Spectral Clustering as Mapping to a Simplex

Peeyush Kumar[*]
Indian Institute of Technology
Madras
Chennai, India

Niveditha L[†]
Department of Applied
Mathematics and
Computational Sciences
PSG College Of Technology
Coimbatore, India

Balaraman Ravindran[‡]
Department of Computer
Science and Engineering
Indian Institute of Technology
Madras
Chennai, India

ABSTRACT

Spectral methods have been widely used to study the structural properties of unlabeled datasets. In this work we describe a clustering approach that exploits the structural properties in the configuration space of objects as well as the spectral sub-space, quite unlike earlier methods. We propose a spectral clustering approach, where we formalize the notion of *clusters* as vertices of a simplex in the spectral subspace. We define clustering as *memberships* of data points to vertices of this simplex. We empirically demonstrate that our method is comparable to the state-of-the-art methods in a variety of domains and outperforms other generic clustering algorithms.

1. INTRODUCTION

In recent years, spectral clustering has become one of the most popular approaches for clustering data. There have been many successful applications of spectral clustering methods on the real world data (e.g., [15, 21, 22, 14, 2], etc.). These methods outperform conventional clustering techniques, yet are simple to implement and can be solved by standard linear algebra tools. Moreover, spectral clustering can be implemented efficiently even for large data sets, as long as the *similarity graph* is sparse, which is usually true for real world applications. Central to the idea of spectral clustering is the *graph Laplacian* which is obtained from the similarity graph ([10]). There are many tight connections between the topological properties of graphs and the graph Laplacian matrices, which spectral clustering methods exploit to partition the data into clusters.

Spectral clustering was made popular by the works of [15] (Normalized Cut Algorithm), [12], [7], etc. Although these methods are known to have many successful applications, they typically work on a case-by-case basis. Though the spectra of the Laplacian preserves the structural properties of the graph, the methods used thereafter to cluster the data in the eigenspace of the Laplacian do not guarantee this. For example [13, 15] uses k-means clustering in the eigenspace of the Laplacian, which will only work if the clusters lie in disjoint convex sets of the underlying eigenspace. While [12] uses projections onto the largest $k - eigenvectors$ to partition the data into clusters, which does not preserve the topological properties of the data lying in the eigenspace of the Laplacian. This is because projection methods do not incorporate geometric constraints due to the underlying structure in the eigenspace: points closer in Eucledian distance may be far apart on the manifold.

In this work we describe a clustering approach that exploits the structural properties in the configuration space of objects as well as the spectral sub-space, quite unlike earlier methods. We construct the Laplacian from the adjacency information. The spectra of this Laplacian is constructed, which encodes the structural properties of the underlying graph. We then find the best transformation of the spectra, such that the transformed basis aligns itself with the clusters of data points in the eigen-space. Then we use a projection method described in Section 4 to find the membership of each of the datapoints to a set of of special points lying on the transformed basis, which we identify as vertices of a simplex, as decribed in Section 3.

An important point to note here is that while the secondary clustering might look similar to the one in N-Cut, it is in fact quite different from the N-Cut algorithm. N-cut performs k-means clustering after projecting the data onto the top-k eigenvectors, while we assign memberships to the identified vertices in the eigenspace. These vertices need not coincide with the centroid of the data points.

Analytically [20] shows that the convex hull of the data points form a simplex in the Laplacian eigenspace of a suitably constructed similarity graph. In disjoint datasets where the similarity matrix is block diagonal, the data points lie on the vertices of this simplex. In datasets which are not completely disjoint but which show some *group* structure one may consider the data points as perturbations near this simplex structure formed by perfect disjoint sets in the Laplacian eigen-subspace.

For first order perturbation, we show that the simplex is linearly transformed and all the data points cluster exactly *at* the vertices of this transformed simplex. For higher order perturbations, which is the most general case, the process of clustering is defined as assigning degrees of association of the data points to each of these vertices. Consequently, a

[*] peeyush@stuckinimagination.com
[†] narasimhan.niveditha@gmail.com
[‡] ravi@cse.iitm.ac.in

good clustering is one that minimizes the deviation from the simplex structure. Clustering is therefore defined as the degree of association of the data-points with each of these *clusters.* Unlike the other methods cited above, our method does not make any assumptions on the structure of the underlying spectral-space, and hence is generalizable across multiple domains. This is indeed what we observe in Section 5.

We take inspiration from the conformal dynamics literature, where [20] does a similar analysis to detect conformal states of a dynamical system. They propose a spectral clustering algorithm PCCA+, which is based on the the principles of Perron Cluster Analysis of the transition structure of the system. We extend their analysis to generic structural similarity data to yield a state-of-the-art spectral clustering algorithm. Using this framework gives us various advantages:

- *A formal notion of clusters as vertices of a simplex in the eigen-subspace of the Laplacian.* The clustering is performed by minimizing deviations from a simplex structure and hence does not require any arbitrary regularization term.
- *Characteristic functions that describe the degree of membership to a given abstract cluster.* We can interpret the membership functions as the *likelihood* of an object belonging to a particular cluster (see [16]). The algorithm could also generate crisp partitioning of the objects into groups, as and when required.
- *Connectivity information between the clusters* which might be required in certain domains. For example one might be interested to know the connectivity information between two documents in a text dataset.

The rest of the paper is organized as follows. Section 2 introduces Spectral Methods. Section 3 describes the notion of clustering as mapping to a simplex structure. Section 4 introduces the PCCA+ algorithm. Finally we empirically demonstrate in Section 5 that our method is comparable to the state-of-the-art methods in a variety of domains and outperforms any other generic clustering algorithm.

