Learning Robust Simplex Sparse Representation Using Alternating Linearized Minimization Method

Zhennan Shi
Department of Computer Science,
Colorado School of Mines
Golden, CO, USA
zhennanshi@mines.edu

Qiushi Wei
Department of Computer Science,
Colorado School of Mines
Golden, CO, USA
qiushiwei@mines.edu

***Abstract*—The construction of graphical representations is crucial for dealing with high-dimensional data in various fields. However, creating such representations often encounters challenges of conventional methods. In this work, we present a novel methodology of a Robust Simplex Sparse Representation (RSSR). Additionally, we address the optimization of the standard simplex by introducing an effective iterative algorithm known as the Proximal Alternating Linearized Minimization method. Our algorithm achieves global convergence and convergence of the objective function and results in a notable reduction in computational expenses while bolstering resilience to outliers and noise. Extensive experiments on both synthetic and real-world datasets validate our method's effectiveness and advantages over state-of-the-art clustering techniques. Our research offers novel insights into data representation and establishes a foundation for more precise analyses of big data.**

***Index Terms*—Graphical representation, Proximal Alternating Linearized Minimization, clustering.**

I. Introduction

In the ever-evolving landscape of machine learning, one fundamental mathematical discipline emerges as a graphical representation [1]. While machine learning has made significant strides in modeling data, it is often closely related to the data itself. Hence, we focus on datasets that can be readily transformed into similar graphical representations using straightforward methods. Numerous techniques for constructing graphical representations have been developed, such as k-Nearest Neighbors (k-NN) [2], [3], Graph Minimum Spanning Tree (MST) [4], Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), ϵ-neighborhood Graph [5], and Fully Connected Graph Self-Tuning Graph Construction [6]. Nevertheless, these methods still have problems. For example, PCA results in information loss, especially when the reduced dimensions do not adequately capture the variability in the original data [7]. The ϵ-neighbor graph method needs to find the appropriate ϵ to establish the edge connections, and the k-nearest neighbor graph relies heavily on k to connect neighbors.

Many existing sparse representation methods, such as cosine similarity, Euclidean distance, and kernel functions, are used to calculate the affinity (similarity or weight) between data points based on their sparse representations [8], [9]. However, constructing sparse graphical representations still faces huge computational overhead for large datasets and may overlook the global structure of the data. In this work, we focus on Sparse Representation Graphs [10], [11], which is one of the most effective graph representation methods renowned for its resilience to noise and outliers. Importantly, this approach captures similarities among high-dimensional data without requiring scale consistency.

This paper makes a significant contribution to solving a non-convex and non-smooth objective function. To address this challenge, we introduce an effective solution update algorithm, known as the Proximal Alternating Linearized Minimization method [12], [13]. Through rigorous experimentation, our proposed Robust Simplex Sparse Representation (RSSR) surpasses several competing methods and exhibits a significant advantage in handling diverse datasets. The contributions of our work can be summarized as follows:

1) The RSSR inherits the advantages of sparse representation and exhibits robustness to issues related to scale inconsistency and outlier noise.
2) The proximal alternating linearized minimization method is tailored for the optimization of the objective function, which ensures the convergence of both the objective and the sequence and easily can find the globally optimal solution.
3) Our approach integrates a simplex constraint into sparse representation, guaranteeing shift invariance and fostering more compact representations. This contributes to improved accuracy in data interpretation and enhances computational efficiency.
4) Extensive experiments comprise two aspects: synthetic datasets and real-world datasets to highlight the remarkable clustering effectiveness of RSSR.

II. Robust Simplex Sparse Representation

In this section, we introduce the background of the RSSR.

A. Background of Sparse Representation

Given a dataset consisting of n vectors with a dimension of d features, it can be arranged into columns of a training sample matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$. Given a new observation $\mathbf{y}$ which should satisfy $\mathbf{y} \approx \sum_{i=1}^{n} \mathbf{x}_i \alpha_i = \mathbf{X}\boldsymbol{\alpha}$, a representation vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$ can be thus computed. In

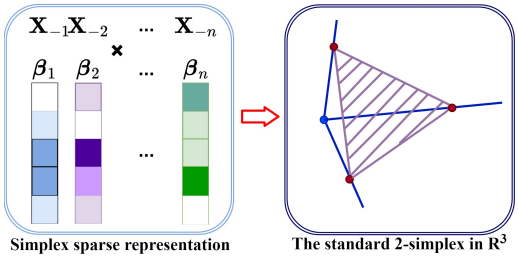

Fig. 1. Framework of the Robust Simplex Sparse Representation Method

order to obtain a sparse solution for $\boldsymbol{\alpha}$, it is equivalent to solving the following optimization problem:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{X}\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \lambda_0 \|\boldsymbol{\alpha}\|_0, \tag{1}$$

where the pseudo-norm ℓ_0 quantifies the number of non-zero elements in $\boldsymbol{\alpha}$, and $\lambda_0 > 0$ serves as a control parameter. The sparse solution in Eq. (1) can be effectively approximated by solving the ℓ_1 minimization problem. Remarkably, standard linear programming techniques enable the solution of ℓ_1 problem in polynomial time.

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1, \text{ s.t. } \mathbf{X}\boldsymbol{\alpha} = \mathbf{y} \text{ or } \min_{\boldsymbol{\alpha}} \|\mathbf{X}\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_1. \tag{2}$$

The sparse representation framework has gained recognition for its ability to handle noise and outliers effectively in image classification tasks. Importantly, it is worth mentioning that this framework does not impose any constraints on the scale consistency of the data vectors. In other words, sparse representation holds the potential to address the challenges of scale inconsistency and outliers highlighted in the introduction.

