Title
Spectral Embedding Norm

Permalink
https://escholarship.org/uc/item/3b08w3wh

Author
Dror, Shahar

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Spectral Embedding Norm

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Electrical Engineering (Communication Theory and Systems)

by

Shahar Dror

Committee in charge:

Professor Gal Mishne, Chair
Professor Paul Siegel, Co-Chair
Professor Piya Pal

2021

The thesis of Shahar Dror is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to extend my sincerest gratitude to my advisor, Dr. Gal Mishne for her help and guidance towards the work contained in this thesis. I have learnt a tremendous amount through the discussions that we have had together. This thesis would not have been what it is without her.

I am also indebted to Dr. Xiuyuan Cheng for her invaluable insights and constructive criticisms.

I am very grateful to the rest of the members of my committee namely, Dr. Paul Siegel and Dr. Piya Pal for their helpful comments and support.

The material in this thesis is currently being prepared for submission for publication of the material. Dror, Shahar; Mishne, Gal; Cheng, Xiuyuan. The thesis author was the primary investigator and author of this material.

ABSTRACT OF THE THESIS

Spectral Embedding Norm

by

Shahar Dror

Master of Science in Electrical Engineering (Communication Theory and Systems)

University of California San Diego, 2021

Professor Gal Mishne, Chair
Professor Paul Siegel, Co-Chair

Anomaly detection is a data partitioning algorithm which separates outliers from normative data points. An *unsupervised learning* approach to this problem does not assume any prior information. Anomaly detection is a primary data analysis task with diverse applications and has been studied under many models and assumptions. Spectral methods such as spectral clustering have been widely used to solve the clustering problem. These methods make use of the leading K eigenvectors of the graph Laplacian matrix to detect K clusters, if the graph has a clear community structure. In a setting where the data consists of unbalanced clusters, as in anomaly detection, the spectral properties are determined by a dominating component. In such cases, traditional graph clustering methods fail, while the *spectral embedding norm* was found to overcome this challenge. This thesis generalizes the spectral embedding norm definition, formerly used, by introducing the capability of a weighted norm. With just one simple natural condition: allowing limited connectivity between the clusters and the background, we prove that this quantity can be used to detect the clusters of interest. Experiments on both synthetic data sets and real-world defect detection images demonstrate the effectiveness of the algorithm and its performance was found to be stable, with respect to parameter choices.

Chapter 1

Introduction

1.1 Clustering and Anomaly Detection

Clustering has always been a fundamental technique in unsupervised learning, for data analysis and exploration. Dividing data points into several groups such that points in the same group are similar and points in different groups are dissimilar, with no prior information, is a general task that is formulated with respect to the data set and intended use of the results. This requires an appropriate choice of clustering algorithm and parameter settings. Different algorithms may differ significantly in their understanding of what constitutes a cluster and how to efficiently find them, since no definition is universally accepted and is often application-dependent. Vertex similarity is the basis of traditional methods,like hierarchical and spectral clustering [For10].

A variant of the clustering problem is the task of anomaly detection which focuses on finding data that does not conform to the expected behavior of normal data points. Equivalently, anomaly detection identifies interesting clusters within the data set. Automatic anomaly detection, based on graph clustering methods, are defined for general weighted graphs and are very successful in various applications, such as detecting sea mines in side-scan sonar images [MC12] and defects in industrial monitoring applications [ZC10].

We consider the setting where the data consists of interesting small clusters and the majority of data points belong to one cluster, namely the background. The unbalanced size of the clusters compared to the background and inter-cluster edges linking the background to the clusters poses challenges for traditional clustering methods. Nadler and Galun [NG07] show that spectral clustering algorithms cannot successfully cluster data sets that contain structures at different scales of size and density. Their analysis is summarized in 2.4.

In this thesis we present a model that distinguishes data points that either belong to a cluster or the background, with a theoretical guarantee, and compare to the popular graph-based approach of Spectral Clustering [SM00]. The model is motivated by applications and will be tested on real-world data.

1.2 Spectral Clustering

Spectral clustering methods are common graph-based approaches to unsupervised clustering of data. The advantages of spectral clustering is that it is very simple to implement and can be solved efficiently by standard linear algebra methods. Several spectral clustering methods propose to construct a graph on the data set and cluster it into k clusters by embedding the high-dimensional data points with the first k dominant eigenvectors of the graph Laplacian matrix [SM00, NJW01].

Anomaly detection can be seen as a special case of clustering in which there is a vast imbalance in the size and density of the background cluster compared to the anomalous clusters. Several works in spectral clustering analyze the limitations of spectral clustering in the presence of multi-scale data with varying density [NG07, VLBB08] and show that in many examples the information provided by the leading eigenvectors is misleading and cannot be used for clustering as they fail to capture the anomaly.

In past years, non-linear techniques for dimensionality reduction have been proposed to preserve local information whilst deriving dominant eigenvectors informative of either the global or local graph dynamics, including ISOMAP [TDSL00], locally linear embedding (LLE) [RS00], Laplacian Eigenmaps [BN03], Hessian Eigenmaps [DG03] and Diffusion Maps [CL06]. The calculation of the embedding relies on the construction on an affinity kernel on the data and its Laplacian matrix. The choice of the affinity kernel corresponds to preserving different properties of the original high-dimensional data. Spectral dimensionality reduction algorithms suffer from a major limitation known as the “repeated eigen-directions”. That is, successive eigenvectors tend to represent directions along the data manifold which were already captured by previous ones [BM17, DTCK18]. This can negatively impact representation if choosing only the first dominant eigenvectors. On account of the redundant information in the leading eigenvectors, our approach looks deep in the graph Laplacian spectrum. In this thesis, we propose a Laplacian-based clustering algorithm adapted for the purpose of outlier detection.

1.3 Anomaly Detection in Images

Automatic anomaly detection in images is concerned with the problem of separating objects from the image background based on its different appearance or statistical properties. This is especially challenging where no prior information on the image normality is present. There are many relevant applications that must rely on unsupervised algorithms that can detect anomalous regions in natural images. Such fields include military applications, medical image analysis and automation of quality assurance processes, which produce large amounts of images. A robust solution that presents the user only with suspicious objects saves valuable time and manpower, as suspicious objects occur very rarely by nature.

Anomaly detection in images is challenging due to several factors:

1. *Large size of the data set* - As technology progresses, so has the image resolution which increases the number of pixels per image. Modern data sets include images with millions

of pixels.

2. *High dimensionality of the data* - Data samples are usually represented using a high-dimensional feature vector such as patches surrounding each pixel.

3. *Noisy features* - Natural images incorporate non-idealities which may be falsely detected as anomalies of interest.

   In this thesis, we will apply our approach to a recently published image outlier detection data set [BFSS19].

4. *Multiple background components* - In many images, the normal data points do not belong to a single cluster.

Detecting anomalies in a single image is possible under the assumption that anomalies are a minor part of the image. Since anomalies cannot be modeled, the attention of common detection algorithms rely on background modeling. There are many approaches to bridge this gap that are based on statistical models, machine learning, saliency based methods, sparse representations, etc. Such methods can be further categorised according to the assumed background model [EDMD19]. Common approaches include:

1. *Stochastic background model* - The background is modeled as a probability density function (PDF) and the anomalies are detected by a binary hypothesis test. [DZ10]

2. *Sparse background model* - The background components are represented by a dictionary constructed under the highly redundant background prior. Anomalies are detected by salient differences in reconstruction. [LZZM15]

3. *Homogeneous background model* - These methods generate an estimate of the background. This characterization can vary locally or apply globally on the whole dataset. Anomalous samples are identified using a distance to the background estimate. [TH99, TH03]

Ideally, a generic unsupervised algorithm should be able to separate anomalies from background pixels, in images, using a rigorous detection mechanism. The above methods do not use any prior information but make assumptions on the image properties that do not apply universally; in general, most methods are designed for a particular application and imply a detection threshold with no theoretical justification [EDMD19]. For a fair evaluation of the proposed method, we set a detection threshold, based on a uniform criterion, and setting a evaluation metric which links the number of false alarms (FA) with a true detection (TD). This detection criterion is commonly used [MC17, MC12, EDMD19].

1.4 Overview of the Thesis

Chapter 2 details the mathematical background on how we will represent the data. This is a common foundation for all spectral analysis methods. The original contribution of this research starts in Chapter 3.

The *spectral embedding norm* is introduced in Chapter 3. This is a positive scalar, calculated for each point in a data set, and separates outliers from the normative points in the data set.

Chapter 4 provides the theoretical results guaranteeing anomaly detection, under detailed conditions. The framework views the affinity matrix as a perturbation of a block affinity matrix where the background and clusters are completely disconnected. We formulate the consequent deformation of the spectral embedding norm and provide guaranteed detection of the clusters by simple thresholding. Detailed proofs leading to the Separation Theorem are in the Appendix.

Chapter 5 presents results of the anomaly detection algorithm on synthetic data sets and real-world multi-defect textured images. We show that the spectral embedding norm enables improved detection and stability when tested on multiple examples. The algorithm involves a parameter choice which defines the weight function of the eigenvectors summed over, and we show that the performance is not sensitive to the parameter choice.

Chapter 2

Mathematical Background

In this chapter, we present mathematical background on the methods we employed in this thesis. We review similarity graphs and the way of representing the data for analysis, and methods for graph construction. This entails choosing scaling parameters and deriving the graph Laplacian. The different graph Laplacians and their basic properties are reviewed.

2.1 Similarity Graph

In modeling data with a graph, the graph describes the notion of similarity between all pairs of data points using a geometric structure.

Given a set of n data points, $x_1,...,x_n \subset \mathbb{R}^p$, the methods begin by formulating a weighted graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ consisting of the set of nodes $\mathcal{V}$, where $|\mathcal{V}| = n$, a set of edges $\mathcal{E}$, and the weight matrix W. Each data point corresponds to a node in $\mathcal{V}$. An edge from node x_i to x_j has weight $W_{i,j}$. For an undirected graph, the edge weights are symmetric, $W_{i,j} = W_{j,i}$, and non-negative, $W_{i,j} \geq 0$.

The weight, $W_{i,j}$, measures the similarity of any two nodes, x_i and x_j. The choice of the weight function should be determined by the application, since it conveys the local geometry of the data set. A common choice is a Gaussian kernel, where σ is a scaling parameter (bandwidth) and $d(\cdot)$ is some distance function on the data.

$$W_{i,j} = \exp\left(-\frac{d^2(x_i,x_j)}{\sigma^2}\right) \tag{2.1}$$

Selecting the appropriate scale of analysis requires prior knowledge of the data set and a single global scale assumes a uniform density across different clusters. To eliminate such assumptions, we make use of an adaptive scaling parameter, varying for each pair of points, as detailed in 2.2.

The degree of a node $x_i \in \mathcal{V}$ is defined as

$$d_i = \sum_{j=1}^{n} W_{i,j}$$

The degree matrix D is defined as the diagonal matrix with the degrees $d_1,...,d_n$ on the diagonal.

2.2 Self-Tuning Scaling Parameter

In 2.1, the scale parameter σ has high impact on the clustering quality. Fixing σ to a value which is too small results in a disconnected graph, where many points are connected only to themselves. Setting σ to be too large, results in connecting together all the points in the graph. In such cases, data outliers are strongly connected to normative data points and are incorporated in the background. This leads to mistaken characterization and detection of anomalies in the data. Possible values for σ include the standard deviation or median of distances between points in the data set. Improved methods select the appropriate scale of analysis locally. In these methods, σ varies with every pair of points, e.g [ZMP04, ZLY11].