2. BACKGROUND

Given a set of data points $x_1, \ldots, x_n$ and some notion of similarity $s_{ij} > 0$ between all pairs of data points x_i and x_j, the intuitive goal of clustering is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar to each other. For this purpose we represent the data in form of the similarity graph $G = (V, E)$. Each vertex v_i in this graph represents a data point x_i. Two vertices are connected if the similarity s_{ij} between the corresponding data points x_i and x_j is positive or larger than a certain threshold, and the edge is weighted by s_{ij}. We build a similarity matrix for the similarity graph as the weighted adjacency matrix:

$$W_{ij} = \begin{cases} s_{ij} & s_{ij} \geq t_c \\ 0 & otherwise \end{cases} \tag{1}$$

It is not trivial to construct a suitable similarity matrix and this is usually domain specific.

The degree of the vertex $v_i \in V$ is defined as $d_i = \sum_j w_{ij}$. We define the Laplacian as

$$\mathcal{L} = D^{-1}W \tag{2}$$

where D is the matrix with d_i on its diagonal and 0 elsewhere. The spectrum of this Laplacian, namely the eigenvalues and eigenvectors encode the structural properties of the graph.

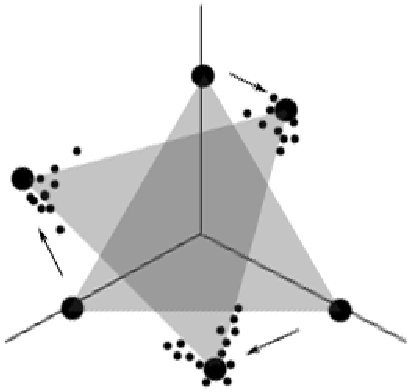


Figure 1: Simplex First order and Higher order Perturbation

3. CLUSTERING AS MAPPING TO THE SIMPLEX STRUCTURE

In this section we describes the notion of clustering as mapping to a simplex structure and provide some derivations and interpretations to better understand the process of clustering. Lemma 1 describes the structure of the simplex. We derive an explicit formula (Lemma 2) for the best transformation operator, such that the transformed simplex (whose vertices lie on the transformed basis) aligns itself with the data points. We then define the notion of clustering as membership to these vertices (Def 1), which are identified as clusters. While [4, 8] derives the results in Lemma 2 using different methods, we give a new approach for interpreting them. We also derive bounds on the deviation of actual data-points from this simplex structure in Lemma 3. We define another quantity, the macro transition matrix, in Def 2 which encodes the connectivity information across clusters, which could be useful to know in various domains.

For a graph with disjoint components i.e., one whose similarity matrix can be reduced to a block diagonal form, the Lapalcian matrix $\tilde{\mathcal{L}}$ has a block structure, where each block is a matrix which corresponds to the Laplacian matrix for the disjoint set of vertices. Each vertex $v_i \in V$ can be mapped to the i^{th} row of the eigenvector matrix $\tilde{Y}_{nk}$; where n is the number of vertices and k is number of eigenvectors, corresponding to the *eigenvalues* $= 1$. As it turns out, the eigenvectors for the Laplacian matrix can be interpreted as an indicator of membership for each object to a suitable disjoint set.

LEMMA 1. *Eigenvectors of the Laplacian $\mathcal{L}$ with a block diagonal structure form the vertices of a simplex $\mathbb{R}^{k-1}$ ([20]).*

PROOF. Each block matrix has its own Laplacian $\hat{\mathcal{L}}$, since the rows of the Laplacian sum to 1, hence a vector with all identical elements is an eigenvector of this system. Transforming this to the full Laplacian matrix, components of each of the eigenvectors $[Y_1, Y_2, \ldots, Y_k]$ of $\mathcal{L}$ are pairwise

Algorithm 1 PCCA+

```
Construct $\mathcal{L}$ from the Similarity matrix $S$
Compute first $n$ (in the descending order) eigenvalues $(d)$ for $\mathcal{L}$
Choose first $k$ eigenvalues for which $\frac{(e_{k+1}-e_k)}{1-e_k} > t_c(Spectral\ Gap\ Threshold)$. Compute the eigenvectors for corresponding eigenvalues $(e_1, e_2, \ldots, e_k)$ and stack them as column vectors in matrix $Y$
Lets denote the rows of $Y$ as $Y(1), Y(2), \ldots Y(N) \in \mathbb{R}^k$.
Define $\pi(1)$ as that index, for which $\|Y(\pi(1))\|_2$ is maximal. Define $\gamma_1 = span\{Y(\pi_1)\}$.
For $i = 2, \ldots, k$: Define $\pi_i$ as that index, for which the distance to the hyperplane $\gamma_{i-1}$, i.e. $\|Y(\pi_i) - \gamma_{i-1}\|_2$, is maximal. Define $\gamma_i = span\{Y(\pi_1), ..., Y(\pi_i)\}$. $\|Y(\pi_i) - \gamma_{i-1}\|_2 = \left\|Y(\pi_i) - \gamma_{i-1}^T((\gamma_{i-1}\gamma_{i-1}^T)^{-1}\gamma_{i-1})Y(\pi_i)^T)\right\|$
```

identical for indices corresponding to the same block. Regarding the rows of Y in $\mathbb{R}^k$ as k distinct points in $\mathbb{R}^k$, they form the vertices of a simplex because by definition the convex hull of k distinct points in $\mathbb{R}^k$ form a simplex $\hat{\sigma}^{k-1}$. □

We call these vertices in the eigenspace $\mathbb{R}^k$ as (first-order) Clusters $\mathbb{C}_k$. Laplacian in systems which exhibit connections across *groups* can be approximated as perturbations on disjoint sets.

