Robust Spectral Clustering: A Locality Preserving Feature Mapping Based on M-estimation

Aylin Taştan, Michael Muma and Abdelhak M. Zoubir
Signal Processing Group
Technische Universität Darmstadt
64283 Darmstadt, Germany
{atastan,muma,zoubir}@spg.tu-darmstadt.de

***Abstract*—Dimension reduction is a fundamental task in spectral clustering. In practical applications, the data may be corrupted by outliers and noise, which can obscure the underlying data structure. The effect is that the embeddings no longer represent the true cluster structure. We therefore propose a new robust spectral clustering algorithm that maps each high-dimensional feature vector onto a low-dimensional vector space. Robustness is achieved by posing the locality preserving feature mapping problem in form of a ridge regression task that is solved with a penalized M-estimation approach. An unsupervised penalty parameter selection strategy is proposed using the Fiedler vector, which is the eigenvector associated with the second smallest eigenvalue of a connected graph. More precisely, the penalty parameter is selected, such that, the corresponding Fiedler vector is Δ-separated with a minimum information loss on the embeddings. The method is benchmarked against popular embedding and spectral clustering approaches using real-world datasets that are corrupted by outliers.**

***Index Terms*—embedding, clustering, spectral clustering, feature mapping, dimension reduction**

I. Introduction

Dimension reduction and feature extraction are fundamental in many clustering algorithms that have been intensively researched for decades [1]-[4]. Spectral clustering (SC) is a simple and effective tool that relies on the eigenfunctions of the Laplace-Beltrami operator on a manifold to discover the intrinsic structure hidden in the data. It has various applications such as in face recognition and image segmentation [5].

A popular way of estimating eigenvectors of a Laplacian is the method of Laplacian eigenmaps [1], which is a manifold learning technique motivated by the correspondence between the graph Laplacian and the Laplace-Beltrami operator on a manifold. The term Laplacian eigenmaps refers to a nonlinear method that embeds high-dimensional feature vectors into a low-dimensional vector space while preserving certain local properties. Locality Preserving Indexing (LPI) is motivated by determining the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator in an attempt at discovering the inherent nonlinear structure. The computational complexity of LPI can mainly be attributed to computing a complete singular value decomposition (SVD) and it has been reduced in [3]-[4], making such approaches attractive in practice. However, in real-world scenarios the data may be corrupted by outliers and noise [6], leading to a performance degradation. Existing robust algorithms for spectral clustering have been proposed to minimize the effect of outliers in representation space, e.g. [7] or in the projection operation [8]. The robust projection operation as in [8], uses the ℓ_1 norm that creates a different eigenbasis and it requires prior information about the data, such as, representative samples. To the best of our knowledge, an unsupervised robust projection algorithm that uses the ℓ_2 norm as in the eigen-decomposition of the original spectral clustering has not been proposed in the literature.

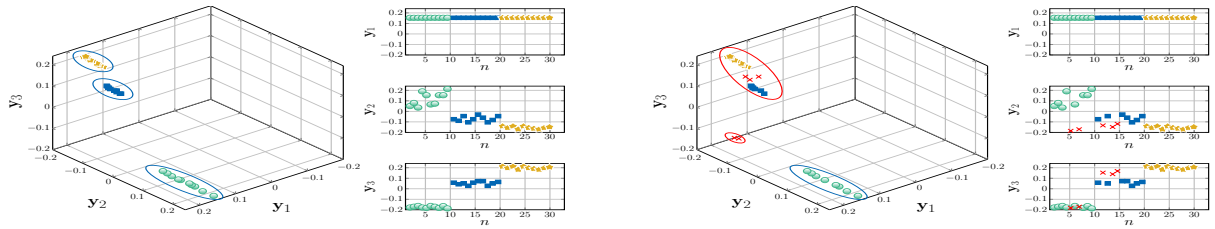
To integrate robustness in spectral clustering, we propose a *robust locality preserving feature mapping (RLPFM)* and *an unsupervised penalty parameter selection algorithm* using the geometric structure of well-spread embeddings. Building upon regularized locality preserving indexing (RLPI), which is a computationally efficient extension of the LPI framework that regularizes the eigenvectors, we propose a robust M-estimation approach to feature embedding to mitigate the effect of outliers on the determination of the group structure. The penalty parameter, which is a key factor for the performance of RLPI, is selected, such that, the estimated Fiedler vector is Δ-separated with minimum information loss.

The remaining paper is organized as follows. Section II briefly revisits LPI while Section III contains the motivation and problem formulation. The proposed robust spectral clustering method is detailed in Section IV. Section V demonstrates the performance of the proposed method in comparison to popular embedding and spectral clustering approaches. Finally, conclusions are drawn in Section VI.