Zhang et al. [ZLY11] introduce a local density measure that scales the distance between points by the number of points in the joint region of the ε-neighborhood around them. This is under the assumption that two points distributed in the same cluster, are in the same region, which has a relatively high density. In practice, this method does not outperform the traditional clustering methods when applied on real data sets [Bea15].

The technique used in this thesis is the Self-Tuning scaling parameter. Zelnik and Perona [ZMP04] propose a scaling method used to compute the affinity between each pair of points. When the input data includes clusters with different local densities, a local scaling parameter should be used to compute the similarity between two points, varying for each pair of points. This allows to handle multi-scale data and background clutter. The self-tuning method calculates a local scaling parameter σ_i for each data point x_i, with the graph weights set to be

$$w_{ij} = \exp\left(-\frac{d^2(x_i, x_j)}{\sigma_i \sigma_j}\right),$$

where $\sigma_i = d(x_i, x_K^{(i)})$ where $x_K^{(i)}$ is the K^{th} closest point to x_i in terms of the distance function d.

2.3 Graph Laplacians

After the similarity matrix is constructed, the next step is forming the corresponding graph Laplacian matrix. There are three forms of graph Laplacian matrices which are, respectively, suitable for different clustering conditions.

1. Unnormalized graph Laplacian

$$L = D - W$$

The main property that distinguishes the unnormalized graph Laplacian is that self edges in the graph do not change L. These edges are defined by the diagonal elements of W.

2. Symmetric normalized graph Laplacian

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$

3. Random walk graph Laplacian

$$L_{rw} = I - D^{-1}W = I - P$$

The matrix P has nonnegative entries and $\sum_{j=1}^{n} P_{i,j} = 1$. Therefore, $P_{i,j}$ can be interpreted as the probability for a random walker to jump from x_i to x_j in a single time step and P is the time-homogeneous discrete-time transition matrix of a Markov chain over the data set.

L_{sym} L_{rw} are similar matrices, with identical eigenvalues and the corresponding eigenvectors satisfy $\psi_{sym} = D^{1/2}\psi_{rw}$.

A common problem of interest is to find a partition of a graph such that the edges between different groups have a very low weight and the edges within a group have high weight. The unnormalized Laplacian serves in the approximation of the minimization of RatioCut (which optimizes an objective relative to the number of nodes in each cluster). While the normalized Laplacian serves in the approximation of the minimization of NCut (which optimizes an objective relative to the volume of each cluster) [VL07].

2.4 Spectral Clustering and its Limitations

There are three similar spectral clustering algorithms, each one applied on a different graph Laplacian mentioned in 2.3 [VL07]. In all three algorithms, the main idea is to change the representation of each data point $x_i \in \mathbb{R}^p$ to a low-dimensional representation in $\mathbb{R}^k$, where $k \ll p$. This thesis compares the performance of the new method, spectral embedding norm, to the grouping algorithm introduced by Shi and Malik [SM00]:

Algorithm 1: Normalized spectral clustering according to Shi and Malik (2000)

```
Input: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with weight matrix $W$, $k$-number of clusters to construct
    1. Compute the unnormalized graph Laplacian, $L = D - W$.
    2. Solve $L\psi = \lambda D\psi$ for the first $k$ eigenvectors with the smallest eigenvalues.
    3. For $i = 1, \ldots, |\mathcal{V}|$, each node $x_i$ is represnted by $(\psi_1(i), \ldots, \psi_k(i)) \in \mathbb{R}^k$.
    4. Cluster the points in $\mathbb{R}^k$ with the $k$-means algorithm into clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$.
Output: Clusters $A_1, \ldots, A_k$ such that $A_j = \{i | x_i \in \mathcal{C}_j\}$
```

Spectral clustering requires the computation of the first k eigenvectors of a graph Laplacian matrix. If the graph is large, an exact computation of the eigenvectors is computationally expensive, as it requires a time $O(n^3)$, where n is the number of nodes in $\mathcal{G}$ [For10]. In such cases, approximations of the leading eigenvectors may be sufficient. Such techniques, notably the power method or Lanczos method [Lan50], are iterative and the speed of convergence for k eigenvectors, relies on the spectral gap $|\lambda_{k+1} - \lambda_k|$, where λ_k and λ_{k+1} are the k^{th} and $(k+1)^{th}$ smallest eigenvectors of L. The larger the spectral gap, the faster the algorithms converge.

A pertinent issue of spectral clustering, to the problem of anomaly detection, is its sensitivity to the data structure. Nadler and Galun [NG07] show that the first few eigenvectors of such adjacency matrices cannot successfully cluster datasets that contain structures at different scales of size and density. Their analysis introduces two characteristic times that depend on the transition matrix of $\mathcal{G}$ to predict the success of spectral clustering. First, the equilibrium time for each cluster, denoted by τ_i^R. Second, the mean first passage time from cluster $\mathcal{C}_i$ to $\mathcal{C}_j$, denoted by $\tau_{i,j}$. Spectral clustering succeeds when the equilibrium times are slower than the mean first passage times, i.e. $\max_i\{\tau_i^R\} < \min_{(i,j)}\{\tau_{i,j}\}$.

2.5 Heat Kernel Signature

The heat kernel signature (HKS) is based on analyzing the heat diffusion process on a compact Riemannian manifold, governed by the equation:

$$\frac{\partial u}{\partial t} = \Delta u(x,t) \tag{2.2}$$

where Δ is the Laplace-Beltrami operator. The solution $u(x,t)$ to (2.2), with the initial condition $u(x,0) = \delta_x$, where δ_x is the Dirac delta function at x (i.e. initially all the heat energy is concentrated in a point x) is called the heat kernel and describes the amount of heat on the surface at point x at time t.

$$u(x,t) = k_t(x,x), \qquad k_t(x,y) = \sum_{j=1}^{\infty} e^{-\zeta_j t} \phi_j(x) \phi_j(y)$$

where $\left\{\zeta_j\right\}_{k=1}^{\infty}$ are eigenvalues and $\left\{\phi_j\right\}_{k=1}^{\infty}$ are corresponding eigenfunctions of the Laplace-Beltrami operator.

Sun *et al.*[SOG09] proposed using the HKS, $k_t(x,x) = \sum_{j=1}^{\infty} e^{-\zeta_j t} \phi_j^2(x)$ as local shape descriptors. The exponential decay of the weights, when t is small, results in HKS being highly dominated by information from leading eigenfunctions, which correspond to global properties of the shape, and suppresses small scale information [ASC11].

Chapter 3

Problem formulation

Given a set of n points $\mathcal{V}$ in the feature space, an undirected weighted graph can be constructed, with affinity matrix $W \in \mathbb{R}^{n \times n}$. The edge weight, between any two nodes $x, y \in \mathcal{V}$, is denoted by $W(x, y)$ and is defined by a symmetric and non-negative function. In our analysis we assume that W has been constructed.

3.1 Setup and Notations

We adopt the framework introduced by Cheng and Mishne in [CM20]. Suppose $\mathcal{V}$ can be partitioned into two disjoint subsets: background and clusters, denoted by $\mathcal{B}$ and $\mathcal{C}$ respectively. The typical scenario which we consider is when the data points in $\mathcal{C}$ are concentrated in K sub-clusters, in the feature space, and points in $\mathcal{B}$ form a disperse manifold composed of the majority of the data points. Therefore, the background dominates the underlying geometry of the data. Let δ be the fraction of data points that lie in $\mathcal{C}$,

$$|\mathcal{C}| = \delta n, \qquad |\mathcal{B}| = (1-\delta)n \tag{3.1}$$

To simplify the analysis, we assume the K sub-clusters are of equal size equal to $\frac{\delta n}{K}$. The result extends to the unequal-size case and the required adjustments are presented in [CM20].

We also assume the inter-cluster connections are weak such that W is close to having $K+1$ blocks. The assumptions will be precisely formulated in Assumption 1 and the connectivity strength will be bound in Theorem 1. Define the matrix W_0 derived from W by removing all connections between $\mathcal{B}$ and $\mathcal{C}$ and let E be the matrix consisting of the edges between $\mathcal{B}$ and $\mathcal{C}$. We consider W as a perturbation of W_0 according to the following dynamic, parameterized by time t,

$$W(t) = W_0 + tE, \qquad t \in [0,1] \tag{3.2}$$

From this point forth, the pre-defined affinity matrix will be denoted as W_1, so that $W(0) = W_0$ and $W(1) = W_1$. This is to prevent any confusion from the time varying matrix $W(t)$ when the time dependence is omitted.

For every point in time, the degree of a node x and the volume of a set A are defined as,

$$d(x,t) = \sum_{y \in \mathcal{V}} W(x,y;t), \qquad \nu(A,t) = \sum_{x \in A} d(x,t) \tag{3.3}$$

We also define lower and upper bounds on the degrees at $t=0$:

$$\underline{d_0} = \min_{x \in \mathcal{V}} d(x,0), \qquad \overline{d_0} = \max_{x \in \mathcal{V}} d(x,0) \tag{3.4}$$

and assume $\underline{d_0} > 0$. By construction of (3.2), the degree $d(x,t)$ of any node monotonically increases over time. Therefore, $\underline{d_0}$ is a global lower bound for all degree nodes at any point in time.

3.2 Embedding Norm

We consider the row-stochastic random-walk matrix $P = D^{-1}W$, where $D \in \mathbb{R}^{n\times n}$ is the diagonal matrix of the row sums of W, $D(x,x) = \sum_{y\in\mathcal{V}} W(x,y)$. Let

$$P\psi_k = \lambda_k\psi_k, \qquad \psi_k^T D\psi_j = \delta_{kj}, \qquad k,j = 1,\dots n \tag{3.5}$$

$\{\lambda_k\}_{k=1}^n$ is the set of eigenvalues of P, $\{\psi_k\}_{k=1}^n$ are the corresponding right eigenvectors and $\delta_{kj} = 1$ when $k = j$ and 0 otherwise.

P is similar to $D^{-1/2}WD^{-1/2}$, which is a real and symmetric matrix, thus has real eigenvalues. Furthermore, P has a Perron eigenvalue of 1. Therefore, the eigenvalues of P are all real and $|\lambda_k| \leq 1$, indexed in decreasing order such that $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

The dependence of $W(t)$ on t, defined in (3.2), results in $P, D, \{\lambda_k\}_{k=1}^n$ and $\{\psi_k\}_{k=1}^n$ also depending on t. The ordering of eigenvalues preserves the inequalities at every time t. Therefore, it does not guarantee continuous change over time.

The *spectral embedding norm* of node $x \in \mathcal{V}$ is defined as:

$$S(x) := S(x;\alpha,\beta) = \sum_{k=1}^{n} f(\lambda_k;\alpha,\beta)\,\psi_k(x)^2 \tag{3.6}$$

where $S(x)$ is the (squared) weighted Euclidean norm of the embedded vector of x in the spectral embedding space using eigenvectors.