B. RSSR Construction

The proposed approach can be employed to compute the similarity matrix $\mathbf{S}$ in this context. Considering a data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where each sample is represented as a vector of dimension d. Specifically, the similarities $\boldsymbol{\beta}_i \in \mathbb{R}^{n-1}$ between the i-th sample and other samples are computed using sparse representation through the optimization problem:

$$\min_{\boldsymbol{\beta}_i} \|\mathbf{X}_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i\|_2^2 + \lambda \|\boldsymbol{\beta}_i\|_1. \tag{3}$$

Here, $\mathbf{X}_{-i} = [\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times (n-1)}$ represents the data matrix without the i column. To ensure the non-negativity of the similarity matrix, we can impose a non-negative constraint on Eq. (3) to minimize the sum of the residue error in sparse representation.

$$\sum_{i=1}^{n} \min_{\boldsymbol{\beta}_i \geq 0} \|\mathbf{X}_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i\|_2^2 + \lambda \|\boldsymbol{\beta}_i\|_1. \tag{4}$$

The parameter λ requires tuning in this approach. Note that Eq. (4) represents the sum of n independent variables. Therefore, for the purpose of our discussion, we will focus on a single term. Furthermore, when the data is shifted by constants $\mathbf{b} = [b_1, \cdots, b_d]^T$, i.e., $\mathbf{x}_k = \mathbf{x}_k + \mathbf{b}$ for any k, the computed similarities will be altered. To ensure shift-invariance in the similarities, we introduce the following equation:

$$\left\|\left(\mathbf{X}_{-i} + \mathbf{b}\mathbf{1}^T\right)\boldsymbol{\beta}_i - (\mathbf{x}_i + \mathbf{b})\right\|_2^2 = \|\mathbf{X}_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i\|_2^2. \tag{5}$$

Eq. (5) implies that $\boldsymbol{\beta}_i^T\mathbf{1} = 1$. By adding this constraint to problem Eq. (4), we obtain:

$$\begin{aligned} &\min_{\boldsymbol{\beta}_i} \|\mathbf{X}_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i\|_2^2 + \lambda \|\boldsymbol{\beta}_i\|_1. \\ &\text{s.t. } \boldsymbol{\beta}_i \geq 0, \boldsymbol{\beta}_i^T\mathbf{1} = 1 \end{aligned} \tag{6}$$

The constraints in problem Eq. (6) render the second term constant. As a result, problem Eq. (6) can be simplified to:

$$\begin{aligned} &\min_{\boldsymbol{\beta}_i} \|\mathbf{X}_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i\|_2^2. \\ &\text{s.t. } \boldsymbol{\beta}_i \geq 0, \boldsymbol{\beta}_i^T\mathbf{1} = 1 \end{aligned} \tag{7}$$

The constraints in problem Eq. (7) form a simplex, leading to the term "simplex representation" for this type of sparse representation. In this paper, we use the Alternating Minimization (AM) method to optimize Eq. (7). Details of applying AM method will be introduced in the rest sections.

III. Simplex Constrained Optimization

To solve the simplex constrained minimization problem, We can construct a Lagrange function, given by:

$$\mathcal{J}(\boldsymbol{\beta}_i, \lambda_i, \boldsymbol{\mu}_i) = \|\mathbf{X}_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i\|_2^2 - \lambda_i(\boldsymbol{\beta}_i^T\mathbf{1} - 1) - \boldsymbol{\mu}_i{}^T\boldsymbol{\beta}_i. \tag{8}$$

In Eq. (8), the new variable λ_i acts as a Lagrange multiplier for the constraint and $\boldsymbol{\mu}_i$ is a Lagrangian coefficient vector. Instead of minimizing $\mathcal{J}$ concerning all parameters at once, we can use the AM method to minimize $\mathcal{J}$ concerning one parameter at a time. This approach yields the following update rules when applied to the objective in Eq. (8):

$$\boldsymbol{\beta}_{i(k+1)} = \mathcal{J}(\boldsymbol{\beta}_i, \lambda_{i(k)}, \boldsymbol{\mu}_{i(k)}), \boldsymbol{\beta}_i \in \mathbf{B}_i \tag{9}$$

$$\lambda_{i(k+1)} = \mathcal{J}(\boldsymbol{\beta}_{i(k)}, \lambda_i, \boldsymbol{\mu}_{i(k)}), \lambda_i \in \boldsymbol{\Lambda}_i \tag{10}$$

$$\boldsymbol{\mu}_{i(k+1)} = \mathcal{J}(\boldsymbol{\beta}_{i(k)}, \lambda_{i(k)}, \boldsymbol{\mu}_i), \boldsymbol{\mu}_i \in \mathbf{U}_i \tag{11}$$

To address the problem in Eq. (9), Eq. (10), and Eq. (11) arising from the constraint on each sequence set, we employ the ALM [12], [13] method. This method has been used in similar non-convex and non-smooth optimization problems. To do this, we add a proximal regularization term to the updated equations, resulting in the following formulation:

$$\begin{aligned} \boldsymbol{\beta}_{i(k+1)} = &\arg\min \langle \boldsymbol{\beta}_i - \boldsymbol{\beta}_{i(k)}, \nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})\rangle \\ &+ \frac{\gamma_1}{2}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i(k)}\|_2^2, \end{aligned} \tag{12}$$

$$\begin{aligned} \lambda_{i(k+1)} = &\arg\min \langle \lambda_i - \lambda_{i(k)}, \nabla\mathcal{J}(\lambda_{i(k)})\rangle \\ &+ \frac{\gamma_2}{2}(\lambda_i - \lambda_{i(k)})^2, \end{aligned} \tag{13}$$

$$\begin{aligned} \boldsymbol{\mu}_{i(k+1)} = &\arg\min \langle \boldsymbol{\mu}_i - \boldsymbol{\mu}_{i(k)}, \nabla\mathcal{J}(\boldsymbol{\mu}_{i(k)})\rangle \\ &+ \frac{\gamma_3}{2}\|(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i(k)})\|_2^2. \end{aligned} \tag{14}$$