II. Locality Preserving Indexing for Spectral Clustering

Suppose that a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ with m denoting the dimension and n the number of feature vectors, can be represented as a graph $G = \{V, E, \mathbf{W}\}$ where V denotes the vertices, E represents the edges, and $\mathbf{W} \in \mathbb{R}^{n \times n}$ is the nonnegative definite affinity matrix that is computed from a similarity measure, e.g. cosine similarity. Spectral clustering [1] maps the original m dimensional feature vectors onto a smaller k dimensional vector space by finding the eigenvectors associated with the k smallest eigenvalues of the following eigen-problem

The authors are with the Signal Processing Group, Technische Universität Darmstadt, Darmstadt, Germany (e-mail: atastan,muma,@spg.tu-darmstadt.de; muma@spg.tu-darmstadt.de; zoubir@spg.tu-darmstadt.de).



(a) (b)

Fig. 1: The eigenvectors associated with $k = 3$ smallest eigenvalues for a data matrix.

$$\mathbf{L}\mathbf{y} = \lambda \mathbf{D}\mathbf{y}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ denotes the eigenvector associated with the eigenvalue λ, $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal weight matrix with overall edge weights $d_{i,i} = \sum_j w_{i,j}$ on the diagonal, and $\mathbf{L} \in \mathbb{R}^{n \times n}$ is the Laplacian matrix defined by $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

According to the Theorem 2 in [2], for an eigenvector $\mathbf{y} \in \mathbb{R}^n$ with $\mathbf{y} = \mathbf{X}^\top \boldsymbol{\beta}$, the LPI finds a transformation vector $\boldsymbol{\beta} \in \mathbb{R}^m$ that is the eigenvector associated with the smallest eigenvalue of the generalized eigen-problem

$$\mathbf{X}\mathbf{L}\mathbf{X}^\top \boldsymbol{\beta} = \lambda \mathbf{X}\mathbf{D}\mathbf{X}^\top \boldsymbol{\beta} \tag{2}$$

with the same eigenvalue λ as in Eq. (1). This fundamental property of LPI, gives identical solutions to the spectral clustering if the data matrix $\mathbf{X}$ is a full rank square matrix. Thus, building upon [3], the LPI basis functions can be determined in two consecutive steps for spectral clustering. First, the k eigenvectors $\mathbf{y}_1, \ldots, \mathbf{y}_k$ associated with the k smallest eigenvalues $\lambda_1 < \cdots < \lambda_k$ in Eq. (1) is computed. Then, for each eigenvector $\mathbf{y}_j \in \mathbb{R}^n$, where $j = 1, \ldots, k$, LPI estimates a transformation vector $\boldsymbol{\beta}_j \in \mathbb{R}^m$ that satisfies $\mathbf{y}_j = \mathbf{X}^\top \boldsymbol{\beta}_j$ by solving the following least squares problem

$$\hat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j}{\operatorname{argmin}} \sum_i^n (\boldsymbol{\beta}_j^\top \mathbf{x}_i - y_{i,j})^2, \tag{3}$$

where $y_{i,j}$ is the ith mapping point in the jth eigenvector $\mathbf{y}_j$ and $\hat{\boldsymbol{\beta}}_j$ is the estimated jth transformation vector.

III. Motivation and Problem Formulation

A. Motivation

To motivate the use of robust methods, this section provides an illustrative discussion of possible outlier effects on spectral clustering. Fig. 1a shows an example, where the data that consists of $n = 30$ feature vectors can be separated into $k = 3$ disjoint clusters by the popular Laplacian eigenmaps method [1], which analyzes the eigenvectors corresponding to the three smallest eigenvalues. The ellipsoids around the yellow, blue, and green feature vectors highlight the discovered clusters. Fig. 1b uses the same dataset, except that six blue and green points have been replaced by outliers that are marked as red crosses. In the context of clustering, outliers are, generally speaking, defined as data points that do not follow the cluster structure that is inherent to the large majority of the data. We can distinguish two different types of outliers: On the one hand, an outlier may be a point that does not have any similarity with any of the clusters. On the other hand, an outlier may also be defined as a point that has considerable similarity with multiple clusters. In both cases, as illustrated in Fig. 1b, the outliers obscure the cluster structure inherent to the eigenvectors. In this example, the popular Laplacian eigenmaps method is not able to correctly split the data into the yellow, blue and green clusters. Instead, it opens up a cluster for the outliers that are not associated with any of the clusters, and it fuses the yellow and blue data points into a single cluster. Robust spectral clustering methods should be designed to be less sensitive to outliers. M-estimation is a widely used robust alternative to least-squares estimation when the data is subject to heavy-tailed noise and outliers [6]. Building upon the concepts of robust statistics [6], we propose an M-estimation approach, that down-weights outlying data points in the objective function, as will be detailed in the next section.