In this work, we will primarily focus on the following weight function,

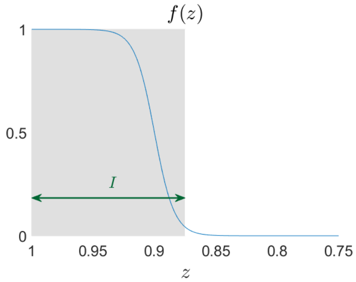
$$f(z) := f(z;\alpha,\beta) = \frac{1}{1+\exp\left(-\frac{z-(1-\alpha)}{\beta}\right)} \tag{3.7}$$

The weight function, $f(z)$, satisfies $|f(z)| \leq 1$ and for $z \in \mathbb{R}$, is a monotonically increasing function taking values in (0,1]. The parameters α and β are user-defined and remain fixed throughout time. Notable limiting function of $f(z)$ is the indicator function obtained as $\lim_{\beta\to 0} f(z) \to \mathbb{1}_{\{\lambda > 1-\alpha\}}$. And for $\lim_{\frac{\alpha}{\beta}\to 0} f(z) \to \exp^{-\frac{1}{\beta}(1-\lambda)}$ which is the heat kernel signature (HKS).

The majority of the theoretical analysis and guarantees, detailed in the next section, apply for a larger range of weight functions. Specifically, any $f(z)$ which is nonnegative, analytic and decays to zero from some eigenvalue until λ_n could be selected for this application.

Let $\varepsilon_0 > 0$ be a near-zero constant and define the domain in which $f(z) \geq \varepsilon_0$ to be I:

$$I := \left[(1-\alpha) + \beta \ln\left(\frac{\varepsilon_0}{1-\varepsilon_0}\right), 1\right]$$



where $\alpha > 0$ and $1/2 > \varepsilon_0$. Note that I includes the leading eigenvalues with high kernel weights. We will denote the set of eigenvalues in I at time $t = 0$ by $\Lambda_I := I \bigcap \{\lambda_k(0)\}_{k=1}^{n}$.

Chapter 4

Theoretical Analysis of Cluster Detection

At $t = 0$ in (3.2), the matrix W_0 has a two block structure, and the spectrum of the graph Laplacian of W_0 also splits into two groups, one residing on $\mathcal{C}$ and the other on $\mathcal{B}$ respectively. However, as t increases, interactions among the eigenvectors develop and the perfect splitting pattern is no longer preserved. The embedding norm varies more stably than individual eigenvectors over time, and serves as a measure by which to separate $\mathcal{C}$ from $\mathcal{B}$ up to time $t = 1$.

4.1 Initial separation of $S(x;\alpha,\beta)$ at $t = 0$

At $t = 0$, the affinity matrix W_0 decomposes into two separated blocks corresponding to $\mathcal{B}$ and $\mathcal{C}$. The block structure of W_0 results in eigenvectors supported either on nodes in $\mathcal{B}$ or $\mathcal{C}$. We denote the set of eigenvectors supported only on $\mathcal{B}$ by $\Psi^{\mathcal{B}}$, and similarly $\Psi^{\mathcal{C}}$. Since we will use $S(\cdot)$ to separate $\mathcal{B}$ and $\mathcal{C}$, we need it to do so at least at $t = 0$, when the two clusters are perfectly separated. Note that this is not guaranteed in cases where the sub-clustering in $\mathcal{C}$ is not perfect or many eigenvalues corresponding to $\Psi^{\mathcal{B}}$ are close to 1. We make the following assumptions:

Assumption 1. *At $t = 0$:*

(a) Λ_I contains K eigenvalues corresponding to K $\mathcal{C}$-eigenvectors and $|\Lambda_I| - K$ $\mathcal{B}$-eigenvectors.

(b) Each of the the K $\mathcal{C}$-eigenvectors (up to rotation) is closely localized on one of the K sub-clusters as follows: There exists $0 \leq \varepsilon_1 \leq 1$ such that for each $\psi \in \Psi^{\mathcal{C}}$, with eigenvalue $\lambda \in \Lambda_I$, there is a unique j, $1 \leq j \leq K$, such that:

$$\frac{1-\varepsilon_1}{\nu(C_j,0)} \leq \psi(x)^2 \leq \frac{1+\varepsilon_1}{\nu(C_j,0)}, \forall x \in C_j$$

$$\psi(x)^2 \leq \frac{\varepsilon_1}{\nu(C,0)}, \forall x \in C \backslash C_j$$

(c) There exists $\varepsilon_2 \geq 0$ such that for any $\psi \in \Psi^{\mathcal{B}}$, with eigenvalue $\lambda \in \Lambda_I$:

$$\psi(x)^2 \leq \frac{1+\varepsilon_2}{\nu(B,0)}, \forall x \in \mathcal{B}$$

The assumption above does not limit the problem as it primarily formulates the typical scenario. At $t=0$, $\mathcal{B}$ and $\mathcal{C}$ are completely disconnected and in the case where the K sub-clusters are also perfectly separated, the first K $\mathcal{C}$-eigenvalues are 1. The subsequent eigenvalues will be strictly less than 1 and depend on the mixing time of the Markov chain (defined by P) within each cluster [LP17]. Since $\mathcal{C}$ consists of a small δ fraction of nodes, even for a non-ideal clustering scenario, the outlier clusters in $\mathcal{C}$ remain localized in the graph. As a result, even when the clustering is not perfect, there still exists a sufficient spectral gap between the first K and the subsequent $\mathcal{C}$-eigenvalues and Λ_I can exclude these eigenvalues. This fulfills (a) and (b). If the K sub-clusters in $\mathcal{C}$ are perfectly separated, (b) is fulfilled for $\varepsilon_1 = 0$. Recall the normalization in (3.5) applies for any $\psi \in \Psi^{\mathcal{B}}$ such that,

$$\sum_{x \in \mathcal{B}} \psi(x)^2 d(x,0) = 1 \tag{4.1}$$

and the leading eigenvector, with eigenvalue 1 is a constant vector with the constant value $\psi(x)^2 = \frac{1}{\nu(B,0)}$. In this situation, the graph Laplacian normalization yields an approximation of the Laplace–Beltrami operator on the submanifold $\mathcal{B}$ [BN03] resulting in the leading $\mathcal{B}$-eigenvectors to be delocalized. Since we assume that $\Lambda_I \ll n$, this fulfills (c) with some small ε_2.

The second assumption is on the proportion of cluster nodes:

Assumption 2. *The constants* $\delta, \varepsilon_0, \varepsilon_1, \varepsilon_2$ *satisfy:*

$$\frac{1-\delta}{\delta}\frac{\underline{d_0}}{\overline{d_0}} > \frac{(1+\varepsilon_2)(|\Lambda_I| - K) + \varepsilon_0(1-\delta)n}{(1-\varepsilon_0)(1-\varepsilon_1)K} \tag{4.2}$$

Theoretically, the above two assumptions guarantee that $S(x;\alpha,\beta)$ separates $\mathcal{C}$ and $\mathcal{B}$ by a margin, g_0, at $t = 0$. We now make a few remarks on the implications of Assumption 2.

Remark 1. Since $|\Lambda_I| \sim K$ and $\delta \ll 1$, the dominating term in Assumption 2 is $(1-\delta)\varepsilon_0 n$. δ and n are data dependent constants whereas ε_0 depends on the decay of $f(z)$.

Remark 2. The condition imposed in Assumption 2 implies that $\varepsilon_0 = O\left(n^{-1}\right)$. The assumption that ε_0 is $O(n^{-1})$ is acceptable since the dependency of β on ε_0 requires β to be $O\left((\ln n)^{-1}\right)$ which is a very relaxed requirement compared to relying on a significant spectral gap traditionally assumed. Now, the data set mainly consists of points that originated from the background manifold but also points that originate from other manifolds generating the points in the clusters. And in practice, the data points may deviate slightly from the manifolds due to noise. This aspect, where not all points lie exactly on the manifold, has been investigated [HAVL05, Sin06, CL06] and it has been shown that for a finite data set with a predefined kernel scaling parameter, the spectral gap of the Laplacian is $O(n^{-1/2})$.

A prototypical case where Assumption 1 and 2 are satisfied is the example in Section 5.1, Fig.5.2(a). For the example shown, $K = 10$ and $n = 5 \cdot 10^3$. In this scenario, where $\alpha \approx 0.093$ and $\beta = 0.005$, We obtain maximal separation, measured by g_0, for $\varepsilon_0 = 6 \cdot 10^{-4}$.
Assuming that at $t = 0$, the separation of the clusters within $\mathcal{C}$, and $\mathcal{B}$ from $\mathcal{C}$ result in ε_1 and ε_2 being small constants. Furthermore, in such cases, if the graph has a balanced degree, i.e.

$\underline{d_0} \approx \overline{d_0}$, then Eq.(4.2) is satisfied $\forall \delta < 0.091$, and specifically, for $\delta = 0.02$. For datasets with higher values of δ, the theoretical analysis requires adjusting ε_0, to decrease $|\Lambda_I|$ and comply with Assumption 2. Nevertheless, we have observed that in practice the smooth embedding norm $S(x;\alpha,\beta)$ can successfully separate $\mathcal{C}$ from $\mathcal{B}$ for higher values of δ, as shown in Section 5.2.

Theoretically, the above two assumptions guarantee that the embedding norm $S(x;\alpha,\beta)$ separates the clusters $\mathcal{C}$ and $\mathcal{B}$ at time $t = 0$, by at least a quantity denoted by g_0 in addition with an upper bound of $S(x;\alpha,\beta)$ over $\mathcal{V}$:

Proposition 1 (Initial separation by $S(x;\alpha,\beta)$)**.** *Under Assumption 1 at time $t = 0$:*

$$(1-\varepsilon_0)(1-\varepsilon_1)\frac{K}{\delta n \overline{d_0}} \leq S(x;\alpha,\beta) \leq \frac{(1+\varepsilon_1)K}{\underline{d_0}\delta n} + \frac{\varepsilon_1}{\underline{d_0}\delta n}(K-1) + \frac{\varepsilon_0}{\underline{d_0}},\ \forall x \in \mathcal{C} \tag{4.3}$$

$$S(x;\alpha,\beta) \leq \frac{1+\varepsilon_2}{\underline{d_0}n(1-\delta)}(|\Lambda_I| - K) + \frac{\varepsilon_0}{\underline{d_0}},\ \forall x \in \mathcal{B} \tag{4.4}$$

$$\sup_{x\in\mathcal{V}} S(x;\alpha,\beta) < \frac{K}{\delta n \underline{d_0}}\left(1 + \varepsilon_0 \frac{\delta n}{K} + 2\varepsilon_1\right) \tag{4.5}$$

If Assumption 2 holds as well, then the initial gap between data points in $\mathcal{B}$ and $\mathcal{C}$ is at least

$$\begin{aligned} g_0 =& \frac{K}{\delta n \underline{d_0}} \cdot \blacklozenge \\ \blacklozenge :=& \frac{\underline{d_0}}{\overline{d_0}}(1-\varepsilon_0)(1-\varepsilon_1) - \frac{\delta}{1-\delta}\frac{|\Lambda_I| - K}{K}(1+\varepsilon_2) - \varepsilon_0\frac{\delta n}{K} \end{aligned} \tag{4.6}$$

that is, $\forall x \in \mathcal{C}$ and $y \in \mathcal{B}$, $S(x;\alpha,\beta) - S(y;\alpha,\beta) \geq g_0 > 0$.

