Learning Strictly Orthogonal p-Order Nonnegative Laplacian Embedding via Smoothed Iterative Reweighted Method

Haoxuan Yang[1*], **Kai Liu**[1*], **Hua Wang**[1] and **Feiping Nie**[2]

[1]Department of Computer Science, Colorado School of Mines, Golden, CO 80401, U.S.A.
[2]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
haoxuan@mymail.mines.edu, {cskailiu, huawangcs, feipingnie}@gmail.com

Abstract

Laplacian Embedding (LE) is a powerful method to reveal the intrinsic geometry of high-dimensional data by using graphs. Imposing the orthogonal and nonnegative constraints onto the LE objective has proved to be effective to avoid degenerate and negative solutions, which, though, are challenging to achieve simultaneously because they are nonlinear and nonconvex. In addition, recent studies have shown that using the p-th order of the ℓ_2-norm distances in LE can find the best solution for clustering and promote the robustness of the embedding model against outliers, although this makes the optimization objective nonsmooth and difficult to efficiently solve in general. In this work, we study LE that uses the p-th order of the ℓ_2-norm distances and satisfies both orthogonal and nonnegative constraints. We introduce a novel *smoothed iterative reweighted method* to tackle this challenging optimization problem and rigorously analyze its convergence. We demonstrate the effectiveness and potential of our proposed method by extensive empirical studies on both synthetic and real data sets.

1 Introduction

Data that reside in high-dimensional space are often intractable due to the computational complexity and the lack of intuition. In traditional Laplacian embedding (LE), the intrinsic subspace/manifold in high-dimensional space can be explored in such a way that the inherent data structures are well preserved and made more apparent due to the fact that the features less related to others will be pruned. LE is a powerful nonlinear graph based embedding method, which was first introduced as "quadratic placement" in 1970s [Hall, 1970]. Recently, the real power of LE was revealed as its relation to graph clustering [Hagen and Kahng, 1992; Chan *et al.*, 1994; Shi and Malik, 2000]. The eigenvectors of the Laplacian matrix provide the approximation to the Ratio Cut spectral clustering [Chan *et al.*, 1994] and it has been proved that LE and ratio cut clustering are mathematically identical [Luo *et al.*, 2009].

However, both positive and negative values in the solution of multi-way clustering tasks make the results hard to interpret directly, because the clustering indicator vectors require nonnegative results. In two-way clustering, this is not a problem, because a linear Ψ-transformation [Ding and He, 2004] of the eigenvectors leads to two genuine indicator vectors (each row has only one nonzero positive entry). Thus, mixed-sign solution is a generic difficulty for multi-way spectral clustering. To solve this problem, a clustering step usually has to be performed after the embedding is learned. That is, in traditional way, the clustering indicator vectors approximated by the eigenvectors of the Laplacian matrix will be grouped by using K-means clustering [Hartigan and Wong, 1979] in the eigenvector space. Thus, the traditional clustering solution provided from this process is neither stable nor intuitive, which may also be very sensitive to data outliers. To tackle this difficulty, Nonnegative Laplacian Embedding (NLE) method [Luo *et al.*, 2009] was proposed by additionally imposing the nonnegative constraints on the embeddings.

Despite the fact that the nonnegativity can be achieved in the NLE method, there are still some difficulties of this model that are not well addressed. It has been noted that the NLE method imposes the nonnegative constraint at the price of relaxing the orthogonality on the learned approximations [Ding *et al.*, 2006], although the orthogonality constraint ($\mathbf{X}^T\mathbf{X} = \mathbf{I}$) is of significant importance to guarantee a good performance. The true meaning of the orthogonality constraint is to prevent degenerate solution ($\mathbf{X} \to 0$). For one dimensional problem, the orthogonality can avoid that the embedded data collapse into a point. For multi-dimensional problem, the orthogonality can prevent data points from collapsing into a subspace with dimensions less than desired.

In this paper, we propose a new approach to learn LE with strictly guaranteed orthogonality and nonnegativity in the solution. Unlike using the auxiliary function method [Lee and Seung, 2001] to derive the solution algorithm for NLE in [Luo *et al.*, 2009], the orthogonality of our solution is rigorously achieved by using the Alternating Direction Method of Multipliers (ADMM) [Bertsekas, 1996; Boyd *et al.*, 2011], leading to a more stable solution and a better performance in the problem of spectral clustering. We also keep the nonnegativity in the constraint, such that the clustering membership can be readily read off from the embedded data due to the nonnegative constraint, *i.e.*, we can

[*]Equal contribution.

consider each row of the solution $\mathbf{X}$ as the posterior clustering probability. In other words, the values in i-th row of the solution can be viewed as the likelihoods that the i-th data point belongs to different clusters, which gives our new approach the soft clustering capability that is crucial in many real-world applications.