$$\tilde{\mathcal{L}} = \mathcal{L} + \epsilon\mathcal{L}^{(1)} + \epsilon^2\mathcal{L}^{(2)} + \ldots$$

where $\mathcal{L}^{(1)}, \mathcal{L}^{(2)}, \ldots$ are respectively the first order and higher order Laplacian perturbation terms, and ϵ is the order artifact. With the ϵ perturbation analysis of this equation the perturbation on the eigenvectors and eigenvalues can similarly be written as

$$\tilde{Y} = Y + \epsilon Y^{(1)} + \mathcal{O}(\epsilon^2)$$

$$\tilde{\Lambda} = \Lambda - \epsilon\Lambda^{(1)} - \mathcal{O}(\epsilon^2)$$

Lemma 2. *First order perturbation term $Y^{(1)} = YB$, $B \in \mathbb{R}^{k\times k}$ is a linear mapping: $\mathbb{R}^k \mapsto \mathbb{R}^k$.*

Proof. consider the i^{th} eigenvector

$$\tilde{\mathcal{L}}\tilde{y}_i = \tilde{\lambda}_i\tilde{y}_i$$

writing it in terms of the perturbation expansion and matching the same order terms (for the first order perturbation) we get

$$\mathcal{L}y_i^{(1)} + \mathcal{L}^{(1)}y_i = \lambda_i y_i^{(1)} - \lambda_i^{(1)}y_i$$

therefore,

$$(\mathcal{L}^{(1)} + \lambda_i^{(1)}\mathcal{I})y_i + (\mathcal{L} - \lambda_i\mathcal{I})y_i^{(1)} = 0$$

Taking a dot product of this equation with another eigenvector y_j, we get

$$\left\langle(\mathcal{L}^{(1)} + \lambda_i^{(1)}\mathcal{I})y_i, y_j\right\rangle + \left\langle(\mathcal{L} - \lambda_i\mathcal{I})y_i^{(1)}, y_j\right\rangle = 0$$

$$\left\langle(\mathcal{L}^{(1)} + \lambda_i^{(1)}\mathcal{I})y_i, y_j\right\rangle + \left\langle y_i^{(1)}, (\mathcal{L} - \lambda_i\mathcal{I})y_j\right\rangle = 0$$

The second term goes to zero because $(\mathcal{L} - \lambda_i\mathcal{I})y_j = 0$, hence we have the first term zero as well. This implies that $(\mathcal{L}^{(1)} + \lambda_i^{(1)}\mathcal{I})$ is linear transformation which takes a vector y_i and either transforms it perpendicular to itself, or to itself, because $\langle y_i, y_j\rangle = 0$. Also

$$y_i^{(1)} = (\lambda_i\mathcal{I} - \mathcal{L})^{-1}(\mathcal{L}^{(1)} + \lambda_i^{(1)}\mathcal{I})y_i$$

Hence we get $y_i^{(1)} = By_i$ □

This implies that the perturbation of the simplex structure can at most be of the order $O(\epsilon^2)$ (see Figure 1). In other words, the simplex structure is preserved for first order perturbations, while for the higher order perturbations the simplex structure perturbs. Hence we have here a formal definition of clusters, in the abstract notion, as vertices $\mathbb{C}_k$ of this simplex structure.

Def 1. *A vertex v_i is said to belong to the cluster $\mathbb{C}_k$ with the perfect membership if $Y(i,:) = \mathbb{C}_k$*

In soft clustering a continuous indicator for membership $\tilde{\chi}_i : V \mapsto [0,1]$, which assigns a grade of membership between 0 and 1 to each vertex $v_i \in V$; $\forall i$. Therefore, a vertex may correspond to different clusters with a different grade of membership. For each vertex $v \in V$ the sum of the grades of membership with regard to the different clusters is 1, i.e.

$$\sum_i^k \tilde{\chi}_i(v) = 1$$

Each vertex is represented by a vector $(\chi_1(v), \ldots, \chi_k(v)) \in \mathbb{R}^k$. Since these vectors are positive and the partition of unity holds, they lie in the standard σ_{k-1} simplex spanned by the k unit vectors of $\mathbb{R}^k$. Therefore, clustering can be seen as a simple linear mapping from the rows of Y to the rows of a membership matrix $\tilde{\chi}$. The linear mapping is expressed by a regular $k \times k$ matrix $\mathcal{A}$:

$$\tilde{\chi} = \tilde{Y}\mathcal{A}$$

This matrix maps the vertices of the simplex contained in the rows of Y onto the vertices of the simplex σ_{k-1}. Therefore, if one finds the indices $\pi_1, \ldots, \pi_n \in [1, N]$ of the vertices in Y one can construct the linear mapping as follows

$$\mathcal{A}^{-1} = \begin{pmatrix} \tilde{Y}_{\pi_1,1} & \cdots & \tilde{Y}_{\pi_1,k} \\ \vdots & \vdots & \vdots \\ \tilde{Y}_{\pi_k,1} & \cdots & \tilde{Y}_{\pi_k,k} \end{pmatrix}$$

[19] shows that solution for $\mathcal{A}$ exists, if and only if the convex hull of the rows of $\tilde{Y}$ is a simplex. From perturbation analysis we know that this is the case with a deviation of order $O(\epsilon^2)$. To partition the data, as and when required, we assign each state to a partition numbered $P(s) = \arg\max_{k=1}^n \tilde{\chi}_k(s)$. □

To estimate the number of clusters k, the spectral gap is used as an indicator of deviation from the simplex structure. Spikes in the eigenvalues indicate the presence of a group structure in the graph.