B. Problem Formulation

Given a dataset of feature vectors $\mathbf{X} \in \mathbb{R}^{m \times n}$, the goal of this work is to embed each feature vector into a k dimensional space where k denotes the specified number of clusters. Robustness implies that the method is not heavily affected by a few outliers in the dataset.

IV. Robust Spectral Clustering

This section is dedicated to robust spectral clustering. We first describe how we use penalized M-estimation for robust locality preserving feature mapping. We then propose a strategy for the penalty parameter selection. Finally, we analyze the computational complexity. The pseudo-code is given in Algorithm 1.

A. M-estimation for Locality Preserving Feature Mapping

Assume that the dataset $\mathbf{X}$ is corrupted by outliers and noise. The mappings in dimension-reduced space can then be written as

$$y_{i,j} = \boldsymbol{\beta}_j^\top \mathbf{x}_i + \epsilon_{i,j}, \tag{4}$$

where $y_{i,j} \in \mathbb{R}$ denotes the mapping point for the ith feature vector $\mathbf{x}_i$ and jth transformation vector $\boldsymbol{\beta}_j$, and $\epsilon_{i,j} \in \mathbb{R}$ represents noise and additive outliers. For an embedding operation from the m dimensional space to the k dimensional space, the error vector $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times 1}$ is constructed by using embedding errors of all feature vectors such that $\epsilon_i = \sum_j^k \epsilon_{i,j}$, where $\epsilon_i \in \boldsymbol{\epsilon}$ denotes the embedding error of the ith feature vector. Then, the transformation vector $\boldsymbol{\beta}_j$ can be computed using penalized ridge regression M-estimation [9] by solving the following zero gradient equation

$$-\sum_{i=1}^n \psi\Big(\frac{\epsilon_i}{\hat{\sigma}}\Big)\Big(\frac{\mathbf{x}_i^\top}{\hat{\sigma}}\Big) + \gamma \boldsymbol{\beta}_j = \mathbf{0}, \tag{5}$$

where γ denotes the penalty parameter, $\hat{\sigma}$ is a robust scale estimate of $\boldsymbol{\epsilon}$ and ψ is a bounded and continuous odd function

called the score-function. A popular M-estimator is defined by Huber's function

$$\psi\Big(\frac{\epsilon_i}{\hat{\sigma}}\Big) = \begin{cases} \frac{\epsilon_i}{\hat{\sigma}}, & \text{for} \quad \big|\frac{\epsilon_i}{\hat{\sigma}}\big| \leq c \\ c\text{sign}\big(\frac{\epsilon_i}{\hat{\sigma}}\big), & \text{for} \quad \big|\frac{\epsilon_i}{\hat{\sigma}}\big| > c \end{cases}, \tag{6}$$

where $\text{sign}(x)$ is the sign function defined as $\text{sign}(x) = x/|x|$, c is the tuning parameter that trades off robustness against outliers and efficiency under a Gaussian distribution (see [6] for a discussion). A common choice for $\hat{\sigma}$ is the normalized median absolute deviation [6] that is defined by

$$\hat{\sigma} = \text{madn}(\boldsymbol{\epsilon}) = 1.4826 \cdot \text{med}|\boldsymbol{\epsilon} - \text{med}(\boldsymbol{\epsilon})|, \tag{7}$$

with $\text{med}(\boldsymbol{\epsilon})$ being the median.

B. Theoretical Analysis

Using matrix-notation, the solution to Eq. (5) can conveniently be written as

$$\hat{\boldsymbol{\beta}}_j = (\mathbf{X}\boldsymbol{\Omega}\mathbf{X}^\top + \gamma\hat{\sigma}^2\mathbf{I})^{-1}\mathbf{X}\boldsymbol{\Omega}\mathbf{y}. \tag{8}$$

Here, $\boldsymbol{\Omega} \in \mathbb{R}^{n\times n}$ is diagonal weighting matrix defined by $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_n)$ with $\omega_i = \omega\Big(\frac{\epsilon_i}{\hat{\sigma}}\Big)$ where

$$\omega\Big(\frac{\epsilon_i}{\hat{\sigma}}\Big) = \begin{cases} \psi\Big(\frac{\epsilon_i}{\hat{\sigma}}\Big)\Big/\Big(\frac{\epsilon_i}{\hat{\sigma}}\Big) & \text{for} \quad \frac{\epsilon_i}{\hat{\sigma}} \neq 0 \\ 1, & \text{for} \quad \frac{\epsilon_i}{\hat{\sigma}} = 0 \end{cases} \tag{9}$$

In the following Theorem, we show the link between the proposed RLPFM algorithm and spectral clustering, whose eigenvectors are obtained as a special case.