Here, $\nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})$ represents the gradient of $\mathcal{J}$ with respect to $\boldsymbol{\beta}_i$ at iteration k, while $\langle \boldsymbol{\beta}_i - \boldsymbol{\beta}_{i(k)}, \nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})\rangle$ refers to the dot product between the gradient $\nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})$ and the offset $\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i(k)}$. Additionally, the regularization terms introduced in the optimization problem help prevent the updated values at each iteration from changing too rapidly. Eq. (12) to Eq. (14) can be solved by setting its derivative to zero. Thus, we obtain the following equations:

$$\boldsymbol{\beta}_{i(k+1)} = \boldsymbol{\beta}_{i(k)} - \frac{1}{\gamma_1}(2\mathbf{X}_{-i}^T(\mathbf{X}_{-i}\boldsymbol{\beta}_{i(k)} - \mathbf{x}_i) - \lambda_i\mathbf{1} - \boldsymbol{\mu}_i), \tag{15}$$

$$\lambda_{i(k+1)} = \lambda_{i(k)} - \frac{1}{\gamma_2}(1 - \boldsymbol{\beta}_{i(k)}^T\mathbf{1}), \tag{16}$$

$$\boldsymbol{\mu}_{i(k+1)} = \boldsymbol{\mu}_{i(k)} - \frac{1}{\gamma_3}(-\boldsymbol{\beta}_i). \tag{17}$$

The pseudocode for updating the three regularization terms is shown in Algorithm 1 below. Here the convergence criteria

Algorithm 1: Alternating Linearized Minimization

Input: Data Matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, λ_i, γ_1, γ_2, γ_3, and K

Initialize $\boldsymbol{\beta}_i$;

while $k \leq$ K **do**

  Update $\boldsymbol{\beta}_{i(k)}$ as Eq. (15);

  Update $\lambda_{i(k)}$ as Eq. (16) ;

  Update $\boldsymbol{\mu}_{i(k)}$ as Eq. (17);

return

is that the relative change of $\|\boldsymbol{\beta}\|$, the Frobenius norm of $\boldsymbol{\beta}_i$, which indicates that $\|\boldsymbol{\beta}_{i(k+1)} - \boldsymbol{\beta}_{i(k)}\|_F$ is less than 10^{-4}. The empirical experiments conducted in this paper show that our algorithm converges fast and always ends within 10 iterations.

From the graphical representation we obtain, the similarity matrix can be further constructed. However, the similarity matrix $\mathbf{S} = [\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_n]$ is not necessarily symmetric. Thus, $\hat{\boldsymbol{\beta}}_i$ is the resulting vector of dimension n of inserting coefficient 0 in the i-th position of $\boldsymbol{\beta}_i$. The coefficient $\boldsymbol{\beta}_{ij}$ represents the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ for $\mathbf{X}_{-i}$, which does not have to be equal to $\boldsymbol{\beta}_{ji}$. To deal with this, the final similarity matrix is obtained by:$\mathbf{W} = \frac{1}{2}(\mathbf{S} + \mathbf{S}^T)$. The proof of the correctness of the final similarity matrix is omitted here.

IV. Convergence Analysis

In this section, we focus on analyzing the convergence of our solution algorithm. We start by establishing the Lipschitz continuity of the algorithm's gradient. This analysis provides the stability and convergence properties of the algorithm.

A. Lipschitz Continuous Gradient

We first show that $\mathcal{J}(\boldsymbol{\beta}_i, \lambda_i, \boldsymbol{\mu}_i)$ has a Lipschitz continuous gradient, which assists in later convergence analysis.

Theorem IV.1. *$\mathcal{J}(\boldsymbol{\beta}_i, \lambda_i, \boldsymbol{\mu}_i)$ has Lipschitz continuous gradient for all for all $\boldsymbol{\beta}_i, \boldsymbol{\beta}_{i(k)} \in \mathbf{B}_i$, $\lambda_i, \lambda_{i(k)} \in \boldsymbol{\Lambda_i}$ and $\boldsymbol{\mu}_i, \boldsymbol{\mu}_{i(k)} \in \mathbf{U}_i$. That is, there exists a constant $L_c \geq 0$ such that:*

$$\begin{aligned}&\|\nabla\mathcal{J}(\boldsymbol{\beta}_i, \lambda_i, \boldsymbol{\mu}_i) - \nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)}, \lambda_{i(k)}, \boldsymbol{\mu}_{i(k)})\|_F \\ &\leq L_c\|(\boldsymbol{\beta}_i, \lambda_i, \boldsymbol{\mu}_i) - (\boldsymbol{\beta}_{i(k)}, \lambda_{i(k)}, \boldsymbol{\mu}_{i(k)})\|_F.\end{aligned} \tag{18}$$

The $L_c \geq 0$ here is referred to as the Lipschitz constant, $\mathcal{J}(\boldsymbol{\beta}_i, \lambda_i, \boldsymbol{\mu}_i)$ is non-negative, convex and twice differentiable.

Proof. It is equivalent to showing $\|\nabla^2\mathcal{J}\|_F \leq L_c$ for all variables. The objective can be rewritten in the following forms applying the Taylor expansion:

$$\begin{aligned}&\mathcal{J}(\boldsymbol{\beta}_{i(k)}, \lambda_{i(k)}) = \mathcal{J}(\boldsymbol{\beta}_i, \lambda_i) + \Delta_{\boldsymbol{\beta}_i}\nabla\mathcal{J}(\boldsymbol{\beta}_i) \\ &+ \Delta_{\lambda_i}\nabla\mathcal{J}(\lambda_i) + \frac{1}{2}(\Delta_{\boldsymbol{\beta}_i}, \Delta_{\lambda_i})\nabla^2\mathcal{J}(\boldsymbol{\beta}_i, \lambda_i)(\Delta_{\boldsymbol{\beta}_i}, \Delta_{\lambda_i}),\end{aligned} \tag{19}$$

and

$$\begin{aligned}&\mathcal{J}(\boldsymbol{\beta}_{i(k)}, \boldsymbol{\mu}_{i(k)}) = \mathcal{J}(\boldsymbol{\beta}_i, \boldsymbol{\mu}_i) + \Delta_{\boldsymbol{\beta}_i}\nabla\mathcal{J}(\boldsymbol{\beta}_i) \\ &+ \Delta_{\boldsymbol{\mu}_i}\nabla\mathcal{J}(\boldsymbol{\mu}_i) + \frac{1}{2}(\Delta_{\boldsymbol{\beta}_i}, \Delta_{\boldsymbol{\mu}_i})\nabla^2\mathcal{J}(\boldsymbol{\beta}_i, \boldsymbol{\mu}_i)(\Delta_{\boldsymbol{\beta}_i}, \Delta_{\boldsymbol{\mu}_i}).\end{aligned} \tag{20}$$