Concluding from Remarks 1 and 2, the leading term in $\blacklozenge$ is approximately $\frac{\underline{d_0}}{\overline{d_0}}(1 - O(\delta))$. This is under the assumptions that $\varepsilon_0, \varepsilon_1$ and ε_2 are small constants, $|\Lambda_I|$ is $O(K)$ and ε_0 is $O(n^{-1})$ (see Remark 2). Furthermore, in such cases, if the graph has balanced degree, i.e., $\underline{d_0} \approx \overline{d_0}$, then $\blacklozenge \approx 1 - O(\delta)$. Resulting in (4.6) to be approximately $g_0 \approx \frac{K}{\delta n \underline{d_0}}$.

4.2 Evolution of $S(x;\alpha,\beta)$

We will prove the stability of $S(x,t;\alpha,\beta)$ over time and show that this quantity can be bounded for every time t. This is an extension of previous work developed by Cheng and Mishne [CM20]. In the existing framework, the separation of $\mathcal{C}$ and $\mathcal{B}$ is guaranteed under certain limitations on the graph spectral properties. The main limitation being an eigen-gap that partitions the graph-Laplacian spectra into two disjoint sets such that no eigen-crossings are permitted between them. Said eigen-gap is time dependant and must remain positive at every time t. The difficulty in this constraint is that the eigenvalues are differentiable with respect to t for $0 \leq t \leq 1$. As the eigenvalues change continuously over time, eigen-crossings do occur due to the strengthening connections between clusters and generally there is no control on the speed of change.

This work eliminates the requirements of an eigen-gap and only depends on the connection strength between $\mathcal{C}$ and $\mathcal{B}$. We define the following measure of the connection between $\mathcal{C}$ and $\mathcal{B}$:

$$C := \sum_{\substack{x\in\mathcal{C} \\ y\in\mathcal{B}}} W(x,y) \tag{4.7}$$

Let $S(x,t)$ denote the variable $S(x,t;\alpha,\beta)$ which also depends on time. The following proposition defines the change in $S(x,t)$ for every $0 \leq t \leq 1$:

Proposition 2 (Evolution of the embedding norm)**.**

$$\begin{aligned}\dot{S}(x;\alpha,\beta) = &\sum_{\substack{k=1 \\ \lambda_k=\lambda_j}}^{n}\sum_{j=1}^{n} \left(\psi_k^T \left(f'(\lambda_k)\dot{W} - (zf(z))'(\lambda_k)\dot{D}\right)\psi_j\right)\psi_k(x)\psi_j(x) \\ &+ \sum_{\substack{k=1 \\ \lambda_k\neq\lambda_j}}^{n}\sum_{j=1}^{n} \left(\psi_k^T \left(\frac{f(\lambda_k)-f(\lambda_j)}{\lambda_k-\lambda_j}\dot{W} - \frac{\lambda_k f(\lambda_k)-\lambda_j f(\lambda_j)}{\lambda_k-\lambda_j}\dot{D}\right)\psi_j\right)\psi_k(x)\psi_j(x)\end{aligned} \tag{4.8}$$

This leads to the main result in this thesis:

Theorem 1 (Separation at $t = 1$)**.** *Under assumptions 1 and 2 and if the following condition is satisfied,*

$$\frac{C}{\underline{d_0}} \leq \frac{1}{4}\frac{1}{1+\frac{2}{\beta}}\ln\left(1+\frac{1}{2}\frac{\blacklozenge}{\left(1+\varepsilon_0\frac{\delta n}{K}+2\varepsilon_1\right)}\right) \tag{4.9}$$

where $\blacklozenge$ *is defined in* (4.6)*, then* $\mathcal{C}$ *and* $\mathcal{B}$ *can be separated by thresholding the embedding norm of points in* $\mathcal{V}$ *at* $t = 1$*:*

$$S(x;\alpha,\beta) > \eta, \quad \forall x \in \mathcal{C}$$

$$S(x;\alpha,\beta) < \eta, \quad \forall x \in \mathcal{B}$$

In practise, η can be set to the $1-\delta$ quantile of $S(x)$ of all nodes.

A remark on the condition of Theorem 1:

Remark 3. The analysis requires C to be small compared to the minimal node degree. We note that this condition may be much stronger than encountered in applications due to the reliance on a lower bound on the initial gap g_0 and the upper bounds derived in (4.9) which are not tight. Further work can be done to improve the upper bound in (4.9) as we show in the demonstrated practical applications in Chapter 5 that the embedding norm can separate $\mathcal{C}$ and $\mathcal{B}$ under more relaxed circumstances.

Chapter 5

Experimental Results

We illustrate the theoretical analysis of Section 4 with a synthetic data set and apply the spectral embedding norm to simulated data from [NG07]. Anomalies are identified by applying a threshold η to the empirical spectral embedding norms, where $\eta = (1-\delta) - quantile$ of the calculated values of S.

To evaluate the accuracy of the anomaly detection, we calculate the F_1 score:

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision + recall}} = \frac{TP}{TP + \frac{1}{2}(FP+FN)}$$

where TP, FP and FN stand for True Positive, False Positive and False Negative respectively.

5.1 Manifold toy example

In the first simulated data set, n points $\{x_i\} \subset \mathbb{R}^2$ are random samples from a manifold-like background $\mathcal{B}$ and clusters in $\mathcal{C}$. The data points are randomly selected according to the following distributions. Each data point $x_{\mathcal{B}} \in \mathcal{B}$ is the sum of two independent random variables, $x_{\mathcal{B}} = y_{\mathcal{B}} + n_{\mathcal{B}}$ where $y_{\mathcal{B}}$ is selected uniformly from the unit sphere and $n_{\mathcal{B}} \sim \mathcal{N}(0, \sigma_{\mathcal{B}}^2 I)$ with $\sigma_{\mathcal{B}} = 0.01$. Each data point $x_{\mathcal{C}} \in \mathcal{C}$ is sampled independently from $\mathcal{N}(\mu_j, \sigma_{\mathcal{C}}^2 I)$ with $\sigma_{\mathcal{C}} = 0.02$, where $1 \leq j \leq K$ is the index of the sub-cluster to which $x_{\mathcal{C}}$ belongs. The cluster centers $\left\{\mu_j\right\}_{j=1}^{K}$ are also selected uniformly at random at a fixed distance of $r = 1.08$ from the origin.

Fig. 5.1 presents a typical realization of the dataset with $n = 5000$ datapoints, $\delta = 0.02$ and $K = 10$ clusters. From 5.1(c) it can be seen that there is not a clear eigen-gap between λ_{10} and λ_{11} and the $F1$ score with the indicator kernel, in 5.1(f), indicates that selecting $I_\alpha = 10$ does not successfully separate $\mathcal{C}$ from $\mathcal{B}$. For such a case, where $\mathcal{C}$ and $\mathcal{B}$ are connected, and δ is small, the informative eigenvectors, on the location of $\mathcal{C}$, correspond to smaller eigenvalues (with high index). The presented example achieves the highest $F1$ score for $\alpha \approx 0.093$. This clear separation is observed through the embedding norms $S(x,\alpha), \forall x \in \mathcal{V}$, in 5.1(d). The separation of $\mathcal{C}$ from $\mathcal{B}$ using different kernels (indicator, sigmoid, HKS) $f(\lambda_k,\alpha)$ in 50 different datasets (with the same parameter settings) are compared in 5.1(f) and 5.1(g). The stability of the sigmoid kernel is advantageous as it is observed to be the least sensitive to algorithmic parameter choices.

5.2 Data sampled from a mixture

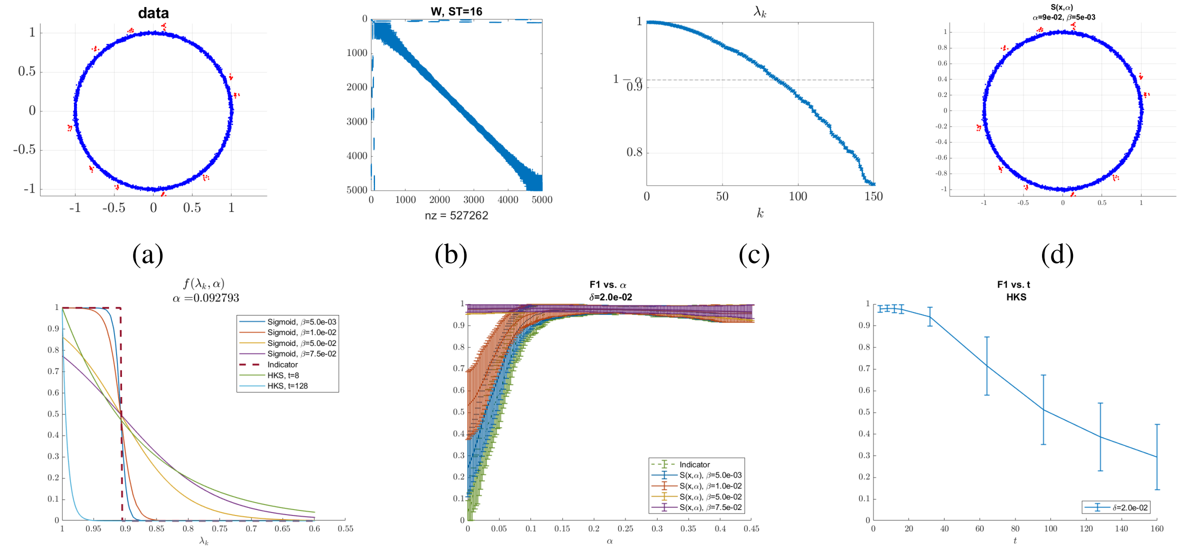
The simulated data is sampled from a mixture of two densities in $\mathbb{R}^2$:

$$p(x) = p(x_1,x_2) = \theta p_G(x_1,x_2) + (1-\theta) p_{L,\varepsilon}(x_1,x_2) \tag{5.1}$$

$p_{L,\varepsilon}$ denotes a uniform density in the rectangle $\Omega = \left\{ (x_1,x_2) | 0 \leq x_1 \leq L, -\varepsilon < x_2 \leq 0 \right\}$, p_G denotes a Gaussian density with expectation $\mu = (\mu_1,\mu_2)$ and covariance matrix $\Sigma = \rho^2 I$ and $0 \leq \theta \leq 1$.

The parameters in the following examples presented are $L = 8, \varepsilon = 0.05, \mu = (2,0.2), \rho = 0.1, n = 1400$ data-points and $K = 1$ where the background manifold, $\mathcal{B}$, is the rectangular strip Ω and the cluster, $\mathcal{C}$, is the Gaussian ball. Since the data-points are randomly sampled from the two densities according to Bern(θ) distribution, $\delta \approx \theta$.