Theorem 1. *Suppose* $\mathbf{y}$ *is the eigenvector of eigen-problem in Eq.* (2)*. Further, let* $\boldsymbol{\Omega} \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m}$ *be two weighting matrices, such that,* $\mathbf{U}^\top\boldsymbol{\Psi}\mathbf{U} = \mathbf{I}$ *and* $\mathbf{V}^\top\boldsymbol{\Omega}\mathbf{V} = \mathbf{I}$*. If* $\mathbf{y}$ *is in the space spanned by the row vectors of the weighted data matrix* $\mathbf{X}^* = \mathbf{X}\boldsymbol{\Omega}$*, the corresponding transformation vector* $\hat{\boldsymbol{\beta}}$ *estimated with RLPFM will be the eigenvector of eigen-problem in Eq.* (2) *as* γ *deceases to zero.*

Proof. See Appendix. □

C. Penalty Parameter Selection

Motivated by the key role of the penalty parameter on the performance of RLPI [3], and the analysis on the geometric structure of well-spread ℓ_2^2-representations in [10], we propose a penalty parameter selection algorithm, such that, every pair of subsets $s_i \in \mathbf{s}$ and $t_j \in \mathbf{t}$ is at least $\Delta = \phi(1/\log^{-2/3} n)$ apart in ℓ_2^2 distance. The Fiedler vector, which is associated to the second smallest eigenvalue is used to define these two sets. Assume that for each $\gamma_i \in \boldsymbol{\gamma} = [\gamma_{\min}, \ldots, \gamma_{\max}] \in \mathbb{R}^N$, there exists a Fiedler vector estimate $\hat{\mathbf{y}}_2^{(\gamma_i)}$ that maps graph vertices onto a real line for the ith penalty parameter γ_i. Based on [10], for a suitable constant κ, the mapping points located on the right and left hand side of κ can be used as initial candidates for sets $\mathbf{s}$ and $\mathbf{t}$. From the Fiedler vector properties [11], the constant κ can be defined as $\kappa = 0$ for $\hat{\mathbf{y}}_2^{(\gamma_i)}$. By definition, the mapping points must vary between zero and one. Thus, after selecting the members of subsets $\mathbf{s}^{(\gamma_i)} \in \mathbb{R}^{N_s}$ and $\mathbf{t}^{(\gamma_i)} \in \mathbb{R}^{N_t}$ associated with γ_i, the sets can be designed using rescaled mappings as

$$\begin{aligned} \mathbf{s}^{(\gamma_i)} &= \{\bar{y}_{2,j}^{(\gamma_i)} : \hat{y}_{2,j}^{(\gamma_i)} > \kappa\} \\ \mathbf{t}^{(\gamma_i)} &= \{\bar{y}_{2,j}^{(\gamma_i)} : \hat{y}_{2,j}^{(\gamma_i)} \leq \kappa\}, \end{aligned} \tag{10}$$

Algorithm 1: Robust Spectral Clustering

Input: A data $\mathbf{X}$ and affinity matrix $\mathbf{W}$, k,$N_{\min}$
Eigenvector Estimation using RLPFM
for $\gamma_i = \gamma_{\min}, \ldots, \gamma_{\max}$ **do**
  Initialization:
  Evaluate the eigenvectors $\mathbf{y}_1, \ldots, \mathbf{y}_k$ as in Eq. (1)
  Get $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k] \in \mathbb{R}^{m\times k}$ for $\mathbf{y}_k = \mathbf{X}^\top\boldsymbol{\beta}_k$
  RLPFM
  Compute the error vector $\boldsymbol{\epsilon} \in \mathbb{R}^n$ using Eq. (4)
  where $\epsilon_i = \sum_j^k \epsilon_{i,j}$ for $\epsilon_i \in \boldsymbol{\epsilon}$
  Compute $\hat{\sigma}$ via Eq. (7)
  Calculate $\omega_i = \omega(\frac{\epsilon_i}{\hat{\sigma}})$, $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega})$, via Eq. (9)
  Solve Eq. (8) and estimate $\hat{\boldsymbol{\beta}}_1^{(\gamma_i)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(\gamma_i)}$
  Estimate $\hat{\mathbf{y}}_1^{(\gamma_i)}, \ldots, \hat{\mathbf{y}}_k^{(\gamma_i)}$ for $\hat{\mathbf{y}}_k^{(\gamma_i)} = \mathbf{X}^\top\hat{\boldsymbol{\beta}}_k^{(\gamma_i)}$
  Δ-separated sets
  Generate sets $\mathbf{s}^{(\gamma_i)}$ and $\mathbf{t}^{(\gamma_i)}$ via Eq. (10)
  while $N_s > N_{\min}$ and $N_t > N_{\min}$ **do**
    Create $\mathbf{r} \in \mathbb{R}^{N_{\mathbf{r}}^{(\gamma_i)}}$ using Eq. (11) and update $N_{\mathbf{r}}^{(\gamma_i)}$
    if $\mathbf{s}^{(\gamma_i)}$ and $\mathbf{t}^{(\gamma_i)}$ are Δ-separated **then**
      break
    end
  end
  Collect $N_{\mathbf{r}}^{(\gamma_i)}$ into a vector $\mathbf{h} \in \mathbb{R}^N$
end
Minimize the $N_{\mathbf{r}}^{(\gamma_i)}$ and estimate $\hat{\gamma}$ using Eq. (12)
Estimate $\hat{\mathbf{B}}^{\hat{\gamma}} = [\hat{\boldsymbol{\beta}}_1^{\hat{\gamma}}, \ldots, \hat{\boldsymbol{\beta}}_k^{\hat{\gamma}}] \in \mathbb{R}^{m\times k}$ for $\hat{\gamma}$
Estimate $\hat{\mathbf{y}}_1^{(\hat{\gamma})}, \ldots, \hat{\mathbf{y}}_k^{(\hat{\gamma})}$ where $\hat{\mathbf{y}}_k^{(\hat{\gamma})} = \mathbf{X}^\top\hat{\boldsymbol{\beta}}_k^{(\hat{\gamma})}$
Partitioning
Get $\hat{\mathbf{c}}_k$ by applying the k-means on $\hat{\mathbf{y}}_1^{(\hat{\gamma})}, \ldots, \hat{\mathbf{y}}_k^{(\hat{\gamma})}$
Output: An estimated label vector $\hat{\mathbf{c}}_k$ for k clusters