Equation (19) and Equation (20) gives us:

$$\begin{aligned}&\|\nabla^2\mathcal{J}(\boldsymbol{\beta_i}, \lambda_i)\|_F \leq \max_{\|\Delta\|_F=1} |[\nabla^2\mathcal{J}(\boldsymbol{\beta_i}, \lambda_i)](\Delta, \Delta)| \leq \\ &|\mathcal{J}(\boldsymbol{\beta}_{i(k)}, \lambda_{i(k)}) - \mathcal{J}(\boldsymbol{\beta_i}, \lambda_i) - \Delta\boldsymbol{\beta}_i\nabla\mathcal{J}(\boldsymbol{\beta}_i) \\ &- \Delta\lambda_i\nabla\mathcal{J}(\lambda_i)|.\end{aligned} \tag{21}$$

This can be expanded as:

$$\begin{aligned}&\|\nabla^2\mathcal{J}(\boldsymbol{\beta_i}, \lambda_i)\|_F \leq| \|\mathbf{X}_{-i}(\boldsymbol{\beta}_i + \Delta\boldsymbol{\beta}_i - \mathbf{x}_i\|_2^2 - \lambda_i((\boldsymbol{\beta}_i + \\ &\Delta\boldsymbol{\beta}_i)^T\mathbf{1} - 1) - \|\mathbf{X}_{-i}\boldsymbol{\beta_i} - \mathbf{x}_i\|_2^2 - \lambda_i(\boldsymbol{\beta_i}^T\mathbf{1} - 1) \\ &- 2(\mathbf{X}_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i)\Delta\boldsymbol{\beta}_i + \lambda_i\mathbf{1}\Delta\boldsymbol{\beta}_i + (\boldsymbol{\beta}_i^T\mathbf{1} - 1)\Delta\lambda_i \\ &|:= \mathbf{C}.\end{aligned} \tag{22}$$

Similarly, Equation (20) can be extended as:

$$\begin{aligned}&\|\nabla^2\mathcal{J}(\boldsymbol{\beta_i}, \boldsymbol{\mu}_i)\|_F \leq| \|\mathbf{X}_{-i}(\boldsymbol{\beta}_i + \Delta\boldsymbol{\beta}_i - \mathbf{x}_i\|_2^2 \\ &- \boldsymbol{\mu}_i{}^T(\boldsymbol{\beta}_i + \Delta\boldsymbol{\beta}_i) - \|\mathbf{X}_{-i}\boldsymbol{\beta_i} - \mathbf{x}_i\|_2^2 - 2(\mathbf{X}_{-i}\boldsymbol{\beta}_i \\ &- \mathbf{x}_i)\Delta\boldsymbol{\beta}_i + \lambda_i\mathbf{1}\Delta\boldsymbol{\beta}_i + (\boldsymbol{\beta}_i)\Delta\boldsymbol{\mu_i} |:= \mathbf{C}.\end{aligned} \tag{23}$$

Thus proving that our objective function has a Lipschitz continuous gradient. □

B. Subsequence Convergence

To analyze the convergence, our objective function is:

$$\min \mathcal{J}(\boldsymbol{\beta}_i, \lambda_i, \boldsymbol{\mu}_i) = h(\boldsymbol{\beta_i}, \lambda_i, \boldsymbol{\mu}_i) + \delta_{\boldsymbol{\beta}}(\boldsymbol{\beta_i}) + \delta_\lambda(\lambda_i) + \delta_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i), \tag{24}$$

where δ is an indicator function. If the element is in the set, the function equals zero, otherwise, it approaches infinity.

Theorem IV.2. *(Subsequence convergence) The solution sequence S_k of Equation* (24) *is bounded and has three important properties:*

1) *Sufficient decrease:*
$$f(S_{k-1}) - f(S_k) \geq \frac{\min(\gamma_1, \gamma_2, \gamma_3) - L_c}{2}\|S_k - S_{k-1}\|_F^2, \tag{25}$$
which implies $\lim_{k\to\infty}\|S_{k-1} - S_k\|_F = 0$.
2) *Convergence:*
$$\lim_{k\to\infty}\|f(S_k)\| = constant. \tag{26}$$
The difference between subsequences,
$$\lim_{k\to\infty}\|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}_k\|_F = 0, \lim_{k\to\infty}\|\lambda_{k-1} - \lambda_k\|_F = 0$$
$$\lim_{k\to\infty}\|\boldsymbol{\mu}_{k-1} - \boldsymbol{\mu}_k\|_F = 0 \tag{27}$$
3) *For any convergent subsequence $\{S_{k'}\}$, the limit point $S^\star$ is a critical point of f and*
$$\lim_{k'\to\infty} f(S_{k'}) = \lim_{k\to\infty} f(S_k) = f(S^\star) \tag{28}$$

Before showing the proof, we provide necessary definitions to establish a clear understanding of the concepts involved.

Definition IV.1. Let f: $\mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semi-continuous function, given the domain of σ as:

$$\operatorname{dom} f := x \in \mathbb{R}^d : \sigma(x) < +\infty. \tag{29}$$

Fréchet subdifferential of $f(\partial f)$ at $x \in \operatorname{dom} f$ is the set of all vectors $z \in \mathbb{R}^d$ representing:

$$\partial f(z) = \{z : \lim_{y \to x} \inf \frac{f(y) - f(x) - < z, y - x >}{\|y - x\|} \geq 0\}. \tag{30}$$

if $x \notin dom f$, then $\partial f(x) = \emptyset$. The point x resulting in a local minimum of a domain, $0 \in \partial f(x)$, is called a limiting critical point of f, or critical point in simple.