Finally, we recognize that the squared ℓ_2-norm distance used in the traditional LE and NLE objectives does not guarantee the optimal embedding [Wang *et al.*, 2015] and is also notoriously known to be sensitive to the outliers [Wang *et al.*, 2012; Nie *et al.*, 2013; Wang *et al.*, 2013c; Nie *et al.*, 2016; Liu *et al.*, 2017]. With strict orthogonality and nonnegativity guaranteed simultaneously in the solution, we are also interested in promoting the robustness of our new NLE model by using the p-th order ($0 < p \leq 2$) of the ℓ_2-norm distance in the objective. As a result, the proposed optimization objective is a quadratic function with both orthonormal and nonnegative constraints, which is highly nonlinear and nonconvex in its feasible domain. The p-th order term further makes the objective nonsmooth and difficult to efficiently optimize in general. To solve this challenging optimization problem, we propose a novel *smoothed iterative reweighted method*. Compared to the iterative reweighted method proposed in [Candes *et al.*, 2008; Nie *et al.*, 2010] to solve the ℓ_1-norm or $\ell_{2,1}$-norm minimization problems, our new optimization framework explicitly adds a smoothness term which can improve numerical stability. Most importantly, as an important theoretical contribution, we rigorously prove the convergence of our new iterative algorithm with the smoothness term, which, though, was not present in [Candes *et al.*, 2008; Nie *et al.*, 2010] and their following works.

To evaluate the proposed robust NLE objective that uses the p-th order of the ℓ_2-norm distances and our new smoothed iterative reweighted method, we have performed extensive empirical studies. The promising experimental results have validated the effectiveness of our new methods.

2 Strictly Orthogonal p-Order Nonnegative Laplacian Embedding

Given a set of n data points, we can represent the pairwise similarities between these data points by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the data points are represented by the vertices $\mathcal{V}$ and $|\mathcal{V}| = n$. Suppose that $\mathbf{W} \in \Re^{n \times n}$ denotes the affinity matrix of the graph $\mathcal{G}$ where w_{ij} measures the similarity between the i-th and the j-th vertices, quadratic placement [Hall, 1970] aims to embed the vertices of the graph into the one-dimensional space with coordinates $(x_1, \ldots, x_n)$, such that if the i-th and the j-th vertices are similar (*i.e.*, w_{ij} is large), they should be adjacent in embedded space, *i.e.*, $(x_i - x_j)^2$ should be small. This can be achieved by the following objective [Hall, 1970]:

$$\min_{\|\mathbf{x}\|_2^2=1} \sum_{i,j} w_{ij} (x_i - x_j)^2 = 2\mathbf{x}^T (\mathbf{D} - \mathbf{W}) \mathbf{x} \ , \quad (1)$$

where $\mathbf{x} = [x_1, \ldots, x_n]^T$, and $\mathbf{D} = diag(d_1, d_2, \ldots, d_n)$ is the degree matrix of the graph with $d_i = \sum_j w_{ij}$.

The one-dimensional quadratic placement in Eq. (1) can be generalized to r-dimensional LE by minimizing the following objective [Luo *et al.*, 2009]:

$$\min_{\mathbf{X}^T\mathbf{X}=\mathbf{I}} \sum_{i,j} w_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \mathbf{tr}\left(\mathbf{X}^T (\mathbf{D} - \mathbf{W}) \mathbf{X}\right) \ , \quad (2)$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \in \Re^{n \times r}$. Obviously, the i-th row of $\mathbf{X}$, *i.e.*, $\mathbf{x}_i^T \in \Re^r$, is the embedding of the i-th data point in the r-dimensional space. Here, the orthonormal constraint of $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ is imposed in Eq. (2) to avoid degenerate solutions, which is critical as analyzed in [Luo *et al.*, 2009; Ding *et al.*, 2006].

To decode the clustering membership from $\mathbf{X}$ in an easier way, Luo *et al.* [Luo *et al.*, 2009] further developed LE by additionally imposing the nonnegative constraint onto the embedding matrix $\mathbf{X}$ by minimizing the following objective:

$$\min_{\mathbf{X}} \ \mathbf{tr}\left(\mathbf{X}^T (\mathbf{D} - \mathbf{W}) \mathbf{X}\right), \quad s.t.\ \mathbf{X} \geq 0, \mathbf{X}^T\mathbf{X} = \mathbf{I} \ . \quad (3)$$

The squared ℓ_2-norm distances used in the both objectives in Eq. (2) and Eq. (3) do not tolerate large value of distance, thus making the distances in the embedded space tend to be even, *i.e.*, not too large but also not too small. Therefore, solving the objective in Eq. (2) or Eq. (3) may not find the optimal embedding such that most of the distances of local data pairs are minimized but a few of them are large [Wang *et al.*, 2015]. Motivated by recent papers that use not-squared ℓ_2-norm distances [Wang *et al.*, 2012; Nie *et al.*, 2013; Wang *et al.*, 2013c; Wang *et al.*, 2014; Nie *et al.*, 2016; Liu *et al.*, 2017] or the p-th order of the ℓ_2-norm distances [Wang *et al.*, 2011; Wang *et al.*, 2013a; Wang *et al.*, 2015] to promote the robustness of learning models, in this paper we propose to solve the following problem to find the optimal spectral embedding from an input graph:

$$\min_{\mathbf{X}} \sum_{i,j} w_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^p \ , \quad s.t.\ \mathbf{X} \geq 0, \mathbf{X}^T\mathbf{X} = \mathbf{I} \ , \quad (4)$$

where $0 < p \leq 2$. Obviously, the NLE method in Eq. (3) [Luo *et al.*, 2009] is a special case of our new method when $p = 2$. More importantly, by setting $p \leq 1$, the method will focus on minimizing most of the distances of local data pairs. Thus, we call Eq. (4) as the proposed strictly orthogonal p-Order Nonnegative Laplacian Embedding (PO-NLE) method.