where $\hat{y}_{2,j}^{(\gamma_i)}$ and $\bar{y}_{2,j}^{(\rho_i)}$ denote the jth element of the estimated Fiedler vector $\hat{\mathbf{y}}_2^{(\gamma_i)}$ and that of rescaled the $\bar{\mathbf{y}}_2^{(\gamma_i)}$ for a candidate penalty parameter γ_i, respectively. If $\bar{\mathbf{y}}_2^{(\gamma_i)}$ is not well-spread, it can contain pairs of points $\bar{y}_{2,i}^{(\gamma_i)} \in \mathbf{s}$ and $\bar{y}_{2,j}^{(\gamma_i)} \in \mathbf{t}$ whose squared Euclidean distance is less than Δ that will be discarded as long as two sets have a sufficient number $N_{\min}$ of mapping points, i.e.,

$$\mathbf{r}^{(\gamma_i)} = \Big\{\bar{y}_{2,i}^{(\gamma_i)}, \bar{y}_{2,j}^{(\gamma_i)} : \|\bar{y}_{2,i}^{(\gamma_i)} - \bar{y}_{2,j}^{(\gamma_i)}\|_2^2 \leq \Delta\Big\}, \tag{11}$$

where $\mathbf{r}^{(\gamma_i)} \in \mathbb{R}^{N_{\mathbf{r}}}$ is a vector of discarded points from subset $\mathbf{s}^{(\gamma_i)}$ and $\mathbf{t}^{(\gamma_i)}$. The penalty parameter γ is estimated by minimizing the number of discarded points as

$$\hat{\gamma} = \underset{\gamma_i=\gamma_{\min},\ldots,\gamma_{\max}}{\arg\min} \{N_{\mathbf{r}}^{(\gamma_i)}\}, \tag{12}$$

where $N_{\mathbf{r}}^{(\gamma_i)}$ denotes number of discarded points for candidate penalty parameter γ_i.

D. Computational Complexity

The computational cost of operations is measured in flam [12] which is a compound operation consisting one addition and one multiplication. If the computational complexity is not specified using flam, the well known Landau notation is used. The RLPFM requires $n(p^2 - k^2)$ to $2n(p^2 - k^2)$ flam for the expansion and npk flam for the contraction phases for the initialization of eigenvectors, where p is the number of Lanczos basis vectors and k is the number of eigenvectors. The weighting operation of M-estimation requires repetitive medians that takes $O(n)$ time. For a sparse matrix, the least squares algorithm, such as, in [13] requires $t(2ns + 3n + 5m)$

Dataset	SC	LPI	RLPI	FastEFM	LSC	RLPFM
Fisheriris [15]	66.0	98.0	98.0	96.6	92.9	98.0
B. Cancer [16]	62.9	88.2	87.4	72.1	85.4	87.0
Ionosphere [17]	64.4	51.9	71.2	68.4	71.5	70.4
Parkinson [18]	50.4	53.2	60.4	61.0	52.1	60.0
Sonar [19]	54.3	55.3	56.3	54.6	51.1	60.6