Proof. According to the definition of the indicator function, we have the function of δ equal to zero if the derived solution belongs to its sequence. Based on Equation (24):

$$h_{L_c}(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}) \geq h(\boldsymbol{\beta}_{i(k-1)}, \lambda_i, \boldsymbol{\mu}_{i(k-1)}), \tag{31}$$

and by the definition of proximal map:

$$\lambda_{i(k)} = \arg\min_{\lambda_i} \delta_{\lambda_i}(\lambda_i) + < \nabla_{\lambda_i} h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}), \Delta_{\lambda_i} > + \frac{\gamma_2}{2} \|\Delta_{\lambda_i}\|_F^2. \tag{32}$$

Since $\lambda_{i(k)} \leq \lambda_{i(k-1)}$, we have:

$$\delta_{\lambda_i}(\lambda_{i(k)}) + < \nabla_{\lambda_i} h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}), \Delta_{\lambda_i} > + \frac{\gamma_2}{2} \|\Delta_{\lambda_i}\|_F^2 \leq \delta_{\lambda_i}(\lambda_{i(k-1)}) + < \nabla_{\lambda_i} h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{(k-1)} \boldsymbol{\mu}_{i(k-1)}), \Delta_{\lambda_i} > + \frac{\gamma_2}{2} \|\Delta_{\lambda_i}\|_F^2 = 0. \tag{33}$$

Applying Equation (33) and Lipschitz continuous gradient definition, and Taylor expansion:

$$\begin{aligned} &h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k)}, \boldsymbol{\mu}_{i(k-1)}) \leq h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}) \\ &+ < \nabla_{\lambda_i} h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}), \Delta_{\lambda_i} > + \frac{L_c}{2} \|\Delta_{\lambda_i}\|_F^2 \\ &\leq h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}) + \frac{L_c}{2} \|\Delta_{\lambda_i}\|_F^2 - \frac{\gamma_2}{2} \|\Delta_{\lambda_i}\|_F^2 \end{aligned} \tag{34}$$

$$h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}) - h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k)}, \boldsymbol{\mu}_{i(k-1)}) \geq \frac{\gamma_2 - L_c}{2} \|\Delta_{\lambda_i}\|_F^2. \tag{35}$$

Similarly, the other two sequence inequalities can be derived as:

$$h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}) - h(\boldsymbol{\beta}_{i(k)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}) \geq \frac{\gamma_1 - L_c}{2} \|\Delta_{\boldsymbol{\beta}_i}\|_F^2, \tag{36}$$

and

$$h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k-1)}) - h(\boldsymbol{\beta}_{i(k-1)}, \lambda_{i(k-1)}, \boldsymbol{\mu}_{i(k)}) \geq \frac{\gamma_3 - L_c}{2} \|\Delta_{\boldsymbol{\mu}_i}\|_F^2. \tag{37}$$

Repeating Equation (35), Equation (36) and Equation (37) together $\forall k$ will give us Equation (25).

$$\begin{aligned} &f(\boldsymbol{\beta}_{i(0)}) - f(\boldsymbol{\beta}_{i(N)}) \\ &\geq \frac{\min(\gamma_1, \gamma_2, \gamma_3) - L_c}{2} \sum_{k=1}^{N} (\|\Delta_{\boldsymbol{\beta}_i}\|_F^2 + \|\Delta_{\lambda_i}\|_F^2 \\ &+ \|\Delta_{\boldsymbol{\mu}_i}\|_F^2), \\ &\sum_{k=1}^{N} (\|\Delta_{\boldsymbol{\beta}_i}\|_F^2 + \|\Delta_{\lambda_i}\|_F^2 + \|\Delta_{\boldsymbol{\mu}_i}\|_F^2) \\ &\leq \frac{2 f(\boldsymbol{\beta}_{i(0)})}{\min(\gamma_1, \gamma_2, \gamma_3) - L_c}. \end{aligned} \tag{38}$$

The sequence $\sum_{k=1}^{N} (\|\Delta_{\boldsymbol{\beta}_i}\|_F^2 + \|\Delta_{\lambda_i}\|_F^2 + \|\Delta_{\boldsymbol{\mu}_i}\|_F^2)$ is also convergent because it is non-decreasing and upper-bounded.

$$\lim_{i \to \infty} (\|\Delta_{\boldsymbol{\beta}_i}\|_F^2 + \|\Delta_{\lambda_i}\|_F^2 + \|\Delta_{\boldsymbol{\mu}_i}\|_F^2) = 0. \tag{39}$$

For critical points:

$$\begin{aligned} \lim_{\bar{k} \to \infty} f(S_{\bar{k}}) &= \lim_{\bar{k} \to \infty} h(\boldsymbol{\beta}_{i\bar{k}}, \lambda_{i\bar{k}}, \boldsymbol{\mu}_i \bar{k}) \\ &+ \delta_{\boldsymbol{\beta}_i}(\boldsymbol{\beta}_{i\bar{k}}) + \delta_{\lambda_i}(\lambda_{i\bar{k}}) + \delta_{\boldsymbol{\mu}_i}(\boldsymbol{\mu}_{i\bar{k}}) \\ &= f(\bar{S}). \end{aligned} \tag{40}$$

