{(n)}$, assuming that the number of components is known. Dependence on this value arises both in that we only obtain a clustering into c clusters, but also importantly in that the distinction of the clusters is only guaranteed within the first c of the eigenvectors of $\mathbf{L}^{(n)}, \mathbf{L}_{\mathrm{N}}^{(n)}$ and $\mathbf{L}_{\mathrm{N}_0}^{(n)}$. The final three results show that c can be consistently estimated by considering scaled sequences of the eigenvalues of the various Laplacian matrices. Combining these with the previous results therefore ensures that the level set components can be consistently estimated using the approach described herein. It is well known that for disconnected graphs, the number of zero eigenvalues of the graph Laplacians corresponds to the number of components (Von Luxburg, 2007). In such cases, the multiplicity of the zero eigenvalue may therefore be used to determine the number of clusters. The following three results show that the eigenvalues also lead to correct identification of the number of clusters when a fully connected graph is used.

Theorem 21 *Let the conditions of Theorem 18 hold. For each* $l \in \left[\left|\widehat{\mathcal{L}(\lambda)}^{(n)}\right|\right]$, *let* $e_l^{(n)}$ *be the* l*-th eigenvalue of* $\mathbf{L}^{(n)}$. *Then we have,*

$$\frac{e_l^{(n)}}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \xrightarrow{a.s.} 0, \text{ for } l \in [c], \quad \frac{e_{c+1}^{(n)}}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \xrightarrow{a.s.} \infty.$$

Proof We again use the beginning of the proof of Theorem 18 to obtain points 1–4 from the results of Lemma 17. By Lemma 9 we thus have,

$$\sum_{l=1}^{c} e_l^{(n)} \leq ncK\left(\frac{\Delta}{\sigma_n}\right)$$
$$e_{c+1}^{(n)} \geq \frac{1}{9n^2(c+1)^4} K\left(\frac{M}{\sigma_n^{\epsilon}}\right) - n^3 cK\left(\frac{\Delta}{\sigma_n}\right).$$

For any $l \in [c]$, we thus have, for n large enough that $\sigma_n^{1-\sqrt{\epsilon}} < \frac{1}{2}\Delta$,

$$\frac{e_l^{(n)}}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \leq \frac{ncK(\Delta/\sigma_n)}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \leq ncA\exp\left(-\left(\frac{\Delta - \sigma_n^{1-\sqrt{\epsilon}}}{\sigma_n}\right)^{\alpha}\right) \leq ncA\exp\left(-\frac{\Delta^{\alpha}}{2^{\alpha}\sigma_n^{\alpha}}\right)$$
$$\leq ncA\exp\left(-\frac{\Delta^{\alpha}}{2^{\alpha}}\log(n)^{\alpha\log(\log(n))}\right) = cAn^{1-\frac{\Delta^{\alpha}}{2^{\alpha}}\log(n)^{\alpha\log(\log(n))-1}} \to 0 \text{ as } n \to \infty.$$

In addition, we have

$$\frac{e_{c+1}^{(n)}}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \geq \frac{K(M/\sigma_n^{\epsilon})}{9n^2(c+1)^4 K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} - \frac{n^3 cK(\Delta/\sigma_n)}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)}$$

Now, since $\epsilon < 1$, we have $\sqrt{\epsilon} > \epsilon$, and so for n large enough, we have $M\sigma_n^{\sqrt{\epsilon}-\epsilon} < \frac{1}{2}$. Therefore, using assumption AK3 and the fact that K is strictly positive,