LEMMA 3. *The perturbation coefficient is bounded by*

$$\epsilon < \frac{1 - \tilde{\lambda}_{min}}{\left|\lambda^{(1)}_{max}\right|}$$

where $\tilde{\lambda}_{min}$ is the smallest eigenvalue of $\tilde{\mathcal{L}}$ and $\left|\lambda^{(1)}_{max}\right| \neq 0$ (this condition just means there are perturbations around the regular zeroth order simplex structure, if there are no perturbations then the Lemma trivially hold as $\epsilon = 0$).

PROOF. Consider

$$\begin{aligned}\tilde{\mathcal{L}} - \tilde{\lambda_i} I &= \mathcal{L} - \lambda_i I - (\tilde{\lambda_i} - \lambda_i) + (\tilde{\mathcal{L}} - \mathcal{L}) \\ &= \left[(1 - (\tilde{\lambda_i} - \lambda_i - (\tilde{\mathcal{L}} - \mathcal{L}))(\tilde{\mathcal{L}} - \lambda_i I)^{-1}\right] (\mathcal{L} - \lambda_i I)\end{aligned}$$

Taking inverse on both sides

$$(\tilde{\mathcal{L}} - \tilde{\lambda_i} I)^{-1} = \\ \left[1 - (\tilde{\lambda_i} - \lambda_i - (\tilde{\mathcal{L}} - \mathcal{L}))(\tilde{\mathcal{L}} - \lambda_i I)^{-1}\right]^{-1} (\mathcal{L} - \lambda_i I)^{-1}$$

The terms inside $[\cdot]^{-1}$ can be defined by a convergent Neumann series. Which in this case we get [8]:

$$\begin{aligned}\left|\tilde{\lambda_i} - \lambda_i\right| + \left\|\tilde{\mathcal{L}} - \mathcal{L}\right\| &< \left\|(\tilde{\mathcal{L}} - \lambda_i I)^{-1}\right\|^{-1} \\ &\leq \left\|((\tilde{\mathcal{L}} - \lambda_i I)^{-1})^{-1}\right\| \\ &= \left\|(\tilde{\mathcal{L}} - \lambda_i I)\right\|\end{aligned}$$

where $\|\cdot\|$ is the spectral norm induced by L2-norm, which is equal to the square root of the largest eigenvalue of the matrix $(\cdot)(\cdot)^T$.
Since

$$\begin{aligned}\left|\tilde{\lambda_i} - \lambda_i\right| &= \left|\epsilon\lambda^{(1)}_i + \epsilon^2\lambda^{(2)}_i + \dots\right| \\ &= \epsilon\left|\lambda^{(1)}_i + \epsilon\lambda^{(2)}_i + \dots\right|\end{aligned}$$

Therefore

$$\epsilon\left|\lambda^{(1)}_i + \epsilon\lambda^{(2)}_i + \dots\right| < \left\|(\tilde{\mathcal{L}} - \lambda_i I)\right\| - \left\|\tilde{\mathcal{L}} - \mathcal{L}\right\|$$

For a given i, such that $\lambda_i = 1$ implies that $\tilde{\lambda_i} > 0$. Hence for all such i each $\lambda^{(n)}_i > 0$. Therefore for such an i

$$\begin{aligned}\epsilon\left|\lambda^{(1)}_i\right| &< \left\|(\tilde{\mathcal{L}} - \lambda_i I)\right\| - \left\|\tilde{\mathcal{L}} - \mathcal{L}\right\| \\ &< \left\|(\tilde{\mathcal{L}} - \lambda_i I)\right\| \\ &\leq 1 - \tilde{\lambda}_{min} \\ \epsilon &< \frac{1 - \tilde{\lambda}_{min}}{\left|\lambda^{(1)}_{max}\right|}\end{aligned}$$

Hence the deviation from the simplex structure is bounded above as shown. □

The connectivity information between the various clusters can also be recovered from the membership function.

DEF 2. *The connectivity information across different clusters, is given by*

$$\mathcal{L}^{macro} = \tilde{\chi}^T \tilde{\mathcal{L}} \tilde{\chi}$$

where $\mathcal{L}^{macro}$ is the Laplacian in a matrix space.

In this representation each cluster is represented by a single node, and connectivity information across the clusters is given $\mathcal{L}^{macro}(i, j)$ for $i \neq j$, while the relative connectivity information within a cluster is given by $\mathcal{L}^{macro}(i, i)$ for all i. The connectivity information might be required in certain domains. For example one might be interested to know the connectivity information between two documents in a text dataset (see 5.2).

4. PCCA+: THE ALGORITHM

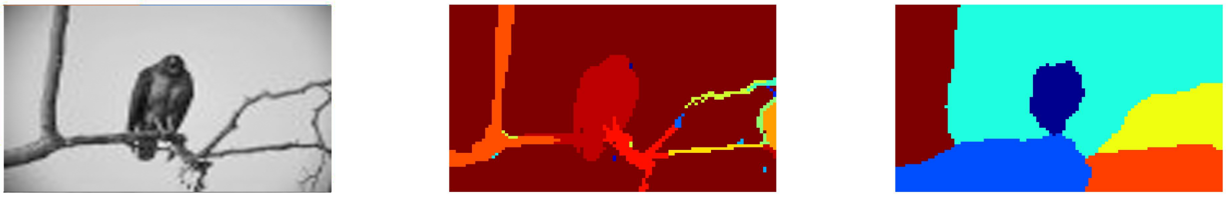
We propose the use of PCCA+ (see Algorithm 1) as a clustering algorithm on unlabeled datasets. Datasets are represented as similarity graph by defining a pairwise similarity function between pairs of objects (pixels, states, vertices, etc.) in the dataset. Using these similarity functions for all the data points, we construct a similarity matrix S. Using this similarity matrix, we define the Laplacian and construct its spectra through eigenanalysis . The spectral gap method is used to estimate the number of clusters k (line 3 in Algorithm 1). This is used to find the simplex in the $\mathbb{R}^k$ eigen-subspace.