This dataset was introduced in [NG07] to demonstrate the limitations of the common



(a) (b) (c) (d)

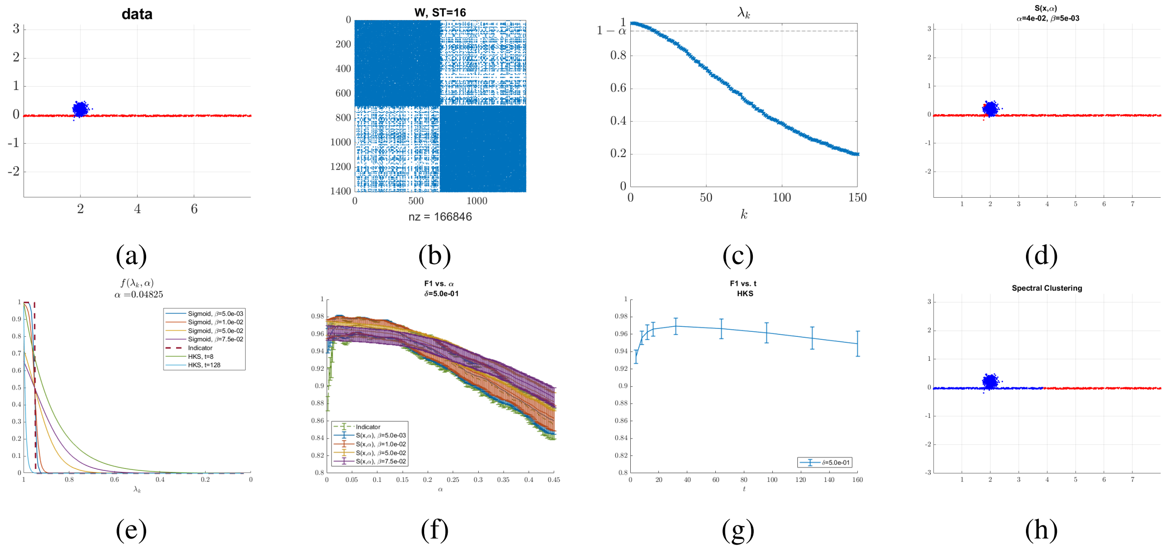
(e) (f) (g)

Figure 5.1: Separation of $\mathcal{C}$ from $\mathcal{B}$. (a) Example of the complete data set with $n = 5 \cdot 10^3$ points. A $\delta = 0.02$ portion of the points form $K = 10$ clusters outside the ring. A $1-\delta$ portion of the points lie near a ring and form the background ($\mathcal{B}$). (b) The affinity matrix W_1. (c) The leading 150 eigenvalues of the transition matrix P. (d) Plot of $S(x;\alpha,\beta)$ for all data points. (e) Plot of kernel weight functions $f(\lambda;\alpha,\beta)$ used for comparison. (f) Mean and standard deviation of F_1 score, measuring the separation of $\mathcal{C}$ and $\mathcal{B}$ by thresholding $S(x;\alpha,\beta)$ using the $1-\delta$ quantile of values. The score is measured for sigmoid and indicator kernel functions, for varying values of α and β. (g) Mean and standard deviation of F_1 score using the HKS kernel function.

graph-based spectral clustering methods. It was shown that for $\theta = 0.5$, spectral clustering based on the second eigenvector of the normalized random-walk Laplacian matrix, method described in [SM00], does not partition $\mathcal{C}$ from $\mathcal{B}$ due to the overlap of points belonging to different densities. Yet, satisfactory separation of the two clusters is not presented in [NG07].

5.2.1 θ=0.5

In fig. 5.2(f) and 5.2(g) we observe that even for large value of δ, the spectral embedding norm is a stable measure w.r.t. the kernel parameters, for separating the points originating from different densities. A comparison between spectral clustering and weighted embedding norm can be seen through the clustering of the data, shown in 5.3(d) and 5.2(h). Spectral clustering partitions the points somewhere along the long strip whereas the embedding norm identifies that the Gaussian ball is separated from the rectangular strip.



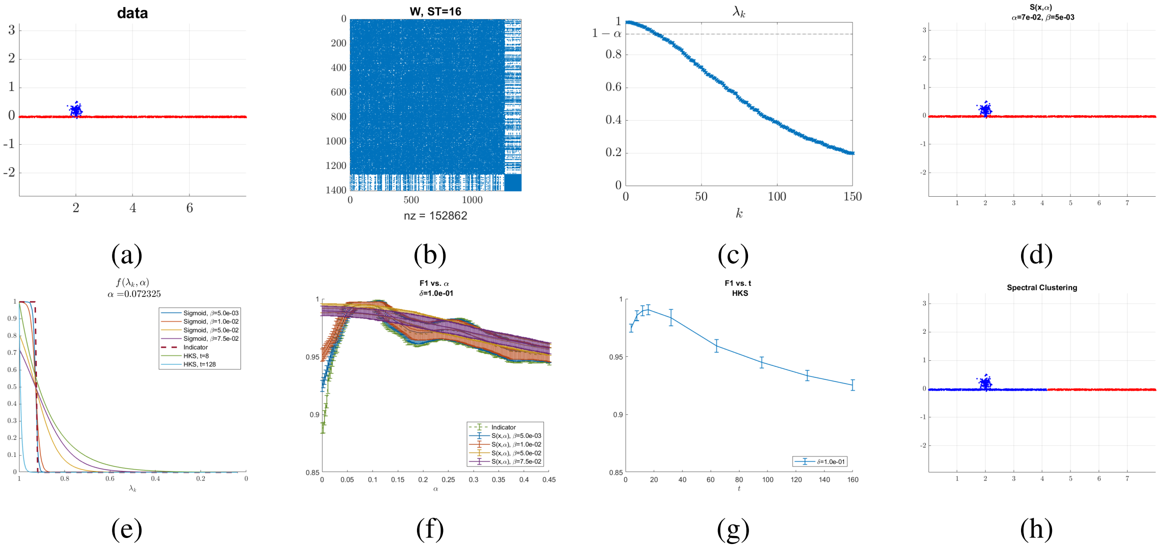
(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5.2: Separation of $\mathcal{C}$ from $\mathcal{B}$ as defined in [NG07]. (a) Example of the complete data set with $n = 1400$ points. A $\delta \approx 0.5$ portion of the points form $K = 1$ cluster above the strip. A $1-\delta$ portion of the points lie near a strip and form the background ($\mathcal{B}$). (b) The affinity matrix W_1. (c) The leading 150 eigenvalues of the transition matrix P. (d) Plot of $S(x;\alpha,\beta)$ for all data points. (e) Plot of kernel weight functions $f(\lambda;\alpha,\beta)$ used for comparison. (f) Mean and standard deviation of F_1 score, measuring the separation of $\mathcal{C}$ and $\mathcal{B}$ by thresholding $S(x;\alpha,\beta)$ using the $1-\delta$ quantile of values. The score is measured for sigmoid and indicator kernel functions, for varying values of α and β. (g) Mean and standard deviation of F_1 score using the HKS kernel function. (h) The result of spectral clustering [SM00].

5.2.2 θ=0.1

For the unbalanced case of $\theta = 0.1$, as expected, with a decrease in δ, we need to use information that is ”deeper” in the spectrum for satisfactory clustering. Similarly to 5.2.1, the embedding norm is shown to partition the points between the strip and the Gaussian ball, fig. 5.3(d), as expected and improving on the spectral clustering, shown in fig. 5.3(h). Concluding that aggregating information beyond the second eigenvector is beneficial for such a scenario with highly connected clusters.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5.3: Separation of $\mathcal{C}$ from $\mathcal{B}$ as defind in [NG07]. (a) Example of the complete data set with $n = 1400$ points. A $\delta \approx 0.1$ portion of the points form $K = 1$ cluster above the strip. A $1-\delta$ portion of the points lie near a strip and form the background ($\mathcal{B}$). (b) The affinity matrix W_1. (c) The leading 150 eigenvalues of the transition matrix P. (d) Plot of $S(x;\alpha,\beta)$ for all data points. (e) Plot of kernel weight functions $f(\lambda;\alpha,\beta)$ used for comparison. (f) Mean and standard deviation of F_1 score, measuring the separation of $\mathcal{C}$ and $\mathcal{B}$ by thresholding $S(x;\alpha,\beta)$ using the $1-\delta$ quantile of values. The score is measured for sigmoid and indicator kernel functions, for varying values of α and β. (g) Mean and standard deviation of F_1 score using the HKS kernel function. (h) The result of spectral clustering [SM00].

5.3 Defect Detection

The image data are real-world industrial inspection scenarios intended for quality control and available at [BFSS19]. This data is designed for optical inspection of textured surfaces of carpets. The data set consists of 89 RGB images of size 1024×1024 pixels with 5 different types of defects, as detailed in Table 5.1. For computational feasibility, the images are first downsampled to the size 256×256 pixels. The next image preprocessing steps vary according to the methods applied:

- **Spectral Embedding Norm -** We extract image patches of size 8×8, with a stride of 1 and apply self-tuning with $K = 16$. Detections are found by applying a threshold to $S(x;\alpha,\beta)$.

- **L2 and SSIM Autoencoder -** We evaluate our algorithm's performance in comparison to that of the convolutional autoencoder architecture as described by Bergmann *et al.*[BLF$^+$18], employing either a per-pixel l_2 loss or a loss based on the structural similiarity index (SSIM). We use the publicly available implementation on GitHub.[*] Image patches are of size 128 × 128, with a stride of 30. Detections are found by applying a threshold to the residual images.

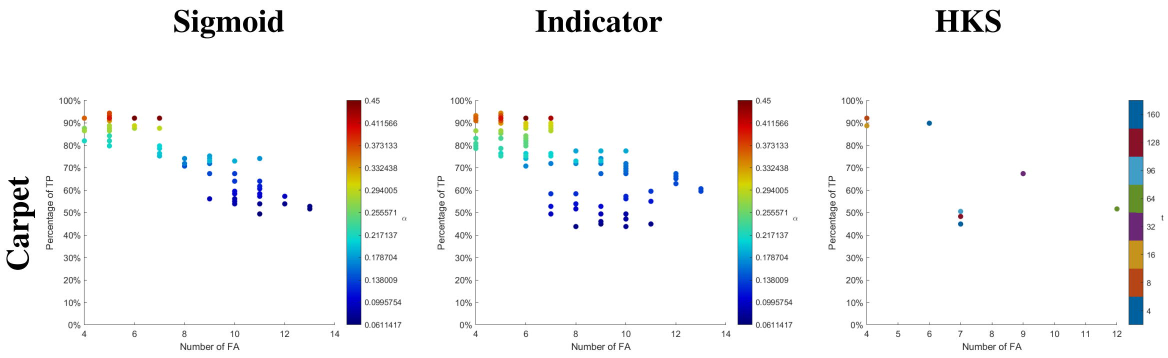
A detection is a connected component (CC) in the binary image (after thresholding), where a CC containing a defect is a true positive (TP) and any other CC's are counted as false alarms (FA). We count all detections on the acrpet defects as a single TP for a given image, whereas there can be more than one FA in an image. The size of the CC can be used to reject noisy detections by discarding small CCs. We threshold the area of the CC to be at least 10 pixels. Using a larger threshold on the size rejects more FAs, but can also result in a decreased amount of TPs, for small sized anomalies. We compare the percentage of TPs for each method for a given FA rate. Results comparing the performance of different kernel weights are given in Fig. 5.4.

Table 5.1: Carpet images data set description.

Defect type	Number of Images
Color	19
Cut	17
Hole	17
Metal Contamination	17
Thread	19
Total	89

Fig. 5.5 presents a visual comparison of selected images, from each defect category, prior to applying a detection threshold. For the scenarios present in the data set, the spectral embedding norm can outperform the Autoencoder performance on detecting small clusters. The embedding norm removes the background in a much ”cleaner” fashion in comparison to the autoencoder.