$$\begin{aligned}
\frac{K(M/\sigma_n^{\epsilon})}{9n^2(c+1)^4 K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} &\geq \frac{1}{9An^2(c+1)^4} \exp\left(\left(\frac{1-M\sigma_n^{\sqrt{\epsilon}-\epsilon}}{\sigma_n^{\sqrt{\epsilon}}}\right)^{\alpha}\right) \\
&\geq \frac{1}{9An^2(c+1)^4} \exp\left(\frac{1}{2^{\alpha}\sigma_n^{\alpha\sqrt{\epsilon}}}\right) \\
&\geq \frac{1}{9An^2(c+1)^4} \exp\left(\frac{1}{2^{\alpha}} \log(n)^{\alpha \log(\log(n))\sqrt{\epsilon}}\right) \\
&= \frac{1}{9A(c+1)^4} n^{\frac{1}{2^{\alpha}} \log(n)^{\alpha \log(\log(n))\sqrt{\epsilon}-1}-2} \to \infty \text{ as } n \to \infty, \\
\text{and } \frac{n^3 cK(\Delta/\sigma_n)}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} &\leq cAn^{3-\frac{\Delta^{\alpha}}{2^{\alpha}} \log(n)^{\alpha \log(\log(n))-1}} \to 0 \text{ as } n \to \infty,
\end{aligned}$$

as before. This proves the result. ■

Theorem 22 *Let the conditions of Theorem 21 hold. Let* $\mathbf{L}_{\mathrm{N}}^{(n)}$ *be the normalised Laplacian of the graph with vertices* $\widehat{\mathcal{L}(\lambda)}^{(n)}$ *and edge weights determined using* K_{σ_n}*. For each* $l \in \left[\left|\widehat{\mathcal{L}(\lambda)}^{(n)}\right|\right]$*, let* $e_l^{(n)}$ *be the* l*-th eigenvalue of* $\mathbf{L}_{\mathrm{N}}^{(n)}$*. Then we have,*

$$\frac{e_l^{(n)}}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \xrightarrow{a.s.} 0, \text{ for } l \in [c], \quad \frac{e_{c+1}^{(n)}}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \xrightarrow{a.s.} \infty.$$

Proof The proof is exactly analogous to the previous proof. ■

Theorem 23 *Let the conditions of Theorem 21 hold, and let* $K(x) = \exp(-x^{\alpha})$*. Let* $\mathbf{L}_{\mathrm{N}_0}^{(n)}$ *be the normalised Laplacian of the graph with vertices* $\widehat{\mathcal{L}(\lambda)}^{(n)}$ *and edge weights determined using* K_{σ_n}*, but with reflexive edges removed. For each* $l \in \left[\left|\widehat{\mathcal{L}(\lambda)}^{(n)}\right|\right]$*, let* $e_l^{(n)}$ *be the* l*-th eigenvalue of* $\mathbf{L}_{\mathrm{N}_0}^{(n)}$*. Then we have,*

$$\frac{e_l^{(n)}}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \xrightarrow{a.s.} 0, \text{ for } l \in [c], \quad \frac{e_{c+1}^{(n)}}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)} \xrightarrow{a.s.} \infty.$$

Proof Using the same approach as in the proof of Theorem 21 we have, now using Lemma 11, that

$$\sum_{l=1}^{c} e_l^{(n)} \leq nc\frac{K\left(\Delta/\sigma_n\right)}{K(M/\sigma_n^{\epsilon})}$$
$$e_{c+1}^{(n)} \geq \frac{K\left(M/\sigma_n^{\epsilon}\right)}{9n^3(c+1)^4} - n^3 c\frac{K\left(\Delta/\sigma_n\right)}{K(M/\sigma_n^{\epsilon})}.$$