Note that for the first order perturbation the simplex is just a linear transformation around the origin, hence in order to find the vertices of the simplex σ^{k-1}, we need to find the k points which could form the convex hull such that the deviation of all the points from this convex hull is minimized. Hence, we start by finding the datapoint which is farthest located from the origin (line 5 in Algorithm 1), say $\tilde{Y}_{\pi_1}$. Then we proceed by finding the data point which is farthest located from the first point, say $\tilde{Y}_{\pi_2}$. We iterate this procedure of finding the datapoints which is located farthest from the consecutive hyperplane constructed by joining the previous data points, until we find k datapoints (see line 6 in Algorithm 1). These datapoints form the vertices of the simplex. While this approach is superficially similar to [18] in fact this is very different since they operate in the data space and we operate in the Eigen subspace.

For example, Figure 1, shows a 3 dimensional eigen-subspace with a σ^2 simplex. The data points are shown as small black dots. Had the system been a completely disjoint system with 3 disjoint sets, the simplex would be aligned with the axes, with the clusters corresponding to the vertices of the simplex (the unit vectors). Since the system is represented as perturbations around this disjoint system, the first order perturbation linearly transforms the simplex, as shown. Because the data consists of higher order perturbations, the datapoints do not exactly map to the vertices of the new simplex, though their deviation is minimized from this simplex. The clusters for this system are the vertices of the new simplex. Thus we have clustering as membership of datapoints to the vertices of the new simplex.

For a graph with pronounced *group* structure, the datapoints will tend to clutter near the vertices of the transformed simplex, while for a graph with high connectivity the datapoints will spread out over the simplex plane. Hence this framework also contains an intrinsic mechanism to return the information about goodness of clustering, which is the distribution of the membership functions for a datapoint across various clusters. Sharp peaks in the value indicates a good clustering while a more uniform value indicates a bad *group* structure and hence a bad clustering.

Hence, unlike the other methods we have various advantages in clustering. *a*) We exploit the local structural prop-



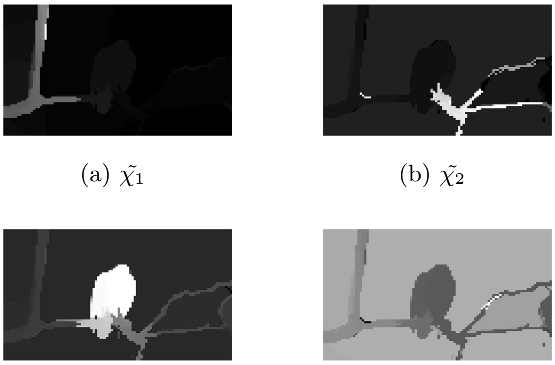
(a) Image

(b) PCCA+

(c) N-Cut

Figure 2: Partitioning into multiple segments. Comparision with N-Cut

erties of the underlying data space by using pairwise similarity functions, while we use spectral methods to encode the global structural properties. *b*) The clustering procedure does not assume anything about the underlying structure and the mapping to a simplex is inherently built in the properties of the Laplacian. *c*) We define a formal notion of clustering as the membership of datapoints to the vertices of the simplex. *d*) We also estimate the connectivity information across clusters. *e*) As shown in Section 3 the perturbations could at most be $\mathcal{O}(\epsilon^2)$, hence the maximum deviation of any datapoint can at most be $\mathcal{O}(\epsilon^2)$. This implies that the clustering procedure is robust to outliers, as in the spectral subspace, they can at most have a deviation of $\mathcal{O}(\epsilon^2)$. An important point to note here is that the standard spectral clustering approach (refer [12]) which projects the data onto the top eigenvectors could be thought of as the zeroth order version of PCCA+ or for a special case when $B = I$ (where B is the transformation operator defined in Lemma 2).



(a) $\tilde{\chi_1}$

(b) $\tilde{\chi_2}$

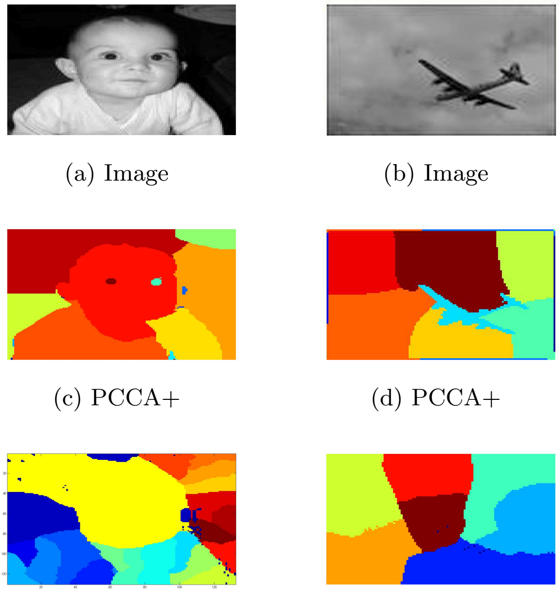
(c) $\tilde{\chi_3}$

(d) $\tilde{\chi_4}$

Figure 3: Membership Functions for some clusters

5. EXPERIMENTS

Other spectral clustering methods had varying amounts of success on different domain. Although few of them have been successfully used across multiple domains. The primary reason for this is because none of them have a principled way of exploiting structural properties encoded in the laplcian. We demonstrate the utility of PCCA+ across multiple domains, while comparing it with other state of the art methods in that domain. We also compare PCCA+ with N-Cut method ([15]) across all these domains.