As a result convergence of $\{f(S_k)\}_{k \geq 0}$, we have

$$\lim_{k \to \infty} \|S_{(k+1)} - S_{(k)}\|_F = 0. \tag{41}$$

To show that $\bar{S}$ is a critical point, we can consider the first-order optimal condition of Eq. (32):

$$\nabla_{\boldsymbol{\beta}_i} h(\boldsymbol{\beta}_{ik-1}, \lambda_{ik-1}, \boldsymbol{\mu}_{ik-1}) + \gamma_1(\boldsymbol{\beta}_{ik} - \boldsymbol{\beta}_{ik-1}) + \partial \delta_{\boldsymbol{\beta}_i}(\boldsymbol{\beta}_{ik}) = 0. \tag{42}$$

similarly, given

$$\nabla_{\lambda_i} h(\boldsymbol{\beta}_{ik}, \lambda_{ik-1}, \boldsymbol{\mu}_{ik-1}) + \gamma_2(\lambda_{ik} - \lambda_{ik-1}) + \partial \delta_{\lambda_i}(\lambda_{ik}) = 0. \tag{43}$$

as well as

$$\nabla_{\boldsymbol{\mu}_i} h(\boldsymbol{\beta}_{ik}, \lambda_{ik-1}, \boldsymbol{\mu}_{ik-1}) + \gamma_2(\lambda_{ik} - \lambda_{ik-1}) + \partial \delta_{\lambda_i}(\lambda_{ik}) = 0. \tag{44}$$

As k approaches infinity:

$$\begin{aligned} &\|\nabla_{\boldsymbol{\beta}_i} h(\boldsymbol{\beta}_{ik}, \lambda_{ik}, \boldsymbol{\mu}_{ik}) + \partial \delta_{\boldsymbol{\beta}_i}(\boldsymbol{\beta}_{ik})\|_F \leq \\ &\|\nabla_{\boldsymbol{\beta}_i} h(\boldsymbol{\beta}_{ik}, \lambda_{ik}, \boldsymbol{\mu}_{ik}) - \nabla_{\boldsymbol{\beta}_i} h(\boldsymbol{\beta}_{ik-1}, \lambda_{ik-1}, \boldsymbol{\mu}_{ik-1})\|_F \\ &+ \gamma_1 \|\boldsymbol{\beta}_{ik} - \boldsymbol{\beta}_{ik-1}\|_F \leq (L_c + \gamma_1) \|S_k - S_{k-1}\| = 0 \end{aligned} \tag{45}$$

and

$$\begin{aligned} &\left\|\nabla_{\lambda_i} h(\boldsymbol{\beta}_{ik}, \lambda_{ik}, \boldsymbol{\mu}_{ik}) + \partial \delta_{\lambda_i}(\lambda_{ik})\right\|_F \leq (L_c + \gamma_2) \|S_k - S_{k-1}\| = 0 \\ &\left\|\nabla_{\boldsymbol{\mu}_i} h(\boldsymbol{\beta}_{ik}, \lambda_{ik}, \boldsymbol{\mu}_{ik}) + \partial \delta_{\boldsymbol{\mu}_i}(\boldsymbol{\mu}_{ik})\right\|_F \leq (L_c + \gamma_3) \|S_k - S_{k-1}\| = 0 \end{aligned} \tag{46}$$

Finally, we have: $(2L_c + \gamma_1 + \gamma_2 + \gamma_3)\|S_k - S_{k-1}\| \geq \operatorname{dist}(0, \partial f(S_k))$, thus $0 \in \partial f(\bar{S})$ since $\partial f(S_{\bar{i}})$ is closed. Therefore, we can say that $\bar{S}$ is a critical point of f. □

Theorem IV.3. *(Sequence convergence) The sequence $\{M_k\}_{k \geq 0}$ generated by Algorithm with a constant step size $(\gamma_1, \gamma_2, \gamma_3 \geq L_c)$ is a global convergent sequence.*

Before Presenting the proof, we will give the following definition:

Definition IV.2. (Kurdyka-Lojasiewicz (KL property)) we say a proper semi-continuous function $h(u)$ satisfies Kurdyka-Lojasiewicz (KL) property, if $\bar{u}$ is a critical point of $h(u)$, then there exists $\delta > 0, \theta \in [0, 1), C_1 > 0$, s.t.

$$|h(\boldsymbol{u}) - h(\overline{\boldsymbol{u}})|^{\theta} \leq C_1 \operatorname{dist}(0, \partial h(\boldsymbol{u})), \forall \boldsymbol{u} \in B(\overline{\boldsymbol{u}}, \delta). \tag{47}$$

It is known that semi-algebraic functions satisfy the KL inequality. Clearly, the objective function f is semi-algebraic as h, $\delta_{\boldsymbol{\beta}_i}, \delta_{\lambda_i}$ and $\delta_{\boldsymbol{\mu}_i}$ are semi-algebraic. Therefore the objective function satisfies the KL property.

Proof. According to the KL property (in the Supplementary), the objective function is subjected to the following:

$$\left[f(S_k) - f(\bar{S})\right]^{\theta} \leq C_2 \operatorname{dist}(0, \partial f(S_k)), \forall k \geq k_0. \tag{48}$$

In order to construct the concavity, we introduce a concave function of $x^{(1-\theta)}$ for some $\theta \in [0, 1)$ with domain $x > 0$,

$$x_2^{1-\theta} - x_1^{1-\theta} \geq (1-\theta) x_2^{-\theta} (x_2 - x_1), \forall x_1 > 0, x_2 > 0. \tag{49}$$

Combined with the sufficient decrease property, Equation (49) can be updated by replacing x_1 and x_2 with $f(S_{k+1}) - f(\bar{S})$

and $f(S_k) - f(\bar{S})$, respectively. Then we have the following:

$$\begin{aligned}
&\left[f(S_k) - f(\bar{S})\right]^{1-\theta} - \left[f(S_{k+1}) - f(\bar{S})\right]^{1-\theta} \\
&\geq (1-\theta)\frac{f(S_k) - f(S_{k+1})}{\left[f(S_k) - f(\bar{S})\right]^{\theta}} \geq \frac{\lambda(1-\theta)}{2C_2}\frac{\|S_k - S_{k+1}\|_F^2}{\operatorname{dist}(0, \partial f(S_k))} \\
&\geq \frac{\lambda(1-\theta)}{2C_2C_3}\frac{\|S_k - S_{k+1}\|_F^2}{\|S_k - S_{k-1}\|_F} \\
&= \kappa\left(\frac{\|S_k - S_{k+1}\|_F^2}{\|S_k - S_{k-1}\|_F} + \|S_k - S_{k-1}\|_F\right) - \kappa\|S_i - S_{k-1}\|_F \\
&\geq \kappa\left(2\|S_k - S_{k+1}\|_F - \|S_k - S_{k-1}\|_F\right).
\end{aligned} \tag{50}$$