The first term in $e_{c+1}^{(n)}$, divided by $K\left(\sigma_n^{-\sqrt{\epsilon}}\right)$ tends to ∞ almost surely, almost exactly as before. The second term in $e_{c+1}^{(n)}$ and in the first c eigenvalues converge to zero fast enough, almost surely, since for n large enough that $M^{\alpha}\sigma_n^{\alpha(1-\epsilon)} < \frac{1}{4}\Delta^{\alpha}$ and $\sigma_n^{\alpha(1-\sqrt{\epsilon})} < \frac{1}{4}\Delta^{\alpha}$, we have

$$\frac{cn^3K(\Delta/\sigma_n)}{K\left(\sigma_n^{-\sqrt{\epsilon}}\right)K(M/\sigma_n^{\epsilon})} = n^3 c\exp\left(-\frac{\Delta^{\alpha}}{\sigma_n^{\alpha}} + \frac{M^{\alpha}}{\sigma_n^{\alpha\epsilon}} + \frac{1}{\sigma_n^{\alpha\sqrt{\epsilon}}}\right)$$
$$\leq n^3 c\exp\left(-\frac{\Delta^{\alpha}}{2\sigma_n^{\alpha}}\right) \to 0 \text{ as } n \to \infty.$$

The rest of the proof follows as before. ■

6. Discussion

In this paper we investigated the relationships between spectral clustering and the problems of maximum margin clustering and estimation of level sets of a probability density. Although these two problems are not usually associated with one another, by applying a maximum margin clustering method to a truncated sample whose low-density points have been removed, it is intuitively the case that such an approach is likely to recover an approximation of the components of a level set of the underlying density. We extended existing theory on the connection between spectral clustering and density level sets by considering multiple versions of spectral clustering, by considering a broader class of kernels including the ubiquitous Gaussian kernel, and importantly achieve consistent estimation with a sequence of scaling parameters which decreases with the sample size. Existing convergence results for spectral clustering assume a fixed bandwidth kernel is used. Although intuitive, as far as we are aware the connection between spectral clustering and maximum margin clustering in the general case has not been made explicit until now.

Acknowledgments

The author would like to express his gratitude to the reviewers, whose insights and recommendations greatly improved the quality of the paper in its final form.

References

Daniel Alspach and Harold Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE transactions on automatic control*, 17(4):439–448, 1972.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Estimating the number of clusters. *Canadian Journal of Statistics*, 28(2):367–382, 2000.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

Lars Hagen and Andrew Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions On Computer-aided Design of Integrated Circuits and Systems*, 11(9):1074–1085, 1992.

John A Hartigan. Clustering algorithms. 1975.

David P Hofmeyr. Improving spectral clustering using the asymptotic value of the normalized cut. *Journal of Computational and Graphical Statistics*, pages 1–13, 2019.

David P Hofmeyr, Nicos G Pavlidis, and Idris A Eckley. Minimum spectral connectivity projection pursuit. *Statistics and Computing*, 29(2):391–414, 2019.

James R Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):1–30, 2014.

Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

Hariharan Narayanan, Mikhail Belkin, and Partha Niyogi. On the relation between low density separation, spectral clustering and graph cuts. In *Advances in Neural Information Processing Systems*, pages 1025–1032, 2006.

Bruno Pelletier and Pierre Pudlo. Operator norm convergence of spectral clustering on level sets. *Journal of Machine Learning Research*, 12(Feb):385–416, 2011.

Richard Peng, He Sun, and Luca Zanetti. Partitioning well-clustered graphs: Spectral clustering works! In *Conference on Learning Theory*, pages 1423–1455, 2015.

Alessandro Rinaldo, Larry Wasserman, et al. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.

Murray Rosenblatt. Stochastic curve estimation. In *NSF-CBMS Regional Conference Series*, volume 3. Institute of Mathematical Sciences, 1991.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Nicolas Garcia Trillos, Dejan Slepcev, James Von Brecht, Thomas Laurent, and Xavier Bresson. Consistency of cheeger and ratio graph cuts. *The Journal of Machine Learning Research*, 17(1):6268–6313, 2016.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007. ISSN 0960-3174. doi: 10.1007/s11222-007-9033-z.

Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

Dorothea Wagner and Frank Wagner. *Between min cut and graph bisection*. Springer, 1993.

Guenther Walther. Granulometric smoothing. *The Annals of Statistics*, pages 2273–2299, 1997.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.