3 Smoothed Iterative Reweighted Method and its Convergence

Although the motivation of our new objective in Eq. (4) is clear, it is nonsmooth and difficult to efficiently solve in general. To solve this challenging optimization problem, in this section we will first introduce a novel *smoothed iterative reweighted method*.

First, let us consider a general problem as follows:

$$\min_{x \in \mathcal{C}} \ f(x) + \sum_i tr\left(\left(g_i^T(x) g_i(x)\right)^{\frac{p}{2}}\right) \ . \quad (5)$$

When $g_i(x)$ is a vector output function, $tr\left(\left(g_i^T(x) g_i(x)\right)^{\frac{p}{2}}\right)$ becomes the following term:

$$tr\left(\left(g_i^T(x) g_i(x)\right)^{\frac{p}{2}}\right) = \|g_i(x)\|_2^p \ . \quad (6)$$

Equation (5) is nonsmooth, thus we turn to solve the following smooth problem [Liu *et al.*, 2017]:

$$\min_{x\in C} f(x) + \sum_i tr\left((g_i^T(x)g_i(x) + \delta\mathbf{I})^{\frac{p}{2}}\right) , \quad (7)$$

where $\delta > 0$ is a small constant. When $\delta \to 0$, Eq. (7) is reduced to Eq. (5) since the following equation holds:

$$\lim_{\delta\to 0} tr\left((g_i^T(x)g_i(x) + \delta\mathbf{I})^{\frac{p}{2}}\right) = \|g_i(x)\|_2^p . \quad (8)$$

Before deriving the algorithm for optimizing the problem in Eq. (7), we need the following lemmas. First, according to the chain rule in calculus, we have:

Lemma 1 *Suppose $g(x)$ is a matrix output function, $h(x)$ is a scalar output function, x is a scalar, vector or matrix variable, then we have:*

$$\frac{\partial h(g(x))}{\partial x} = \frac{\sum_{i,j}\frac{\partial h(g(x))}{\partial g_{ij}(x)}\partial g_{ij}(x)}{\partial x} = \left(\frac{\partial h(g(x))}{\partial g(x)}\right)^T \frac{\partial g(x)}{\partial x} . \quad (9)$$

According to the chain rule in Lemma 1, we can easily derive the following two lemmas:

Lemma 2 *Suppose $g(x)$ is a scalar, vector or matrix output function, x is a scalar, vector or matrix variable, then we have:*

$$\begin{aligned}&\frac{\partial tr((g^T(x)g(x) + \delta\mathbf{I})^{\frac{p}{2}})}{\partial x}\\ =\ & p\,(g^T(x)g(x) + \delta\mathbf{I})^{\frac{p-2}{2}} g^T(x)\frac{\partial g(x)}{\partial x} .\end{aligned} \quad (10)$$

Lemma 3 *Suppose $g(x)$ is a scalar, vector or matrix output function, x is a scalar, vector or matrix variable, $\mathbf{D}$ is a constant and $\mathbf{D}$ is symmetrical if $\mathbf{D}$ is a matrix, then we have:*

$$\frac{\partial tr(g^T(x)g(x)\mathbf{D})}{\partial x} = 2\mathbf{D}g^T(x)\frac{\partial g(x)}{\partial x} . \quad (11)$$

Now we derive the algorithm to optimize the problem in Eq. (7). The Lagrangian function of the problem in Eq. (7) is:

$$\mathcal{L}(x,\lambda) = f(x) + \sum_i tr((g_i^T(x)g_i(x) + \delta\mathbf{I})^{\frac{p}{2}}) - r(x,\lambda) , \quad (12)$$

where $r(x,\lambda)$ is a Lagrangian term for the constraint $x \in \mathcal{C}$. By setting the derivative of Eq.(12) w.r.t. x to zero, we have

$$\begin{aligned}&\frac{\partial L(x,\lambda)}{\partial x}\\ &= f'(x) + \sum_i \frac{\partial tr((g_i^T(x)g_i(x) + \delta\mathbf{I})^{\frac{p}{2}})}{\partial x} - \frac{\partial r(x,\lambda)}{\partial x}\\ &= 0 .\end{aligned} \quad (13)$$

According to Lemma 2, Eq.(13) can be rewritten as

$$\begin{aligned}&f'(x) + \sum_i p(g_i^T(x)g_i(x) + \delta\mathbf{I})^{\frac{p-2}{2}} g_i^T(x)\frac{\partial g_i(x)}{\partial x}\\ &\quad - \frac{\partial r(x,\lambda)}{\partial x} = 0 .\end{aligned} \quad (14)$$

Algorithm 1 The algorithm to solve the problem (7)

Initialize $x \in \mathcal{C}$.
while not converge **do**
1. For each i, calculate

$$\mathbf{D}_i = \frac{p}{2}(g_i^T(x)g_i(x) + \delta\mathbf{I})^{\frac{p-2}{2}} , \quad (18)$$