Accordingly:

$$\begin{aligned}
&2\|S_k - S_{k+1}\|_F - \|S_k - S_{k-1}\|_F \\
&\leq \alpha\left(\left[f(S_k) - f(\bar{S})\right]^{1-\theta}\right) - \left[f(S_{k+1}) - f(\bar{S})\right]^{1-\theta}\Big).
\end{aligned} \tag{51}$$

with $C_3 := 2L_c + \gamma_1 + \gamma_2, \kappa := \frac{\gamma_2(1-\theta)}{2C_2C_3}, \beta := \left(\frac{\gamma_2(1-\theta)}{2C_2C_3}\right)^{-1}$.

Summing the above inequalities up from some $\tilde{k} > k_0$ to infinity yields:

$$\sum_{k=\tilde{k}}^{\infty}\|S_k - S_{k+1}\|_F \leq \left\|S_{\tilde{k}} - S_{\tilde{k}-1}\right\|_F + \alpha\left[f(S_{\tilde{k}}) - f(\bar{S})\right]^{1-\theta} \tag{52}$$

, which indicates that $\sum_{k=\tilde{i}}^{\infty}\|S_k - S_{k+1}\|_F < \infty$. Following the standard argument, it is obvious that:

$$\lim_{t\to\infty}\sup\|S_{t_1} - S_{t_2}\|_F = 0, \text{ s.t.} \quad t_1, t_2 \geq t. \tag{53}$$

, which implies that the sequence S_k is Cauchy, thus convergent. The limit point set $C(S_0)$ is a singleton $\bar{S}$. □

Theorem IV.4. *(Convergence rate) The convergence rate is at least sub-linear.*

Proof. Since $\{S_k\}$ converges to $\bar{S}$, i.e., $\lim_{k\to\infty} S^k = \bar{S}$. From Equation (52) and the triangle inequality:

$$\begin{aligned}
\left\|S_{\tilde{k}} - \bar{S}\right\|_F &\leq \sum_{i=\tilde{k}}^{\infty}\|S_k - S_{k+1}\|_F \\
&\leq \left\|S_{\tilde{k}} - S_{\tilde{k}-1}\right\|_F + \alpha\left[f(S_{\tilde{k}}) - f(\bar{S})\right]^{1-\theta}.
\end{aligned} \tag{54}$$

which indicates the convergence rate of $S_{\tilde{k}} \to \bar{S}$ is at least as fast as the rate that $\left\|S_{\tilde{k}}S_{\tilde{k}-1}\right\|_F + \alpha\left[f(S_{\tilde{k}}) - f(\bar{S})\right]^{1-\theta}$. In particular, the second term $\alpha\left[f\left(\tilde{S}_k\right)\right]^{(1-\theta)}$ can be controlled:

$$\begin{aligned}
\alpha\left[f(S_{\tilde{k}}) - f(\bar{S})\right]^{\theta} &\leq \alpha C_2 \operatorname{dist}\left(0, \partial f(S_{\tilde{k}})\right) \\
&\leq \alpha C_2(2A_0 + \lambda + \|\mathbf{X}\|_F)\left\|S_{\tilde{k}} - S_{\tilde{k}-1}\right\|_F.
\end{aligned} \tag{55}$$

Thus we have:

$$\begin{aligned}
\sum_{i=\tilde{k}}^{\infty}\|S_k - S_{k+1}\|_F &\leq \left\|S_{\tilde{k}} - S_{\tilde{k}-1}\right\|_F \\
&+ \alpha C_2(2A_0 + \lambda + \|\mathbf{X}\|_F)\left\|S_{\tilde{k}} - S_{\tilde{k}-1}\right\|_F^{\frac{1-\theta}{\theta}}.
\end{aligned} \tag{56}$$

We divide the following analysis into three cases based on the value of the KL exponent θ:

Case 1 if $\theta = 0$, we set $Q := k \in \mathbb{N} : S_{k+1} \neq S_k$ and take k in Q. When k is large enough, then we have:

$$\|S_{k+1} - S_k\|_F^2 := C_4 > 0. \tag{57}$$

At the same time, we have:

$$\begin{aligned}
f(S_{k+1}) - f(S_k) &\geq \frac{\min(\gamma_1, \gamma_2, \gamma_3) - L_c}{2}\|S_{k+1} - S_k\|_F^2 \\
&= \frac{\min(\gamma_1, \gamma_2, \gamma_3) - L_c}{2}C_4.
\end{aligned} \tag{58}$$

Since $f(S_k)$ is known to converge to 0, the above equation shows that Q is finite and sequence S_k converges in a finite number of steps.

Case 2: $\theta \in \left(0, \frac{1}{2}\right]$, which corresponds to $\frac{(1-\theta)}{\theta} \geq 1$. We define $P_{\tilde{k}} = \sum_{k=\tilde{k}}^{\infty}\|S_{k+1} - S_k\|_F$,

$$P_{\tilde{k}} \leq P_{\tilde{k}-1} - P_{\tilde{k}} + \alpha C_2(2A_0 + \lambda + \|\mathbf{X}\|_F)\left[P_{\tilde{k}-1} - P_{\tilde{k}}\right]^{\frac{1-\theta}{\theta}}, \tag{59}$$

Since $P_{\tilde{k}-1} - P_{\tilde{k}} \to 0$, there exists a positive integer k_1 such that $P_{\tilde{k}-1} - P_{\tilde{k}} < 1, \forall\bar{k} \geq k_1$. Thus,

$$\begin{aligned}
&P_{\tilde{k}} \leq \left[1 + \alpha C_2(2A_0 + \lambda + \|\mathbf{X}\|_F)\right]\left(P_{\tilde{k}-1} - P_{\tilde{k}}\right), \\
&\forall\bar{k} \geq \max\{k_0, k_1\}.
\end{aligned} \tag{60}$$