(a) Image

(b) Image

(c) PCCA+

(d) PCCA+

(e) N-Cut

(f) N-Cut

Figure 4: Segmentation for some other images

5.1 Image Segmentation

Figure 2a shows an image we would like to segment. The procedure for clustering is as follows

1. Construct a similarity graph $G = (V, E)$ by taking each pixel as a node and connecting each pair of pixels by an edge. The similarity value should reflect the likelihood of 2 pixels belonging to the same group. We define the similarity matrix in terms of the radial basis function of the brightness of the pixels adjacent to each other as follows

$$W_{ij} = \begin{cases} \exp - \frac{\|F_i - F_j\|^2}{\sigma_I^2} & if\ \|X_i - X_j\|^2 \leq 1 \\ 0 & otherwise \end{cases} \tag{3}$$

where X_i and X_j are the spatial location of the pixels, F_i and F_j are the intensity values for brightness of the pixels. This similarity matrix gives a non zero value for

Table 1: Normalized Cut Values for PCCA+ and N-Cut. (Lower is good)

Image	PCCA+	Ncut
Bird	2.1616e-005	0.2099
Baby	4.7245e-004	0.0126
Aircraft	0.0029	0.0128

the pixel i connected to a pixel j which is located on any of the 8 sites on a square lattice around the pixel i. For a colored RGB image, the corresponding grayscale image is used here for image segmentation. Please note that this construction of the similarity matrix is quite different from the construction of the similarity matrix by [15], we have only 1 parameter to adjust which is the width of the intensity gaussian mixture as opposed to 6 parameters in [15].

2. We find the number of clusters using spectral gap method by finding the top k eigenvalues for which $\frac{(e_{k+1}-e_k)}{1-e_k} > t_c$, where t_c is the spectral gap threshold.

3. We apply PCCA+ algorithm (Algorithm 1) to obtain the membership matrix for the graph. We show in Figure 3 the plot of membership functions for some of the clusters identified. We show in Figure 2 the partitioning of the image into discrete segments, where each segment is differently color coded. We also show in the same Figure 2, the results obtained using N-Cut (by performing k-means in the eigen-space for secondary clustering) by carefully choosing the best parameters[1]. Note that PCCA+ provides clusters which could separate the background from the objects which is structurally a very complex group, while N-Cut segmented the background into different groups.

We also compare the Normalized Cut Values of PCCA+ and N-Cut in Table 1. We observe that PCCA+ gives very good average Normalized Cut values ~ 0.

5.2 Text Clustering

Document clustering is one of the most crucial techniques to organize the documents in an unsupervised manner. Many clustering methods have been applied to clustering documents into categories, such as k-means [11], naive Bayes or Gaussian mixture model [1, 9], single-link [6], and DBSCAN [5]. From different perspectives, these clustering methods can be classified into agglomerative or divisive, hard or fuzzy, deterministic or stochastic. The typical data clustering tasks are directly performed in the data space. However, the document space is always of very high dimensionality. Due to the consideration of the curse of dimensionality, it is desirable to first project the documents into a lower-dimensional subspace in which the semantic structure of the document space becomes clear. Literature on spectral clustering shows its capability to handle highly nonlinear data. Also, its strong connections to differential geometry make it capable of discovering the manifold structure of the document space.

For the experiments, three standard document collections were used in our experiments: Reuters-21578, Newsgroups20 and TDT2. Reuters-21578 corpus[2] contains 21,578 documents in 135 categories. The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups[3]. The TDT2 corpus[4] consists of data collected during the first half of 1998 and taken from six sources. It consists of 11,201 on-topic documents which are classified into 96 semantic categories. From the original corpus, the documents appearing in multiple categories are removed. The pruned TDT2 dataset contains 9394 documents containing top 30 categories, Reuters-21578 contains 8293 documents belonging to 65 categories and 20Newsgroup contains 18846 documents that belongs to 20 groups. All the dataset are Document-term matrix where each row represents a document. We perform TF-IDF. This gives the weighted document-term matrix. Then we calculate the distance between any two documents using cosine similarity where each document is a term-vector.

Consider a set of documents $x_1, x_2, \ldots, x_n \in \mathbb{R}^m$. Assume x_i has been normalized to 1.

1. To construct the adjacency graph, suppose the i^{th} node corresponds to the document x_i. We put an edge between nodes i and j if x_i is among p nearest neighbors of x_j or x_j is among p nearest neighbors of x_i.

2. We construct the similarity matrix S as follows: If nodes i and j are connected, $S_{ij} = \mathbf{x_i^T x_j}$; Otherwise, $S_{ij} = 0$.

3. We apply PCCA+ algorithm (Algorithm 1) to obtain the membership matrix for the documents to clusters. The maximum membership criteria is used to cluster documents. We perform 3 sets of experiments to evaluate the cluster quality identified by PCCA+. In the first set of experiments using the unlabeled dataset, we chose the number of clusters using the eigen-gap measure. It was observed that using this measure the number of clusters were exactly equal to the number of categories in the labeled data. Table 4 shows the purity measure for a few datasets. In the second set of experiments, we compare PCCA+ with other clustering based on LSI [3], spectral clustering method, LPI [2] and Nonnegative Matrix Factorization clustering method [24, 23] (refer Table 3, Table2). The evaluations are conducted for the cluster numbers ranging from two to ten. For each given cluster number k, 50 test runs are conducted on different randomly chosen clusters. The clustering performance is evaluated by comparing the obtained label of each document with that provided by the document corpus. Given a document x_i, let r_i and s_i be the obtained cluster label and the label provided by the corpus, respectively (refer [24]). The accuracy measure is defined as $\frac{\sum_{i=1}^{n} \delta(s_i, map(r_i))}{n}$, where n is the total num-