2. Update x by solving the problem:

$$\min_{x\in\mathcal{C}} f(x) + \sum_i tr(g_i^T(x)g_i(x)\mathbf{D}_i) . \quad (19)$$

end while

If we can find a solution x that satisfies the Eq.(14), we usually find a local or global optimal solution to the problem in Eq. (7) according to the Karush-Kuhn-Tucker (KKT) conditions [Boyd and Vandenberghe, 2004]. However, directly finding a solution x that satisfies Eq.(14) is not a easy task. In this paper, following [Candes *et al.*, 2008; Nie *et al.*, 2010] we propose an iterative algorithm to find it using the following observation: if

$$\mathbf{D}_i = \frac{p}{2}(g_i^T(x)g_i(x) + \delta\mathbf{I})^{\frac{p-2}{2}} \quad (15)$$

is a constant, Eq.(14) is reduced to:

$$f'(x) + \sum_i 2\mathbf{D}_i g_i^T(x)\frac{\partial g_i(x)}{\partial x} - \frac{\partial r(x,\lambda)}{\partial x} = 0 , \quad (16)$$

which is equivalent to solving the following problem:

$$\min_{x\in\mathcal{C}} f(x) + \sum_i tr(g_i^T(x)g_i(x)\mathbf{D}_i) . \quad (17)$$

Based on the observation above, we can first guess a solution x. Then we calculate $\mathbf{D}_i$ using the current solution of x and update x by the optimal solution of the problem in Eq. (17) by calculating $\mathbf{D}_i$. We iteratively perform this procedure until it converges, which is summarized in Algorithm 1.

The convergence of Algorithm 1 is guaranteed by the following theorem. The proof of Theorem 1 will be supplied in the extended journal version of this paper due to space limit.

Theorem 1 *The Algorithm 1 will monotonically decrease the objective of the problem (7) in each iteration until the algorithm converges.*

In the convergence, the equality in Eq. (14) will hold, thus the KKT condition of problem (7) is satisfied. Therefore, the Algorithm 1 will converge to a local optimum solution to the problem (7). If the problem (7) is convex, the Algorithm 1 will converge to a global optimum solution.

Here we **note** that the iterative reweighted method introduced in [Candes *et al.*, 2008; Nie *et al.*, 2010] solves the nonsmooth ℓ_1-norm or $\ell_{2,1}$-norm minimization problems. However, the method described in [Candes *et al.*, 2008; Nie *et al.*, 2010] does not explicitly use the smoothness constant (*i.e.*, $\delta\mathbf{I}$ in Eq. (7)). Without this smoothness term,

the algorithm is heavily impacted by the singularity problem due to inverted matrices that divide 0s, which routinely leads to inferior learning performances. To improve the numerical stability, in [Nie *et al.*, 2010] and the following works by the same group of authors [Wang *et al.*, 2013b; Nie *et al.*, 2013], a smoothness term was informally added for empirical purpose. But they only theoretically proved the convergence of the algorithm that does not use the smoothness term and did not provide any theoretical analysis on the objectives that use the smoothness term. As an important theoretical contribution of this paper, we formally introduce the smoothness term (*i.e.*, $\delta\mathbf{I}$ in Eq. (7)) into our algorithm and theoretically prove the convergence of our algorithm in which the smoothness term leads to much more stable solutions. We call Algorithm 1 as the proposed *Smoothed Iterative Reweighted Method*, which can be broadly used to solve a variety of difficult machine learning problems that minimize the objectives using the p-th order of ℓ_2-norm distances, the p-th order of ℓ_p-norm distances, or the p-th order of the Schatten p-norm.

4 Algorithm to Solve the Problem in Eq. (4)

Equipped with Algorithm 1, we can derive the solution algorithm to our new PO-NLE objective in Eq. (4) now. According to Step 2 of Algorithm 1, the key step to solve Eq. (4) is to solve the following problem:

$$\min_{\mathbf{X}} \sum_{i,j} w_{ij} d_{ij} \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 \ , \ s.t.\ \mathbf{X} \geq 0, \mathbf{X}^T\mathbf{X} = \mathbf{I} \ , \tag{20}$$

where $d_{ij} = \frac{p}{2}\left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 + \delta\right)^{\frac{p-2}{2}}$ and $\delta \to 0$.

Denote $\widetilde{\mathbf{W}}_{ij} = w_{ij}d_{ij}$ and let $\widetilde{\mathbf{D}}$ be the diagonal matrix with the i-th diagonal entry as $\sum_j \tilde{w}_{ij} = w_{ij}d_{ij}$. The problem in Eq. (20) can be written as following:

$$\begin{aligned}\min_{\mathbf{X}} \sum_{i,j} \tilde{w}_{ij} \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 = \mathbf{tr}\left(\mathbf{X}^T\left(\widetilde{\mathbf{D}} - \widetilde{\mathbf{W}}\right)\mathbf{X}\right) \ , \\ s.t.\ \mathbf{X} \geq 0, \mathbf{X}^T\mathbf{X} = \mathbf{I} \ .\end{aligned} \tag{21}$$

Obviously, Eq. (21) is identical to the NLE objective in Eq. (3), which was proposed in [Luo *et al.*, 2009]. In [Luo *et al.*, 2009], a solution algorithm was derived using the auxiliary function method [Lee and Seung, 2001]. However, as analyzed in [Ding *et al.*, 2005; Ding *et al.*, 2006] the orthogonal constraint indeed are not guaranteed, which, though, is very important to avoid degenerate solutions [Ding *et al.*, 2006]. Thus, instead of using the solution algorithm provided in [Luo *et al.*, 2009], we derive the solution algorithm to solve Eq. (21) using the ADMM method.