TABLE I: k-means partitioning performance for real-world datasets. The average probability of detection shown in $\%$

where t is the number of iterations and s is the average number of nonzero features. However, if the matrix is dense, Cholesky decomposition requires $O(n^3)$ and in particular $\frac{1}{6}n^3$ flam [12]. Lastly, the Δ-seperated sets step requires $O(n\log n)$ time for sorting and a maximum of n flam for discarding for each candidate γ. In summary, for a sparse matrix the RLPFM step requires from

$$N_\gamma t(2ns + 3n + 5m) + n(p^2 - k^2 + pk + N_\gamma)$$

to

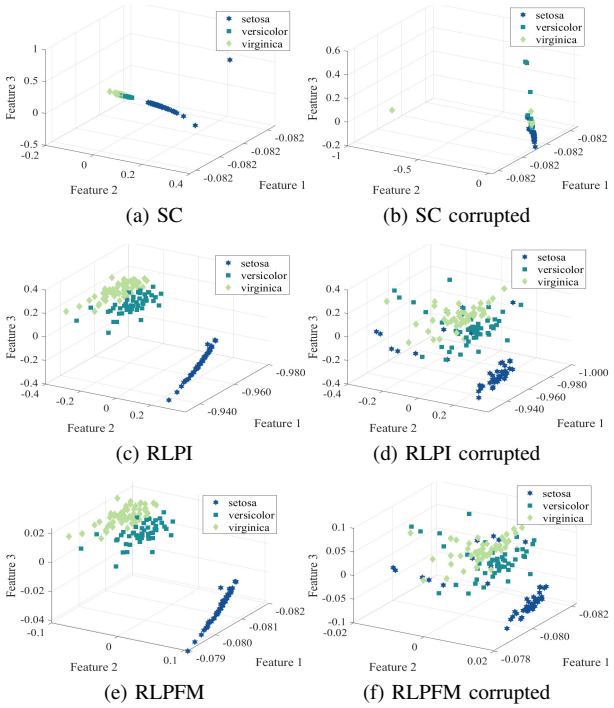
$$N_\gamma t(2ns + 3n + 5m) + n(2p^2 - 2k^2 + pk + N_\gamma)$$

flam in addition to $O(N_\gamma n)$, $O(N_\gamma n\log n)$ for repetitive medians and sorting where N_γ is the number of candidate penalty parameters.

V. Experimental Results

In this section, the proposed RLPFM is compared with five state-of-the-art methods including embedding approaches LPI [2] and RLPI [3] and spectral clustering approaches SC [1], fast large-scale spectral clustering via explicit feature mapping (FastEFM) [4], large scale spectral clustering with landmark-based sparse representation (LSC) [14]. The numerical experiments are performed with real-world databases Fisher's iris (Fisheriris) [15], diagnostic breast cancer (B. Cancer) [16], ionosphere [17], replicated acoustic features of Parkinson disease (Parkinson) [18], and connectionist bench (Sonar) [19] from the UCI machine learning repository. The parameter $N_{\min}$ for Δ-separated sets is defined as $N_{\min} = \frac{n}{10}$ where different values of $N_{\min}$ do not have a huge impact as long as $N_{\min}$ is a reasonably small value. To analyze performance numerically, average clustering accuracy $\bar{P}_{\text{acc}}$ is calculated by averaging clustering results for $N_E = 100$ repetitions and RLPI is performed with the proposed penalty parameter selection method to provide a fair comparison.

The clustering accuracy results are summarized for six different methods on five real-world datasets using k-means partitioning in Tab. I. As can be seen, the SC shows poor performance in terms of average clustering accuracy of 59.6% whereas almost all other clustering approaches have an average accuracy greater than 70%. The proposed RLPFM outperforms all its competitors with 75.2% and RLPI follows it by a narrow margin reaching 74.7% which indicates that the proposed penalty parameter selection algorithm is a promising approach that can be used in other regularized feature mapping algorithms. We also implemented a simple plug-in robustification that replaces k-means by k-medoids, however, it did not improve the partitioning results, and is therefore not reported in detail.



(a) SC (b) SC corrupted (c) RLPI (d) RLPI corrupted (e) RLPFM (f) RLPFM corrupted

Fig. 2: Estimated example feature spaces for Fisheriris dataset.