[1] The parameter values for N-Cut are taken from http:://note.sonots.com/SciSoftware/ NcutImageSegmentation.html, where the author claims that these parameter values for N-Cut produce the best results

[2] Reuters-21578 corpus is available at http://www.daviddlewis.com/ resources/testcollections/reuters21578/

[3] The homepage of 20 Newsgroups dataset is http://qwone.com/ jason/20Newsgroups/

[4] Nist Topic Detection and Tracking corpus is at http://www.nist.gov/speech/tests/tdt/tdt98/index.html

Table 2: Accuracy Measure for TFTD2

K	K-means	LSI	LPI	LE	NMF-NCW	PCCA+
2	0.871	0.913	0.963	0.923	0.925	0.9948
3	0.775	0.815	0.884	0.816	0.807	0.9810
4	0.732	0.773	0.843	0.793	0.787	0.9528
5	0.671	0.704	0.780	0.737	0.735	0.9424
6	0.655	0.683	0.760	0.719	0.722	0.9403
7	0.623	0.651	0.724	0.694	0.689	0.9331
8	0.582	0.617	0.693	0.650	0.662	0.7953
9	0.553	0.587	0.661	0.625	0.623	0.859
10	0.545	0.573	0.646	0.615	0.616	0.817
Avg	0.667	0.702	0.657	0.730	0.730	0.913

Table 3: Accuracy Measure for Reuters

K	K-means	LSI	LPI	LE	NMF-NCW	PCCA+
2	0.989	0.992	0.998	0.998	0.985	0.9715
3	0.974	0.985	0.996	0.996	0.953	0.9794
4	0.959	0.970	0.996	0.996	0.964	0.9801
5	0.948	0.961	0.993	0.993	0.980	0.9693
6	0.945	0.954	0.993	0.992	0.932	0.970
7	0.883	0.903	0.990	0.988	0.921	0.9703
8	0.874	0.890	0.989	0.987	0.908	0.9699
9	0.852	0.870	0.987	0.984	0.895	0.9698
10	0.835	0.850	0.982	0.979	0.898	0.9697
Avg	0.918	0.931	0.982	0.990	0.937	0.9722

ber of documents, $\delta(x, y)$ is the delta function, and map(r_i), is the permutation mapping function, that maps each cluster label r_i to the equivalent label from the data corpus. We observe that PCCA+ provides better quality clusters as compared to other clustering techniques in most of the cases while comparable in others. The third set of experiments is performed using the Reuters-21578 dataset, in this experiment 7 most frequent categories are considered but we do not remove documents belonging to multiple categories. After removing documents whose label sets or main texts are empty, 8,866 documents are retained where only 3.37% of them are associated with more than one class labels. After randomly removing documents with only one label, a text categorization data set containing 1998 documents is obtained (Table 5 provides details of the used categories). We run PCCA+ on this document and generate the connectivity information across categories using the definition provided in Def 2. The macro-transition operator (normalized) which quantifies relation between categories is shown in Table 6. The interesting observation to note is that the macro-transition operator provides high values across categories with seemingly related nature, for example the pairs Money fx-Trade, Trade-Crude, Trade-Interest have high values, while the pair Grain-Earn have low connectivity values.

We observe that PCCA+ is competitive with the best algorithm for all the values of number of clusters. We also observe in Table 4 that PCCA+ has a very high purity measure for document clustering.

Table 4: Purity measure for Document Clusters

Dataset	Number of docs	K	Purity
TDT2	9394	30	0.9344
Reuters	8293	65	0.9694
Newsgroup	18846	20	0.9023

Table 5: Details of the used Reuters21578 top 7 Categories

Category	Number of Docs
Earn	831
Acquisition	482
Money-fx	299
Grain	153
Crude	128
Trade	154
Interest	261

5.3 Synthetic Datasets

We demonstrate the goodness of quality of clustering obtained by using PCCA+ on various synthetic datasets with different structural properties.

1. Conisder the set of datapoints $x_1, x_2, \ldots, x_n \in \mathbb{R}^m$: The similarity matrix is constructed as follows

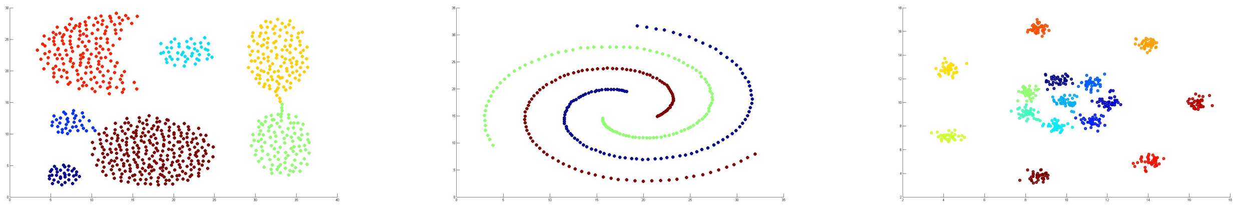
$$S_{ij} = \begin{cases} \exp - \frac{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}{2} & if \; \|x_i - x_j\| \\ & \leq threshold \text{ and } i \neq j \\ 0 & otherwise \end{cases}$$

   where Σ is the covariance matrix

2. We find the number of clusters using spectral gap method by finding the top k eigenvalues for which $\frac{(e_{k+1} - e_k)}{1 - e_k} > t_c$, where t_c is the spectral gap threshold.