Denoting $\mathbf{L} = \widetilde{\mathbf{D}} - \widetilde{\mathbf{W}}$ for brevity[1], we can write the objective in Eq. (21) as following:

$$\min_{\mathbf{X}} \mathbf{tr}\left(\mathbf{X}^T\mathbf{L}\mathbf{X}\right), \quad s.t.\ \mathbf{X}^T\mathbf{X} = \mathbf{I}, \mathbf{X} \geq 0 \ . \tag{22}$$

[1]In practice, due to the zero mode of the Laplacian matrix of a graph [Wang *et al.*, 2010], we compute $\mathbf{L} = \widetilde{\mathbf{D}} - \widetilde{\mathbf{W}} + \frac{\widetilde{\mathbf{W}}_{++}}{n^2}\mathbf{e}^T\mathbf{e}$ to ensure that $\mathbf{L}$ is positive definite, where $\widetilde{\mathbf{W}}_{++} = \sum_{i,j}\widetilde{\mathbf{W}}$ and $\mathbf{e}$ is the vector with all entries to be 1.

We can solve Eq. (22) by solving the following equivalent optimization problem:

$$\min_{\mathbf{X},\mathbf{Y}} \mathbf{tr}\left(\mathbf{Y}^T\mathbf{L}\mathbf{X}\right), \quad s.t.\ \mathbf{Y} = \mathbf{X}, \mathbf{Y}^T\mathbf{Y} = \mathbf{I}, \mathbf{X} \geq 0 \ , \tag{23}$$

where the constraint of $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ in Eq. (22) is implicitly enforced by the constraints of $\mathbf{Y} = \mathbf{X}$ and $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}$.

Following the ADMM optimization framework, we need to solve the following optimization problem:

$$\begin{aligned}\min_{\mathbf{X},\mathbf{Y},\mathbf{\Lambda}} \mathbf{tr}\left(\mathbf{Y}^T\mathbf{L}\mathbf{X}\right) + \frac{\mu}{2}\left\|\mathbf{Y} - \mathbf{X} + \frac{1}{\mu}\mathbf{\Lambda}\right\|_{\mathrm{F}}^2, \\ s.t. \quad \mathbf{Y}^T\mathbf{Y} = \mathbf{I}, \mathbf{X} \geq 0 \ ,\end{aligned} \tag{24}$$

in which we introduced the Lagrangian multiplier $\mathbf{\Lambda}$ for the constraint of $\mathbf{Y} = \mathbf{X}$. The detailed procedures to solve Eq. (24) using the ADMM method will be supplied in the extended journal version of this paper due to space limit.

5 Experiment and Results

In this section we empirically evaluate our new PO-NLE method on one synthetic data set, four data sets from the UCI Machine Learning Data Repository, and three image data sets. We will compare our new method against its counterparts: NLE, Normalized Cut (NCut) [Shi and Malik, 2000] and Laplacian Embedding (LE).

In our evaluations, we use clustering accuracy and clustering purity to measure the performance of the compared methods. We also study the robustness of our method on the real world data sets when they are contaminated with noise. The performance variations when we increase the value of p will be shown to validate our hypothesis that the optimal solution is usually obtained when p is less than 2 and close to 1 (it depends on data sets), given that noises are present in the data. Orthogonality of the solution will be illustrated and compared against the NLE method in [Luo *et al.*, 2009].

5.1 Experiments on a Synthetic Data Set

To illustrate the effectiveness of our new PO-NLE method, we create a synthetic data set as follows. We first randomly generate 3 data points as centroids in the 30-dimensional space. Then we generate 3 groups of data points and each group consists of 39 data points which are randomly distributed around one of the three centroids. A threshold is set to make the distance of groups large enough. As shown in Figure 1, different colors (red, black and blue) and shapes are used to represent different groups of data points. We randomly initialize $\mathbf{X}$ ($0 \leq \mathbf{X} \leq 1$) and set $\rho = 1.02$, $\mu = 0.1$ and $p = 0.8$ in our algorithm. we use K-Nearest Neighbors with heat kernel to construct our adjacency matrix $\mathbf{W}$. The variation of the objective value when our algorithm iterates are shown as the red curve in Figure 1. For visualization purpose, we set $r = 3$, *i.e.*, we embed the original data into the 3-dimensional space using our new PO-NLE algorithm. The x, y and z axes of the 3D plots in the figure correspond to the first, second and third row in matrix $\mathbf{X}$, respectively.