[*] www.github.com/AdneneBoumessouer/MVTec-Anomaly-Detection



(a) (b) (c)

Figure 5.4: Percentage of true positives for given number of false alarms for detections of size greater than 10 pixels for different kernel functions.

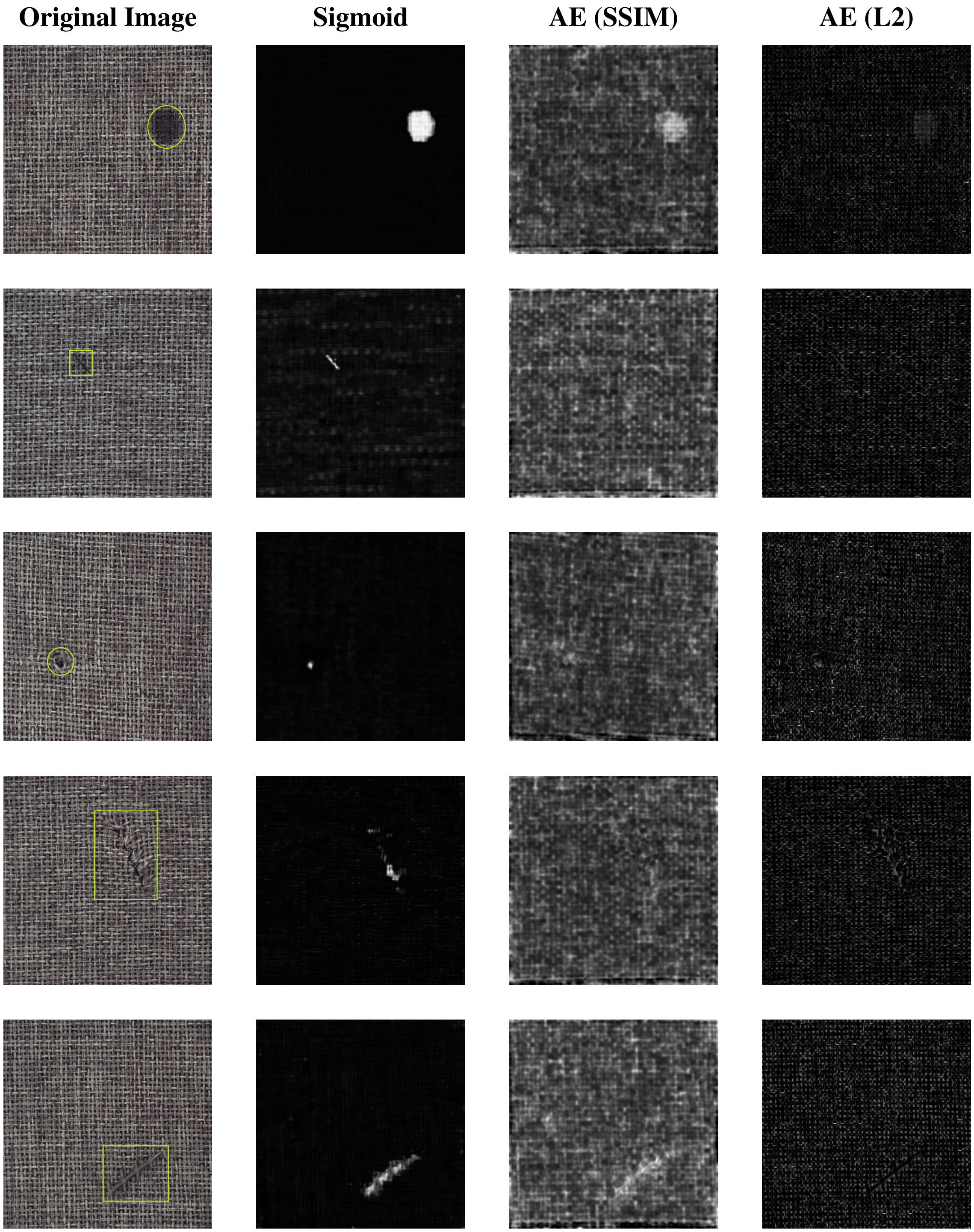


Figure 5.5: Anomaly detection comparison between spectral embedding norm and autoencoder residual images with SSIM and l_2 loss on exemplary images from each carpet defect category.

Chapter 6

Conclusion and Future Work

This thesis introduces a new method that addresses the problem of data segmentation. By focusing on the theoretical analysis, we prove that this quantity can be used to separate clusters from the background in unbalanced settings, including extreme cases such as outlier detection. Our result thus provides a way to go beyond dominating eigenvectors of the graph Laplacian to unbalanced data clustering tasks with theoretical verification.

In Chapter 3, the fundamental tool used to solve the detection problem at hand is presented: a generalization of the original spectral embedding norm. The new definition considers a weighted sum of the contributing eigenvectors of the graph Laplacian. With only limited constraints on the weight function, we can incorporate a wide family of spectral shape descriptors that can successfully be tailored to various applications.

In Chapter 4, a detailed theoretical analysis of the proposed anomaly detection scheme is proved, under generic assumptions. The main result of the thesis is Theorem 1 which is proved in the Appendix. An extension of this work that can be researched, is to provide the Separation Theorem under an extended analysis which relaxes the constraints of the weight function. Specifically, removing the demand of a monotonic function allows a nonconsecutive selection of eigenvectors which contribute to the sum of the spectral embedding norm. We conjecture that this can provide a guarantee of greater separation between the background and clusters under a weaker connectivity condition.

Finally, in Chapter 5, the experimental results demonstrate the robustness of the algorithm in multiple data sets. We begin by applying the method on a synthetic example, adopted from prior work introducing the spectral embedding norm. This establishes the achievable stability of the spectral embedding norm when adapting the weight function to the user application. Next, we demonstrate the superiority of the algorithm by calculating the spectral embedding norm on a data set determined to be complex for spectral methods. The successful and stable detection of outliers, for varying examples, indicates that the theoretical conditions are more restrictive than what occurs in practical applications. Furthermore, this implies that the spectral embedding norm can be used in a broad range of applications with stronger inter-cluster connections.

Appendix A

Proof of Proposition 1

Proof. Proof of (4.3). At $t = 0$, the eigenvectors are supported either on $\mathcal{B}$ or on $\mathcal{C}$. Therefore, $\forall x \in \mathcal{C}$:

$$S(x;\alpha,\beta) = \sum_{\psi_k \in \Psi^{\mathcal{C}}} f(\lambda_k;\alpha,\beta)\,\psi_k(x)^2 \tag{A.1}$$

We can partition the eigenvalues to two groups: eigenvalues in Λ_I and eigenvalues not in Λ_I.

$$S(x;\alpha,\beta) = \sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \in \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 + \sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \notin \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 \tag{A.2}$$

Next, we will bound each sum separately. For the well-separated case at $t = 0$, $\forall \lambda_k \in \Lambda_I$, $(1 - \varepsilon_0) \leq f(\lambda_k) \leq 1$.

$$(1-\varepsilon_0) \sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \in \Lambda_I}} \psi_k(x)^2 \leq \sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \in \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 \leq \sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \in \Lambda_I}} \psi_k(x)^2 \tag{A.3}$$

By Assumption 1(b), each eigenvector in $\Psi^{\mathcal{C}}$, with eigenvalue in Λ_I, is associated with a sub-cluster $C_j,\ 1 \le j \le K$ and since the rotation preserves the squared sum,

$$\sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \in \Lambda_I}} \psi_k(x)^2 = \sum_{j=1}^{K} \psi_j(x)^2 \tag{A.4}$$

Suppose that $x \in C_{j_x}$, with associated eigenvector ψ_{j_x}. Then by Assumption 1(b),

$$\frac{1-\varepsilon_1}{\nu(C_{j_x},0)} \le \sum_{j=1}^{K} \psi_j(x)^2 \le \frac{1+\varepsilon_1}{\nu(C_{j_x},0)} + (K-1)\frac{\varepsilon_1}{\nu(C,0)} \tag{A.5}$$

For any sub-cluster, C_j, $1 \le j \le K$

$$\nu(C_j,0) = \sum_{x \in C_j} d(x,0) \le \overline{d_0}|C_j| = \overline{d_0}\frac{\delta n}{K} \tag{A.6}$$

Combined with the trivial lower bound of $\sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \notin \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 \ge 0$, this yields the lower bound in (4.3).

Similarly, $\nu(C_j,0) = \sum_{x \in C_j} d(x,0) \ge \underline{d_0}|C_j| = \underline{d_0}\frac{\delta n}{K}$ and $\nu(C,0) \ge \underline{d_0}\delta n$. Substituted in the right hand side of (A.3) gives

$$\sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \in \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 \le (1+\varepsilon_1)\frac{K}{\underline{d_0}\delta n} + (K-1)\frac{\varepsilon_1}{\underline{d_0}\delta n} \tag{A.7}$$

By definition of Λ_I, for all $\lambda_k \notin \Lambda_I$, $f(\lambda_k) \le \varepsilon_0$. Furthermore, recall the normalization defined in (3.5), $\forall x \in \mathcal{V}$,

$$\sum_{k=1}^{n} \psi_k(x)^2 = \frac{1}{d(x)} \le \frac{1}{\underline{d_0}} \tag{A.8}$$

giving $\sum_{\substack{\psi_k \in \Psi^{\mathcal{C}} \\ \lambda_k \notin \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 \le \frac{\varepsilon_0}{\underline{d_0}}$. Combined with (A.7), this gives the upper bound in (4.3).

The embedding norm of any $x \in \mathcal{B}$ can be similarly split into two sums,

$$S(x;\alpha,\beta) = \sum_{\substack{\psi_k \in \Psi^{\mathcal{B}} \\ \lambda_k \in \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 + \sum_{\substack{\psi_k \in \Psi^{\mathcal{B}} \\ \lambda_k \notin \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 \tag{A.9}$$

By Assumption 1(c), at $t = 0$,

$$\sum_{\substack{\psi_k \in \Psi^{\mathcal{B}} \\ \lambda_k \in \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 \leq \frac{1+\varepsilon_2}{\nu(B,0)} \sum_{\substack{\psi_k \in \Psi^{\mathcal{B}} \\ \lambda_k \in \Lambda_I}} f(\lambda_k) \leq \frac{1+\varepsilon_2}{\nu(B,0)}(|\Lambda_I| - K) \tag{A.10}$$

where we have used that $f(\lambda_k) \leq 1$ for all λ_k and by Assumption 1(a), there are $|\Lambda_I| - K$ eigenvalues in Λ_I corresponding to $\mathcal{B}$-eigenvectors. We further bound $\nu(B,0)$,

$$\nu(B,0) = \sum_{x \in \mathcal{B}} d(x,0) \geq \underline{d_0}|\mathcal{B}| = \underline{d_0}(1-\delta)n \tag{A.11}$$

Substituting in (A.10) and applying the same justification as for the proof of (4.3) for $\sum_{\substack{\psi_k \in \Psi^{\mathcal{B}} \\ \lambda_k \notin \Lambda_I}} f(\lambda_k)\,\psi_k(x)^2 \leq \frac{\varepsilon_0}{\underline{d_0}}$, proves (4.4).

This concludes the proof of Proposition 1 since Assumption 2 guarantees that the lower bound of (4.3) is greater than the upper bound of (4.4), and then (4.5) and (4.6) directly follow.

□

Appendix B

Proof of Proposition 2

Proof. We prove the validity of (4.8) through the contour integral over the resolvent.