Table 6: Macro connectivity information for Reuters21578 top 7 Categories

	Earn	Acquisition	Money-fx	Grain	Crude	Trade	Interest
Earn	0.106	0.093	0.031	0.052	0.259	0.255	0.202
Acquisition	0.021	0.104	0.058	0.075	0.243	0.279	0.216
Money-fx	0.005	0.056	0.184	0.074	0.228	0.283	0.166
Grain	0.005	0.038	0.039	0.287	0.189	0.169	0.269
Crude	0.015	0.062	0.06	0.095	0.291	0.268	0.205
Trade	0.013	0.07	0.073	0.083	0.261	0.328	0.169
Interest	0.009	0.043	0.034	0.107	0.164	0.138	0.501



(a) Aggregation(7 clusters)

(b) Spiral(3 clusters)

(c) R15(15 clusters)

Figure 5: Clustering on Synthetic Datasets using PCCA+

3. We apply PCCA+ algorithm (Algorithm 1) to obtain the membership matrix for the data points to clusters. Figure 5 shows the clusters obtained using PCCA+. Table 7 shows the purity measure of the clusters obtained using PCCA+ for the same number of clusters k as present in the labeled data (Note that the spectral gap method always identified the same number of clusters a present in the labeled data)

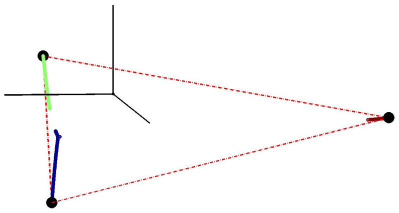


Figure 6: Simplex identified by PCCA+ for the spiral dataset in 5b. The data points are shown as colored dots and are clustered around the vertices of the transformed simplex (shown in black dots). The original basis is shown in black lines.

We observe in Table 7 that PCCA+ obtained clustering with a very high purity measure, even with datasets of different structural properties as is seen in the original data space. We also plot the simplex identified by PCCA+ for the case of spiral dataset in Figure 6. Few observations to note here are: a) the simplex is linearly transformed from its corresponding regular simplex structure, which shows the first order perturbation, and b) the data points(plotted in colored markers) around the vertices of the transformed simplex shows higher order perturbations around the simplex structure.

Table 7: Purity measure for Synthetic Dataset

Dataset	Number of datapoints	K	Purity
R15	600	15	99.7
Spiral	312	3	100.0
Aggregation	788	7	99.6

6. DISCUSSION AND CONCLUSION

In this work we introduced the notion of clustering that better exploits the structural similarity information in datasets. We propose using the vertices of a simplex in the Eigen subspace of the Laplacian of the structure information as abstract clusters. We demonstrated the use of the spectral clustering algorithm, PCCA+, for deriving memberships to such abstract clusters across a variety of domains. The experiments show that we are competitive with the state-of-the-art clustering methods in three domains, namely, image segmentation, text clustering, and synthetic datasets. We significantly outperform other generic clustering methods in these domains. We also have promising results in few other domains like Community Detection and Spatial Abstraction. We believe that this approach represents a new direction in clustering and should open the doors for more robust and efficient spectral methods in different domains.

One of the chief criticisms of spectral clustering methods is the higher time complexity. Since we were dealing with largely sparse similarity matrices, we were able to employ the Lancsoz's algorithm ([17]) and achieved $\mathcal{O}(n)$ per round complexity with typically far fewer rounds than the data size. The Lancsoz algorithm is also amenable to parallel implementations and we are exploring this direction to achieve better speedup. While the spectral gap method for determining the number of clusters usually works well in practice we are exploring the relations between the cluster structure and the graph spectra to derive provably robust methods. Also as mentioned earlier, we can also compute the connectivity between the clusters. This information can be used, depending on the domain to derive more efficient planners, stochastic block models, or super pixel information. This is a very promising line of inquiry that we are currently pursuing in some of the domains

7. REFERENCES

[1] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 96–103, New York, NY, USA, 1998. ACM.

[2] D. Cai, X. He, J. Han, and S. Member. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 2005.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the american society for information science*, 41(6):391–407, 1990.

[4] P. Deuflhard and M. Weber. Robust Perron Cluster Analysis in Conformation Dynamics. Technical report.

[5] M. Ester, H. peter Kriegel, J. S, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[6] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[7] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral, 2000.

[8] T. Kato. *Perturbation Theory for Linear Operators*. 1966.

[9] X. Liu, Y. Gong, W. Xu, and S. Zhu. Document clustering with cluster refinement and model selection capabilities. In *In Proceedings of the 25th International ACM SIGIR conference on research and development in information retrieval*, pages 191–198. ACM Press, 2002.

[10] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 2007.

[11] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[12] M. Meila and J. Shi. A random walks view of spectral segmentation. 2001.

[13] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, 2001.

[14] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Inf. Process. Manage.*, 2006.

[15] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[16] N. D. Singpurwalla and J. M. Booker. Membership functions and probability measures of fuzzy sets. *Journal of the American Statistical Association*, 2004.

[17] A. R. Thomas Ericsson. *Mathematics Of Computation, American Mathematical Society*, 1980.

[18] C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage. Descriptive matrix factorization for sustainability adopting the principle of opposites. *Data Min. Knowl. Discov.*, 2012.

[19] M. Weber, W. Rungsarityotin, and A. Schliep. Characterization of transition states in conformational dynamics using fuzzy sets. Technical report, March 2002.

[20] M. Weber, W. Rungsarityotin, and A. Schliep. Perron Cluster Analysis and Its Connection to Graph Partitioning for Noisy Data. Technical report, November 2004.

[21] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs.

[22] A. P. Wolfe and A. G. Barto. Identifying useful subgoals in reinforcement learning by local graph partitioning. In *In Proceedings of the Twenty-Second International Conference on Machine Learning*, 2005.

[23] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 267–273, New York, NY, USA, 2003. ACM.

[24] H. Zha, X. He, C. Ding, H. Simon, and M. Gu. Spectral relaxation for k-means clustering. pages 1057–1064. MIT Press, 2001.