From Figure 1, we observe that the objective function monotonically decreases in each iteration, which empirically

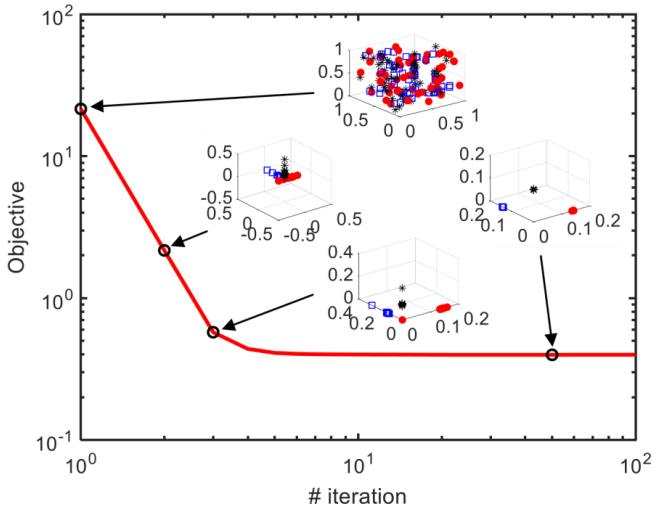


Figure 1: The objective function of our PO-NLE method on the synthetic data with the result of 3D plots illustrating the clustering structure on checkpoints. The x, y and z axis of the 3D plots correspond to the first, second and third column in matrix $\mathbf{X}$, respectively.

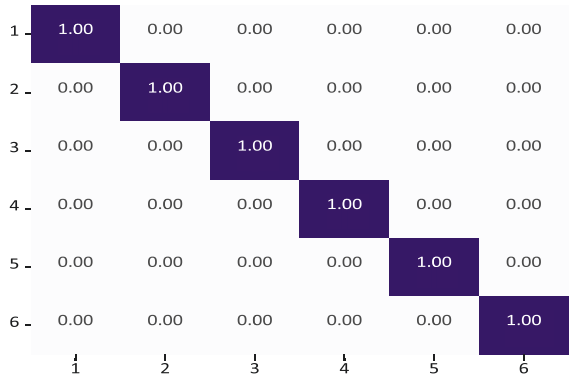
confirms the convergence of the solution algorithm to solve PO-NLE derived by our new smoothed iterative reweighted method. Moreover, for each checkpoint shown by the black circles on the objective curve, the clustering structure of the experimental data becomes more and more clear in the 3D plots when the algorithm iterates. The three clusters of data points gradually find a solution to separate themselves apart and fall on different axes. Note that, due to the nonnegative constraints on $\mathbf{X}$, data points will finally converge on the positive part of each axis. This observation clearly demonstrate the effectiveness of the proposed new method.

5.2 Studies of the Orthogonality of the Solutions of Our New Method

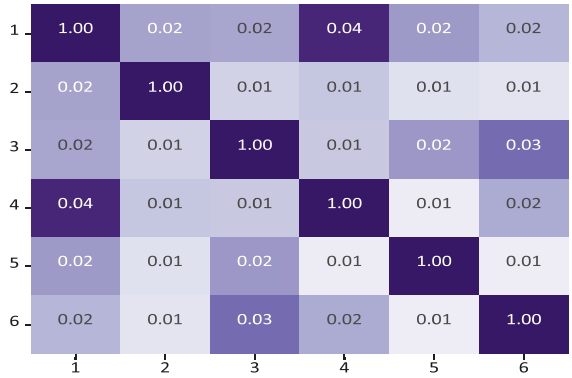
An important improvement of our new method over the NLE method [Luo *et al.*, 2009] is that the orthogonality of our solution is rigorously guaranteed, which, as analyzed in [Ding *et al.*, 2005; Ding *et al.*, 2006] is very important to avoid degenerate solutions. Thus, in this subsection we empirically study the orthogonality of the solutions of our new method and compare them against the solutions obtained from the NLE method. Figure 2 compare the visualizations of $\mathbf{X}^T\mathbf{X}$ learned from the two compared methods on the Glass data set by the heatmaps. The heatmap of our new method is on the top and that of the NLE method is at the bottom. From the results we can see that the learned embeddings from our method are strictly orthogonal as shown in Figure 2(a), which will in return lead to better clustering performances and robustness after embedding. In contrast, the NLE method failed to guarantee the orthogonality, as can be seen in Figure 2(b).

5.3 Experiments on Noiseless Real Data Sets

Now we compare our new method, NLE, NCut and LE on the seven standard data sets as summarized in Table 1. Each data set will be tested by different algorithms independently for 200 times. For NCut and LE algorithms, we run K-means



(a) Visualization of $\mathbf{X}^T\mathbf{X}$ learned by our method.



(b) Visualization of $\mathbf{X}^T\mathbf{X}$ learned by the NLE method.

Figure 2: The comparison of orthogonality between our method and NLE on glass data set.

clustering with random initialization for 50 times and report the best results.

The performances of the compared methods are reported in the top half of Table 2, from which we can see that our method clearly outperforms all other competing methods, especially on those comparatively noisier data sets. Due to the nonnegative solutions of our new method, we do not need any additional clustering step. Instead, the clustering membership can be readily read off directly from the learned embeddings. The strictly guaranteed orthogonality constraint avoids degenerate solution and helps improve the performance com-

Dataset	# Size	# Dimension	# Class
MINIST	5000	784	10
AT&T	400	10304	40
Caltech101	332	900	5
Ionosphere	351	34	2
Wine	178	13	3
Iris	150	4	3
Glass	214	9	6

Table 1: Dataset descriptions.