A. Robustness

To evaluate robustness against outliers, we contaminated the Fisheriris dataset as follows. The outliers were generated as $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \sigma\mathbf{r}$, where $\mathbf{r}$ denotes a vector of uniformly distributed random numbers in the interval $U(0, 1)$, σ is a constant, $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ are the original and corrupted ith feature vector for a randomly selected i, respectively. The examples of estimated eigenvectors for $k = 3$ clusters are shown in Fig. 2 for the original and corrupted cases. For the corrupted case, the examples shown for $\sigma = 5$ and the number of outliers in per cluster $N_{\text{out}} = 15$. Fig. 2a shows that, even the original Fisheriris dataset results in an outlier in the SC mappings that causes the method to break down. Figs. 2c and 2e show that both the RLPI and the proposed RLPFM produce similar and accurate mapping results for the original data. For corrupted data, Figs. 2d and 2f show that RLPFM and RLPI approximately preserve the cluster structure, and RLPFM reduces the effect of outliers by mapping them closer to the cluster centers.

The clustering accuracy is detailed according to different σ and N_{out} values in Fig. 3 and Fig. 4, respectively. Even though most of the algorithms have a clustering accuracy of more than 90% in the beginning, the performance of the competitors drops significantly after $\sigma = 3$. The proposed method is also more robust for an increasing number of outliers while its main competitor RLPI follows it by approximately margin of 10%.

VI. Conclusion

We proposed an unsupervised RLPFM including a penalty parameter selection approach for spectral clustering. The eigenvectors of a Laplacian matrix were reweighted and

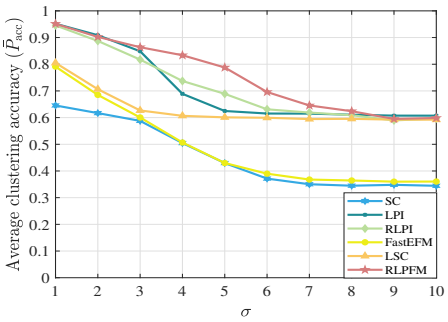


Fig. 3: $\bar{P}_{\text{acc}}$ for increasing number of σ values ($N_{\text{out}} = 15$).

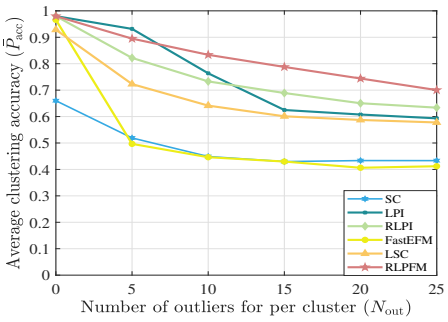


Fig. 4: $\bar{P}_{\text{acc}}$ for increasing N_{out} (σ=5).

penalized by optimizing the penalty parameter, such that, the corresponding Fiedler vector is Δ-separated with minimum information loss. The method was benchmarked on different real-world datasets and it showed promising performance compared to five popular competitors, especially in terms of robustness against outliers and noise.

APPENDIX

Suppose that $\text{rank}(\mathbf{X}) = \tau$, the SVD of $\mathbf{X}$ is $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{\Sigma} = \text{diag}(\Sigma_1, \ldots, \Sigma_\tau)$, $\mathbf{U} \in \mathbb{R}^{m\times\tau}$, $\mathbf{V} \in \mathbb{R}^{n\times\tau}$ and that $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$. Then, for the weighted singular value decomposition (WSVD) [20]

$$\mathbf{X}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{\Omega},$$

where $\mathbf{\Omega} \in \mathbb{R}^{n\times n}$ is square positive definite symmetric weight matrix such that $\mathbf{V}^\top\mathbf{\Omega}\mathbf{V} = \mathbf{I}$. Let $\mathbf{V}^*$ be a weighted matrix whose columns are weighted orthonormal eigenvectors of $\mathbf{V}$ as $\mathbf{V}^* = \mathbf{\Omega}\mathbf{V}$. Then, the orthogonality term can be also written as $\mathbf{V}^\top\mathbf{V}^* = \mathbf{I}$. If $\mathbf{y}$ is in the space spanned by column vectors of $\mathbf{V}^*$, $\mathbf{y}$ is spanned by row vectors of the weighted data matrix $\mathbf{X}^*$. Thus, $\mathbf{y}$ can be represented as a unique linear combination of column vectors $\mathbf{V}^*$ where column vectors of $\mathbf{V}^*$ are linearly independent. For a set of combination coefficients $\mathbf{b} \in \mathbb{R}^\tau$

$$\mathbf{V}^*\mathbf{b} = \mathbf{y} \;\Rightarrow\; \mathbf{\Omega}\mathbf{V}\mathbf{b} = \mathbf{y} \;\Rightarrow\; \mathbf{V}^\top\mathbf{\Omega}\mathbf{V}\mathbf{b} = \mathbf{V}^\top\mathbf{y} \;\Rightarrow\; \mathbf{b} = \mathbf{V}^\top\mathbf{y}$$