For $z \in \mathbb{C}$ and not an eigenvalue of P define

$$R(z) = (W - zD)^{-1}$$

where the time dependence is omitted. Recall from (3.5), the right eigenvectors of $P = D^{-1}W$, Ψ, satisfy $W\Psi = D\Psi\Lambda$ where $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$ and the columns of Ψ are D normalized, *i.e.* $\Psi^T D\Psi = I$. One can verify the equivalent form of $R(z)$ as

$$R(z) = \left(D\Psi\Lambda\Psi^T D - zD\right)^{-1} = \Psi\left(\Lambda - zI\right)\Psi^T = \sum_{k=1}^{n} \frac{\psi_k \psi_k^T}{\lambda_k - z}$$

Let $\mathbf{S} = \sum_{k=1}^{n} f\left(\lambda_k\right) \psi_k \psi_k^T$, then

$$\mathbf{S} = -\frac{1}{2\pi i} \oint_{\Gamma} f\left(z\right) R(z) dz \tag{B.1}$$

where the contour Γ is a simple closed contour such that all the eigenvalues of P are inside Γ throughout time t and singularities of $f\left(z\right)$, obtained at $z = (1-\alpha) - i\pi\beta(1+2\pi m), m \in \mathbb{Z}$, are

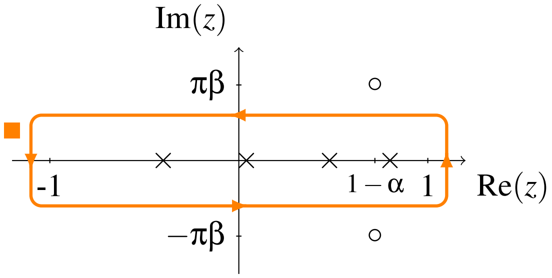


Figure B.1: Figure of contour Γ. $\times$ indicate eigenvalues and $\circ$ indicate singularity points of $f(z,\alpha)$, $z \in \mathbb{C}$.

outside of Γ for all t (Figure B.1).

Differentiating both sides of (B.1) w.r.t. t:

$$\dot{\mathbf{S}} = -\frac{1}{2\pi i}\oint_{\Gamma} f\,(z)\,\dot{R}(z)dz \tag{B.2}$$

And by differentiating

$$(W - zD)\,R(z) = I \tag{B.3}$$

One obtains,

$$\dot{R}(z) = -R(z)\left(\dot{W} - z\dot{D}\right)R(z)$$

Substituting back to (B.2):

$$
\begin{aligned}
\dot{\mathbf{S}} &= \frac{1}{2\pi i}\oint_{\Gamma} f(z)\, R(z)\left(\dot{W} - z\dot{D}\right) R(z) dz \\
&= \frac{1}{2\pi i}\oint_{\Gamma} f(z)\left(\sum_{k=1}^{n}\frac{\psi_k\psi_k^T}{\lambda_k - z}\right)\left(\dot{W} - z\dot{D}\right)\left(\sum_{j=1}^{n}\frac{\psi_j\psi_j^T}{\lambda_j - z}\right) dz \\
&= \frac{1}{2\pi i}\oint_{\Gamma} f(z)\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{\psi_k\psi_k^T\left(\dot{W} - z\dot{D}\right)\psi_j\psi_j^T}{(\lambda_k - z)\left(\lambda_j - z\right)} dz \\
&= \frac{1}{2\pi i}\oint_{\Gamma} f(z)\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{\psi_k^T\left(\dot{W} - z\dot{D}\right)\psi_j}{(\lambda_k - z)\left(\lambda_j - z\right)}\psi_k\psi_j^T dz \\
&= \sum_{k=1}^{n}\sum_{j=1}^{n}\frac{1}{2\pi i}\oint_{\Gamma} f(z)\frac{\psi_k^T\left(\dot{W} - z\dot{D}\right)\psi_j}{(\lambda_k - z)\left(\lambda_j - z\right)}\psi_k\psi_j^T dz \\
&= \sum_{k=1}^{n}\sum_{j=1}^{n}\left(\psi_k^T\left(\alpha_{kj}\dot{W} - \beta_{kj}\dot{D}\right)\psi_j\right)\psi_k\psi_j^T
\end{aligned}
\tag{B.4}
$$

where

$$
\alpha_{kj} = \frac{1}{2\pi i}\oint_{\Gamma}\frac{f(z)}{(\lambda_k - z)\left(\lambda_j - z\right)} dz \qquad \beta_{kj} = \frac{1}{2\pi i}\oint_{\Gamma}\frac{z f(z)}{(\lambda_k - z)\left(\lambda_j - z\right)} dz
$$

By Cauchy's integral formula:

$$
\alpha_{kj} = \begin{cases} f'(z)|_{z=\lambda_k} & \lambda_k = \lambda_j \\ \frac{f(\lambda_k) - f(\lambda_j)}{\lambda_k - \lambda_j} & \lambda_k \neq \lambda_j \end{cases} \tag{B.5}
$$

$$
\beta_{kj} = \begin{cases} (z f(z))'|_{z=\lambda_k} & \lambda_k = \lambda_j \\ \frac{\lambda_k f(\lambda_k) - \lambda_j f(\lambda_j)}{\lambda_k - \lambda_j} & \lambda_k \neq \lambda_j \end{cases} \tag{B.6}
$$

Substituting back to (B.4):

$$
\begin{aligned}
\dot{\mathbf{S}} = & \sum_{k=1}^{n} \sum_{\substack{j=1 \\ \lambda_k = \lambda_j}}^{n} \left(\psi_k^T \left(f'(\lambda_k) \dot{W} - (f(\lambda_k) + \lambda_k f'(\lambda_k)) \dot{D} \right) \psi_j \right) \psi_k \psi_j^T \\
& + \sum_{k=1}^{n} \sum_{\substack{j=1 \\ \lambda_k \neq \lambda_j}}^{n} \left(\psi_k^T \left(\frac{f(\lambda_k) - f(\lambda_j)}{\lambda_k - \lambda_j} \dot{W} - \frac{\lambda_k f(\lambda_k) - \lambda_j f(\lambda_j)}{\lambda_k - \lambda_j} \dot{D} \right) \psi_j \right) \psi_k \psi_j^T
\end{aligned} \tag{B.7}
$$

Since $S(x;\alpha,\beta) = \mathbf{S}(x,x)$, the claim follows by evaluating (B.7) at the entry (x,x) on both sides. □

Appendix C

Proof of Theorem 1

Proof. An initial gap at $t = 0$, of value g_0, between the embedding norms of points in $\mathcal{C}$ and $\mathcal{B}$ is guaranteed under Assumptions 1 and 2. We will show that a gap is preserved throughout time, for any $0 \leq t \leq 1$, depending on the connection strength between $\mathcal{C}$ and $\mathcal{B}$, as formulated in condition (4.9). This will be done by bounding the perturbation of the embedding norms, derived in Proposition 2.

Introducing the notation:

$$\bar{S}(t) := \sup_{x \in \mathcal{V}} S(x,t;\alpha,\beta) \tag{C.1}$$

We will bound the change in size of $S(x,t)$, for any $x \in \mathcal{V}$:

$$|S(x,t) - S(x,0)| \leq \int_0^t \left|\dot{S}(x,\tau)\right| d\tau \tag{C.2}$$

First, we bound the integrand,

$$
\begin{aligned}
\left|\dot{S}(x,t)\right| \leq & \sum_{\substack{k=1 \\ \lambda_k=\lambda_j}}^{n} \sum_{j=1}^{n} \left(\left|f'(\lambda_k)\right| \left|\psi_k^T \dot{W} \psi_j\right| + \left|(zf(z))'(\lambda_k)\right| \left|\psi_k^T \dot{D} \psi_j\right| \right) \left|\psi_k(x)\right| \left|\psi_j(x)\right| \\
& + \sum_{\substack{k=1 \\ \lambda_k\neq\lambda_j}}^{n} \sum_{j=1}^{n} \left(\left|\frac{f(\lambda_k)-f(\lambda_j)}{\lambda_k-\lambda_j}\right| \left|\psi_k^T \dot{W} \psi_j\right| + \left|\frac{\lambda_k f(\lambda_k)-\lambda_j f(\lambda_j)}{\lambda_k-\lambda_j}\right| \left|\psi_k^T \dot{D} \psi_j\right| \right) \left|\psi_k(x)\right| \left|\psi_j(x)\right|
\end{aligned}
\tag{C.3}
$$

Recall that the kernel function is a monotonically increasing function, over $\mathbb{R}$, and in the studied scenario, the kernel weights are evaluated on eigenvalues of P, which are all real and limited to the interval [-1,1] (justification in Section 3.2).

For $\lambda_k = \lambda_j$, we apply the following upper bounds:

$$
\begin{aligned}
& f'(\lambda_k) = \frac{1}{\beta} f(\lambda_k)\left(1-f(\lambda_k)\right) < \frac{1}{\beta} f(\lambda_k) \\
& (zf(z))'(\lambda_k) = f(\lambda_k) + \lambda_k \frac{1}{\beta} f(\lambda_k)\left(1-f(\lambda_k)\right) < \left(1+\frac{1}{\beta}\right) f(\lambda_k)
\end{aligned}
$$

For $\lambda_k \neq \lambda_j$, and a continuous and differentiable kernel function $f(z)$, by the Mean value theorem, there exists a point c between λ_k and λ_j such that:

$$
\frac{f(\lambda_k)-f(\lambda_j)}{\lambda_k-\lambda_j} = f'(c) = \frac{1}{\beta} f(c)(1-f(c))
$$

Since $f(z)$ is monotonically increasing with z, $f(c) \leq f(\max(\lambda_k, \lambda_j))$. Therefore,

$$
\frac{f(\lambda_k)-f(\lambda_j)}{\lambda_k-\lambda_j} < \frac{1}{\beta} f(\max(\lambda_k, \lambda_j))
$$

Similarly,

$$
\frac{\lambda_k f(\lambda_k)-\lambda_j f(\lambda_j)}{\lambda_k-\lambda_j} < \left(1+\frac{1}{\beta}\right) f(\max(\lambda_k, \lambda_j))
$$

All sides of the inequalities are nonnegative $\forall \lambda_k, \lambda_j$. Therefore, the inequality is preserved when applying absolute value on all terms. Hence we obtain,

$$\left|\dot{S}(x,t)\right| \leq \sum_{k=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{\beta}f(\max(\lambda_k,\lambda_j))\left|\psi_k^T\dot{W}\psi_j\right| + \left(1+\frac{1}{\beta}\right)f(\max(\lambda_k,\lambda_j))\left|\psi_k^T\dot{D}\psi_j\right|\right)\left|\psi_k(x)\right|\left|\psi_j(x)\right| \tag{C.4}$$

where we have substituted the trivial equality $\lambda_k = \max(\lambda_k, \lambda_j)$, for $\lambda_k = \lambda_j$. Applying the trivial bound $f(\max(\lambda_k,\lambda_j)) \leq f(\lambda_k) + f(\lambda_k)$ for a non-negative kernel function and exploiting the symmetry with respect to the indices k, j, (C.4) is upper bounded by

$$\left|\dot{S}(x,t)\right| \leq 2\sum_{k=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{\beta}f(\lambda_k)\left|\psi_k^T\dot{W}\psi_j\right| + \left(1+\frac{1}{\beta}\right)f(\lambda_k)\left|\psi_k^T\dot{D}\psi_j\right|\right)\left|\psi_k(x)\right|\left|\psi_j(x)\right| \tag{C.5}$$