Data sets	Clustering accuracy↑								Clustering Purity↑							
	Ours		NLE		NCut		LE		Ours		NLE		NCut		LE	
	Best	Ave	Best	Ave	Best	Ave	Best	Ave	Best	Ave	Best	Ave	Best	Ave	Best	Ave
MINIST	**0.7450**	**0.5946**	0.6700	0.5013	0.6540	0.5909	0.6630	0.5874	**0.7880**	**0.6811**	0.7140	0.5553	0.6597	0.4933	0.6438	0.5121
AT&T	**0.8250**	**0.7719**	0.7825	0.6885	0.7525	0.6383	0.7325	0.6882	**0.8675**	**0.8304**	0.8325	0.7476	0.7325	0.6782	0.7250	0.6487
Caltech101	**0.8342**	**0.7719**	0.7131	0.5131	0.5972	0.4965	0.6663	0.5957	**0.9644**	**0.8719**	0.8383	0.5512	0.5943	0.5625	0.6612	0.5768
Ionosphere	**0.8604**	**0.8065**	0.8120	0.6267	0.7236	0.6449	0.7493	0.6475	**0.9658**	**0.8129**	0.8462	0.7017	0.7175	0.6255	0.7413	0.6580
Wine	**0.7303**	**0.7088**	0.7022	0.6464	0.6910	0.6686	0.6685	0.6585	**0.8034**	**0.7092**	0.7753	0.6698	0.7058	0.6497	0.7702	0.6259
Iris	**0.9667**	**0.8945**	0.9600	0.7591	0.9067	0.7824	0.9000	0.7144	**0.9600**	**0.9045**	0.9600	0.7962	0.9067	0.8080	0.9000	0.7340
Glass	**0.5888**	**0.4646**	0.5748	0.4451	0.4439	0.3703	0.5093	0.4102	**0.7710**	**0.6384**	0.5935	0.4832	0.6176	0.5037	0.5335	0.4512
MINIST	**0.5460**	**0.4458**	0.4590	0.3625	0.4210	0.3795	0.4430	0.3928	**0.6520**	**0.5879**	0.5550	0.4369	0.5230	0.4736	0.5070	0.4661
AT&T	**0.7275**	**0.6630**	0.6800	0.5797	0.6575	0.5718	0.6550	0.5606	**0.7800**	**0.7408**	0.7175	0.6529	0.7075	0.6408	0.6975	0.6324
Caltech101	**0.8199**	**0.7681**	0.5839	0.4291	0.5524	0.4177	0.5025	0.4160	**0.8993**	**0.8260**	0.5839	0.4465	0.5573	0.4312	0.5036	0.4285
Ionosphere	**0.7692**	**0.5923**	0.6211	0.5250	0.5755	0.5249	0.5783	0.5270	**0.8889**	**0.7123**	0.6279	0.5305	0.5795	0.5311	0.5848	0.5318
Wine	**0.6292**	**0.5077**	0.5506	0.4318	0.5787	0.4308	0.5730	0.4237	**0.6461**	**0.5537**	0.5506	0.4450	0.5347	0.4400	0.6067	0.4346
Iris	**0.7867**	**0.6679**	0.6733	0.4958	0.6200	0.4689	0.6467	0.4755	**0.8667**	**0.7078**	0.6733	0.5183	0.6800	0.4934	0.6333	0.4953
Glass	**0.5421**	**0.4586**	0.4159	0.3249	0.4299	0.3305	0.3738	0.2914	**0.7383**	**0.6165**	0.4486	0.3565	0.4626	0.3592	0.3832	0.3171

Table 2: Best and average (Ave) clustering accuracy and purity by our method, NLE, NCut and LE over 200 trials. "↑" means that the bigger number are the better. **Top:** the results on noiseless data (Section 5.3); **bottom:** the results on noisy data (Section 5.4).

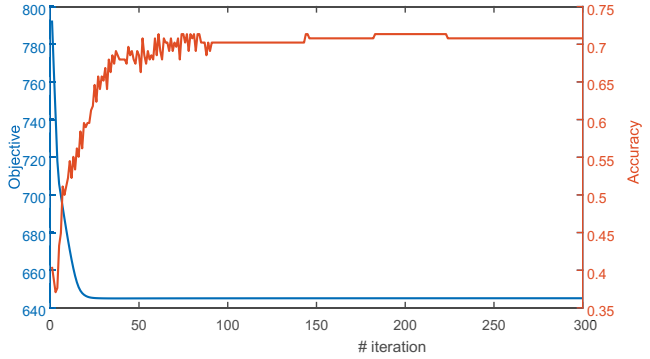
pared with loosely constrained NLE method which does not have such desirable property. To illustrate the convergence of the objective function of our new method, Figure 3(a) and Figure 3(b) show a typical run of our algorithm on two UCI benchmark data sets. As can be seen from the figures, when the algorithm iterates and the objective value decreases, the accuracy shows a relatively smoothly increasing line.