Substituting $\mathbf{b} = \mathbf{V}^\top\mathbf{y}$ into $\mathbf{V}^*\mathbf{b} = \mathbf{y}$ yields $\mathbf{V}^*\mathbf{V}^\top\mathbf{y} = \mathbf{y}$. Using the pseudo inverse of data matrix $\mathbf{X}^\dagger$ and weighted data matrix $(\mathbf{X}^*)^\dagger$ which can be written as

$$\mathbf{X}^\dagger = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top \quad \text{and} \quad (\mathbf{X}^*)^\dagger = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{\Psi},$$

for $\gamma \to 0$, $\mathbf{X}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{\Omega}$ and $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, Eq. (5) gives

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{\Psi}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{\Omega}\mathbf{y} \\ &= \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathbf{\Omega}\mathbf{y} \\ &= \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^\top\mathbf{y}.\end{aligned}$$

Further, if we insert $\hat{\boldsymbol{\beta}}$ into $\hat{\mathbf{y}} = \mathbf{X}^\top\hat{\boldsymbol{\beta}}$

$$\hat{\mathbf{y}} = \mathbf{X}^\top\hat{\boldsymbol{\beta}} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^\top\mathbf{y} = \mathbf{y}$$

$\hat{\boldsymbol{\beta}}$ is the eigenvector of eigen-problem in Eq. (2).

ACKNOWLEDGMENT

The work of A. Taştan is supported by the Republic of Turkey Ministry of National Education. The work of M. Muma has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY centre and is supported by the 'Athene Young Investigator Programme' of Technische Universität Darmstadt, Hesse, Germany.

REFERENCES

[1] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373-1396, 2003.

[2] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing" in *IEEE Trans. Knowl. Data Eng.*, vol. 17, pp. 1624-1637, 2005.

[3] D. Cai, X. He, W. V. Zhang and J. Han, "Regularized locality preserving indexing via spectral regression," in *Proc. 16th ACM Conf. Inf. Knowl. Manage.*, pp. 741-750, 2007.

[4] L. He, N. Ray, Y. Guan and H. Zhang, "Fast large-scale spectral clustering via explicit feature mapping," *IEEE Trans. Cybern.*, vol. 49, pp. 1058-1071, 2018.

[5] X. Wang, B. Qian and I. Davidson, "On constrained spectral clustering and its applications," *Data Min. Knowl. Discovery*, vol. 28, pp. 1-30, 2014.

[6] A. M. Zoubir, V. Koivunen, E. Ollila and M. Muma, *Robust statistics for signal processing*, Cambridge, 2018.

[7] Z. Li, F. Nie, X. Chang, L. Nie, H. Zhang and Y. Yang, "Rank-constrained spectral clustering with flexible embedding," *IEEE Trans. Neural Networks Learn. Syst.* , vol. 29, pp. 6073-6082, 2018.

[8] X. Peng, Z. Yi and H. Tang, "Robust subspace clustering via thresholding ridge regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 29, 2015.

[9] S. A. Razavi, E. Ollila, V. Koivunen, "Robust greedy algorithms for compressed sensing," *Proc. of the 20th European Signal Process. Conf.*, pp. 969-973, 2012.

[10] S. Arora, S. Rao and U. Varizani, "Expander flows, geometric embeddings and graph partitioning" *J. ACM*, vol. 56, pp. 1-37, 2009.

[11] D. A. Spielman and S. -Hua Teng, "Spectral partitioning works: Planar graphs and finite element meshes", in *Proc. 37th Conf. Found. Comput. Sci.*, pp. 96-105, 1996.

[12] G. W. Stewart, *Matrix Algorithms: Volume I Basic Decompositions*, Society for Industrial and Applied Mathematics, 1998.

[13] C. C. Paige and M. A. Saunders, "LSQR: Sparse linear equations and least squares problems," *ACM Trans. Math. Software*, vol. 8, pp. 195–209, 1982.

[14] D. Cai, and X. Chen, "Large scale spectral clustering with landmark-based sparse representation," *IEEE Trans. Cybern.*, vol. 45, pp. 1669-1680, 2014.

[15] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

[16] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation applied to breast cytology diagnosis," in *Proc. Natl. Acad. Sci*, vol. 87, pp. 9193-9196, 1989.

[17] V. G. Sigilitto, S. P. Wing, L. V. Hutton and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," *Johns Hopkins APL Tech. Dig.*, vol. 10, pp. 262–266, 1989.

[18] L. Naranjo, C. J. Perez, Y. Campos-Roca and J. Martin, "Addressing voice recording replications for Parkinson's disease detection," *Expert Syst. Appl.*, vol. 46, pp. 286-292, 2016.

[19] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Networks*, vol. 1, pp. 75-89, 1988.

[20] E.f. Galba, "Weighted singular decomposition and weighted pseudoinversion of matrices," *Ukrainian Math. J.*, vol. 48, pp. 1618-1622, 1996.