Next, we derive upper bounds for each of the terms in the series, individually:

$$\begin{aligned}
&\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{1}{\beta}f(\lambda_k)\left|\psi_k^T\dot{W}\psi_j\right|\left|\psi_k(x)\right|\left|\psi_j(x)\right| \\
&\qquad=\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{1}{\beta}f(\lambda_k)\left|\psi_k^T E\psi_j\right|\left|\psi_k(x)\right|\left|\psi_j(x)\right| \qquad \left(\dot{W}=E\right) \\
&\qquad\leq\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{1}{\beta}f(\lambda_k)\left(\sum_{\substack{y\in\mathcal{V}\\ z\in\mathcal{V}}}E(y,z)\left|\psi_k(y)\right|\left|\psi_j(z)\right|\right)\left|\psi_k(x)\right|\left|\psi_j(x)\right| \\
&\qquad=\frac{1}{\beta}\sum_{\substack{y\in\mathcal{V}\\ z\in\mathcal{V}}}E(y,z)\left(\sum_{k=1}^{n}\sum_{j=1}^{n}f(\lambda_k)\left|\psi_k(y)\right|\left|\psi_k(x)\right|\left|\psi_j(z)\right|\left|\psi_j(x)\right|\right)
\end{aligned} \tag{C.6}$$

Applying Cauchy-Schwartz inequality,

$$
\begin{aligned}
\sum_{k=1}^{n}\sum_{j=1}^{n} f(\lambda_k)\,|\psi_k(y)|\,|\psi_k(x)|\,\left|\psi_j(z)\right|\,\left|\psi_j(x)\right| &= \left(\sum_{k=1}^{n} f(\lambda_k)\,|\psi_k(y)|\,|\psi_k(x)|\right)\left(\sum_{j=1}^{n}\left|\psi_j(z)\right|\,\left|\psi_j(x)\right|\right)\\
&\leq \sqrt{\sum_{k=1}^{n} f(\lambda_k)\,|\psi_k(y)|^2}\sqrt{\sum_{l=1}^{n} f(\lambda_l)\,|\psi_l(x)|^2}\sqrt{\sum_{j=1}^{n}\left|\psi_j(z)\right|^2}\sqrt{\sum_{m=1}^{n}|\psi_m(x)|^2}\\
&= \sqrt{\underbrace{\sum_{k=1}^{n} f(\lambda_k)\,|\psi_k(y)|^2}_{S(y,t)}}\sqrt{\underbrace{\sum_{l=1}^{n} f(\lambda_l)\,|\psi_l(x)|^2}_{S(x,t)}}\sqrt{\underbrace{\sum_{j=1}^{n}\left|\psi_j(z)\right|^2}_{\frac{1}{d(z)}}}\sqrt{\underbrace{\sum_{m=1}^{n}|\psi_m(x)|^2}_{\frac{1}{d(x)}}}\\
&\leq \sqrt{S(y,t)}\sqrt{S(x,t)}\sqrt{\frac{1}{\underline{d_0}}}\sqrt{\frac{1}{\underline{d_0}}} \qquad \left(d(z),d(x)\geq \underline{d_0}\right)\\
&\leq \frac{1}{\underline{d_0}}\bar{S}(t)
\end{aligned}
\tag{C.7}
$$

Substituting in (C.6),

$$
\begin{aligned}
\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{1}{\beta}f(\lambda_k)\left|\psi_k^T\dot{W}\psi_j\right|\,|\psi_k(x)|\,\left|\psi_j(x)\right| &\leq \frac{1}{\beta}\sum_{\substack{y\in\mathcal{V}\\ z\in\mathcal{V}}} E(y,z)\left(\frac{1}{\underline{d_0}}\bar{S}(t)\right)\\
&\leq \frac{2C}{\underline{d_0}}\frac{1}{\beta}\bar{S}(t)
\end{aligned}
\tag{C.8}
$$

In a similar fashion, one can show that the remaining term in (C.5) is upper bounded by

$$
\begin{aligned}
&\sum_{k=1}^{n}\sum_{j=1}^{n}\left(1+\frac{1}{\beta}\right)f(\lambda_k)\left|\psi_k^T\dot{D}\psi_j\right|\,|\psi_k(x)|\,\left|\psi_j(x)\right|\\
&\qquad\leq\left(1+\frac{1}{\beta}\right)\sum_{y\in\mathcal{V}}\left|\dot{D}(y)\right|\left(\sum_{k=1}^{n}\sum_{j=1}^{n} f(\lambda_k)\,|\psi_k(y)|\,|\psi_k(x)|\,\left|\psi_j(y)\right|\,\left|\psi_j(x)\right|\right)
\end{aligned}
\tag{C.9}
$$

Applying (C.7),

$$\sum_{k=1}^{n}\sum_{j=1}^{n}\left(1+\frac{1}{\beta}\right)f(\lambda_k)\left|\psi_k^T\dot{D}\psi_j\right|\left|\psi_k(x)\right|\left|\psi_j(x)\right| \leq \left(1+\frac{1}{\beta}\right)\frac{1}{\underline{d_0}}\bar{S}(t)\sum_{y\in\mathcal{V}}\left|\dot{D}(y)\right|$$
$$=\left(1+\frac{1}{\beta}\right)\frac{1}{\underline{d_0}}\bar{S}(t)\sum_{\substack{y\in\mathcal{V}\\ z\in\mathcal{V}}}E(z,y) \tag{C.10}$$
$$\leq\frac{2C}{\underline{d_0}}\left(1+\frac{1}{\beta}\right)\bar{S}(t)$$

Adding both terms, (C.8) and (C.10), together we have,

$$\left|\dot{S}(x,t)\right| \leq \frac{4C}{\underline{d_0}}\left(1+\frac{2}{\beta}\right)\bar{S}(t) \tag{C.11}$$

Substituting in (C.2), $\forall x\in\mathcal{V}$:

$$|S(x,t)-S(x,0)| \leq \tilde{C}\int_0^t \bar{S}(\tau)d\tau, \qquad \tilde{C} := \frac{4C}{\underline{d_0}}\left(1+\frac{2}{\beta}\right) \tag{C.12}$$

Next, we upper bound $\bar{S}(t)$.

Suppose that $\bar{S}(t) = S(x_0,t)$ for some x_0. Then

$$\bar{S}(t)-\bar{S}(0) = S(x_0,t)-\bar{S}(0) \leq S(x_0,t)-S(x_0,0) \leq \sup_{x\in\mathcal{V}}\left(S(x,t)-S(x,0)\right)$$

Since (C.12) holds for all $x\in\mathcal{V}$:

$$\bar{S}(t)-\bar{S}(0) \leq \tilde{C}\int_0^t \bar{S}(\tau)d\tau$$

By the integral form of Grönwall's inequality, $\forall t\in[0,1]$

$$\bar{S}(t) \leq \bar{S}(0)e^{\tilde{C}t}$$

Combining the result with (C.12):

$$|S(x,t) - S(x,0)| \leq \tilde{C} \int_0^t \bar{S}(0) e^{\tilde{C}\tau} d\tau = \bar{S}(0) \left(e^{\tilde{C}t} - 1 \right)$$

By the proposition of initial separation of $S(x;\alpha,\beta)$, at $t = 0$ the separation between $\mathcal{C}$ and $\mathcal{B}$ is at least g_0 (4.6). So the threshold exists as long as

$$g_0 \geq 2\bar{S}(0) \left(e^{\tilde{C}t} - 1 \right)$$

Which is satisfied by the condition that

$$\underline{\frac{C}{d_0}} \leq \frac{1}{4} \frac{1}{1 + \frac{2}{\beta}} \ln \left(1 + \frac{1}{2} \frac{\blacklozenge}{\left(1 + \varepsilon_0 \frac{\delta n}{K} + 2\varepsilon_1 \right)} \right) \tag{C.13}$$

□

The material in this thesis is currently being prepared for submission for publication of the material. Dror, Shahar; Mishne, Gal; Cheng, Xiuyuan. The thesis author was the primary investigator and author of this material.

Bibliography

[ASC11] Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 1626–1633. IEEE, 2011.

[Bea15] Mario Beauchemin. A density-based similarity matrix construction for spectral clustering. *Neurocomputing*, 151:835–844, 2015.

[BFSS19] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019.

[BLF$^+$18] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

[BM17] Yochai Blau and Tomer Michaeli. Non-redundant spectral dimensionality reduction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 256–271. Springer, 2017.

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[CL06] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[CM20] Xiuyuan Cheng and Gal Mishne. Spectral embedding norm: Looking deep into the spectrum of the graph laplacian. *SIAM Journal on Imaging Sciences*, 13(2):1015–1048, 2020.

[DG03] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[DTCK18] Carmeline J Dsilva, Ronen Talmon, Ronald R Coifman, and Ioannis G Kevrekidis. Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study. *Applied and Computational Harmonic Analysis*, 44(3):759–773, 2018.

[DZ10] Bo Du and Liangpei Zhang. Random-selection-based anomaly detector for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 49(5):1578–1589, 2010.

[EDMD19] Thibaud Ehret, Axel Davy, Jean-Michel Morel, and Mauricio Delbracio. Image anomalies: A review and synthesis of detection methods. *Journal of Mathematical Imaging and Vision*, 61(5):710–743, 2019.

[For10] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[HAVL05] Matthias Hein, Jean-Yves Audibert, and Ulrike Von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph laplacians. In *International Conference on Computational Learning Theory*, pages 470–485. Springer, 2005.

[Lan50] Cornelius Lanczos. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA, 1950.

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[LZZM15] Jiayi Li, Hongyan Zhang, Liangpei Zhang, and Li Ma. Hyperspectral anomaly detection by the use of background joint sparse representation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2523–2533, 2015.

[MC12] Gal Mishne and Israel Cohen. Multiscale anomaly detection using diffusion maps. *IEEE Journal of selected topics in signal processing*, 7(1):111–123, 2012.

[MC17] Gal Mishne and Israel Cohen. Iterative diffusion-based anomaly detection. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1682–1686. IEEE, 2017.

[NG07] Boaz Nadler and Meirav Galun. Fundamental limitations of spectral clustering. In *Advances in neural information processing systems*, pages 1017–1024, 2007.

[NJW01] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 849–856, 2001.

[RS00] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[Sin06] Amit Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

[SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[SOG09] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

[TDSL00] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[TH99] D-M Tsai and C-Y Hsieh. Automated surface inspection for directional textures. *Image and Vision computing*, 18(1):49–62, 1999.

[TH03] Du-Ming Tsai and Tse-Yun Huang. Automated surface inspection for statistical textures. *Image and Vision computing*, 21(4):307–323, 2003.

[VL07] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[VLBB08] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

[ZC10] Maria Zontak and Israel Cohen. Defect detection in patterned wafers using anisotropic kernels. *Machine Vision and Applications*, 21(2):129–141, 2010.

[ZLY11] Xianchao Zhang, Jingwei Li, and Hong Yu. Local density adaptive similarity measurement for spectral clustering. *Pattern Recognition Letters*, 32(2):352–358, 2011.

[ZMP04] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. *Advances in neural information processing systems*, 17:1601–1608, 2004.