5.4 Experiments on Noisy Real Data Sets

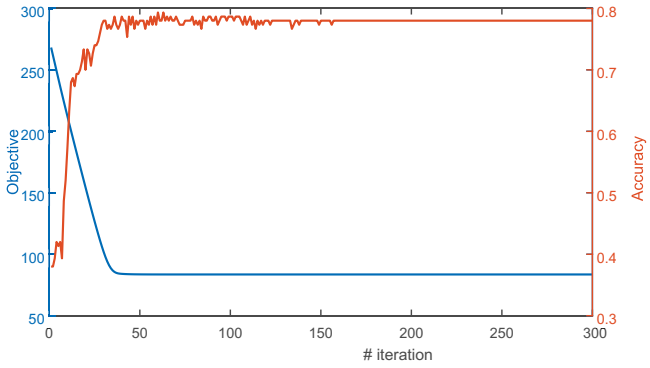
To study the impacts of the value of p in our new embedding model, we randomly contaminate 20% of the data points in all 7 data sets and we run our method with increasing p on those data sets. For each p, we run 200 times for the same contaminated data and original data respectively. Other algorithms are also tested for 200 times on each data set for comparison. The performances of the clustering methods on contaminated noisy data sets are reported in the bottom half of Table 2. Among all the best and the average values of clustering accuracy and clustering purity, our method is consistently better than its counterparts. The results of our approach generally decreases less than other methods on the contaminated data sets, especially for those noisier data sets.

6 Conclusion

In this paper, we proposed a new robust Laplacian embedding approach that uses the p-th order of the ℓ_2-norm distances in the objective and strictly satisfies both orthogonality and nonnegativity constraints at the same time. This results in an objective that is neither convex nor smooth, which is difficult to efficiently solve in general. We thereby proposed a novel *smoothed iterative reweighted method* to solve this challenging optimization problem, in which a smoothness term is formally and explicitly introduced for improved numerical stability. As an important theoretical contribution of this paper, we rigorously we proved the convergence of our new algorithm with the smoothness term. Using this new and improved optimization framework, our objective can be elegantly solved. We have performed extensive experiments, in which the superior performance of our new method has demonstrated its effectiveness and the potential to give a new perspective for nonlinear graph based clustering tasks.



(a) Wine data set.



(b) Iris data set.

Figure 3: A typical run of our algorithm on two data sets with iteration ranging from 1 to 300 to illustrate the convergence of objective function and accuracy of the clustering result.

Acknowledgments

All correspondence should be addressed to: Hua Wang (huawangcs@gmail.com). This work was partially supported by National Science Foundation under Grants IIS-1652943 and IIS-1849359.

References

[Bertsekas, 1996] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 1996.

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[Candes *et al.*, 2008] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted l-1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[Chan *et al.*, 1994] Pak K Chan, Martine DF Schlag, and Jason Y Zien. Spectral k-way ratio-cut partitioning and clustering. *IEEE Transactions on computer-aided design of integrated circuits and systems*, 13(9):1088–1096, 1994.

[Ding and He, 2004] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.

[Ding *et al.*, 2005] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610. SIAM, 2005.

[Ding *et al.*, 2006] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006.

[Hagen and Kahng, 1992] Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE transactions on computer-aided design of integrated circuits and systems*, 11(9):1074–1085, 1992.

[Hall, 1970] Kenneth M Hall. An r-dimensional quadratic placement algorithm. *Management science*, 17(3):219–229, 1970.

[Hartigan and Wong, 1979] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[Lee and Seung, 2001] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[Liu *et al.*, 2017] Yun Liu, Yiming Guo, Hua Wang, Feiping Nie, and Heng Huang. Semi-supervised classifications via elastic and robust embedding. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, 2017.

[Luo *et al.*, 2009] Dijun Luo, Chris Ding, Heng Huang, and Tao Li. Non-negative laplacian embedding. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 337–346. IEEE, 2009.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[Nie *et al.*, 2013] Feiping Nie, Hua Wang, Heng Huang, and Chris HQ Ding. Early active learning via robust representation and structured sparsity. In *IJCAI*, pages 1572–1578, 2013.

[Nie *et al.*, 2016] Feiping Nie, Hua Wang, Cheng Deng, Xinbo Gao, Xuelong Li, Heng Huang, et al. New ℓ_1-norm relaxations and optimizations for graph clustering. In *AAAI*, pages 1962–1968, 2016.

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

[Wang *et al.*, 2010] Hua Wang, Chris Ding, and Heng Huang. Directed graph learning via high-order co-linkage analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 451–466. Springer, 2010.

[Wang *et al.*, 2011] Hua Wang, Feiping Nie, and Heng Huang. Learning instance specific distance for multi-instance classification. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[Wang *et al.*, 2012] Hua Wang, Feiping Nie, and Heng Huang. Robust and discriminative distance for multi-instance learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2919–2924. IEEE, 2012.

[Wang *et al.*, 2013a] Hua Wang, Feiping Nie, Weidong Cai, and Heng Huang. Semi-supervised robust dictionary learning via efficient $\ell_{2,0^+}$-norms minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1145–1152, 2013.

[Wang *et al.*, 2013b] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *International conference on machine learning*, pages 352–360, 2013.

[Wang *et al.*, 2013c] Hua Wang, Feiping Nie, and Heng Huang. Robust and discriminative self-taught learning. In *International conference on machine learning*, pages 298–306, 2013.

[Wang *et al.*, 2014] Hua Wang, Feiping Nie, and Heng Huang. Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization. In *International Conference on Machine Learning*, pages 1836–1844, 2014.

[Wang *et al.*, 2015] Hua Wang, Feiping Nie, and Heng Huang. Learning robust locality preserving projection via p-order minimization. In *AAAI*, pages 3059–3065, 2015.