Let $\rho = \frac{1+\alpha C_2(2A_0+\lambda+\|\mathbf{X}\|_F)}{2+\alpha C_2(2A_0+\lambda+\|\mathbf{X}\|_F)} \in (0,1)$, and combining with Equation (54) results in

$$\begin{aligned}
&P_{\tilde{k}} \leq \left[1 + \alpha C_2(2A_0 + \lambda + \|\mathbf{X}\|_F)\right]\left(P_{\tilde{k}-1} - P_{\tilde{k}}\right), \\
&\text{where } \bar{k} = \max\{k_0, k_1\}.
\end{aligned} \tag{61}$$

Case 3: $\theta \in \left(\frac{1}{2}, 1\right)$, which indicates $\frac{1-\theta}{\theta} \leq 1$. Based on previous results, we have:

$$\begin{aligned}
&P_{\tilde{k}} \leq \left[1 + \alpha C_2(2A_0 + \lambda + \|\mathbf{X}\|_F)\right]\left[P_{\tilde{k}-1} - P_{\tilde{k}}\right]^{\frac{1-\theta}{\theta}}, \\
&\forall\bar{k} \geq \max\{k_0, k_1\}.
\end{aligned} \tag{62}$$

Similarly, we get:

$$P_{\tilde{k}}^{\frac{1-2\theta}{1-\theta}} - P_{k-1}^{\frac{1-2\theta}{1-\theta}} \geq \zeta, \forall k \geq \bar{k}, \text{for some } \zeta > 0. \tag{63}$$

Then repeating and summing up the upper inequality from $\bar{k} = \max\{k_0, k_1\}$ to any $k > \bar{k}$, we get:

$$\begin{aligned}
P_{\tilde{k}} &\leq \left[P_{\tilde{k}-1}^{\frac{1-2\theta}{1-\theta}} + \zeta(\tilde{k} - \bar{k})\right]^{\frac{\theta-1}{2\theta-1}} \\
&= \mathcal{O}\left((\tilde{k} - \bar{k})^{\frac{\theta-1}{2\theta-1}}\right).
\end{aligned} \tag{64}$$

We thus prove the sub-linear convergence:

$$\left\|S_k - \bar{S}\right\|_F \leq \mathcal{O}\left((\tilde{k} - \bar{k})^{\frac{\theta-1}{2\theta-1}}\right), \forall k > \bar{k}. \tag{65}$$

□

TABLE I
DATA SET DESCRIPTION

Data Set	Size	Dimension	Classes
Wine	178	13	3
Iris	150	4	3
Glass	214	9	6
Breast Cancer	569	30	2
Albalone	4177	8	3
AT&T	400	4096	40
USPS	7291	256	10

V. EXPERIMENT AND RESULTS

In this section, we apply the RSSR model to both synthetic data and real-world datasets to evaluate the model's performance. Additionally, we compare the clustering performance of the RSSR model with that of the original dataset.

A. Synthetic Data Experiment

To visually demonstrate the effectiveness of our proposed algorithm, we conduct experiments on a synthetic dataset

TABLE II

CLUSTERING PERFORMANCE OF DIFFERENT METHODS ON REAL-WORLD DATA SETS. THE EVALUATION METRICS ARE AVERAGED WITH 50 RUNS.

ACC	**Representation**	**NMF**	**LSTM**	**K-means**	**NCut**	**STSC**	**Hierarchical**	**EgoNetSplitter**	**MNMF**	**SymmNMF**	**GEMSEC**
Wine	Original	0.882	0.360	0.376	0.438	0.354	0.337	0.158	**0.539**	0.163	0.333
	RSSR	**0.888**	**0.764**	**0.972**	**0.618**	**0.371**	**0.376**	**0.202**	0.494	**0.484**	**0.473**
Iris	Original	0.587	0.127	0.833	0.847	**0.387**	0.687	0.073	0.500	0.087	0.487
	RSSR	**0.647**	**0.667**	**0.867**	**0.887**	0.320	**0.740**	**0.133**	**0.533**	**0.093**	**0.522**
Glass	Original	0.256	0.126	0.304	**0.372**	0.140	**0.336**	0.137	0.145	0.112	0.383
	RSSR	**0.435**	**0.136**	**0.388**	0.341	**0.154**	0.327	**0.210**	**0.383**	**0.136**	**0.458**
Breast Cancer	Original	0.580	**0.362**	0.634	0.376	0.496	0.374	0.011	0.509	0.069	0.524
	RSSR	**0.647**	0.359	**0.645**	**0.622**	**0.503**	**0.376**	**0.019**	**0.536**	**0.076**	**0.634**
NMI	**Representation**	**NMF**	**LSTM**	**K-means**	**NCut**	**STSC**	**Hierarchical**	**EgoNetSplitter**	**MNMF**	**SymmNMF**	**GEMSEC**
Wine	Original	**0.708**	0.248	0.876	**0.572**	0.002	0.018	0.480	0.089	0.375	0.266
	RSSR	0.701	**0.924**	**0.893**	0.562	**0.005**	**0.549**	**0.508**	**0.252**	**0.385**	**0.276**
Iris	Original	0.692	0.734	0.657	0.690	**0.290**	**0.713**	0.491	0.214	0.386	**0.227**
	RSSR	**0.720**	**0.771**	**0.696**	**0.730**	0.010	0.704	**0.522**	**0.258**	**0.404**	**0.253**
Glass	Original	0.424	0.486	0.471	**0.310**	0.132	0.101	0.506	0.286	0.466	0.326
	RSSR	**0.480**	**0.531**	**0.530**	0.100	**0.156**	**0.345**	**0.512**	**0.319**	**0.492**	**0.387**
Breast Cancer	Original	0.014	0.038	0.024	0.005	0.290	0.002	0.057	0.214	0.022	0.007
	RSSR	**0.027**	**0.041**	**0.037**	**0.730**	**0.503**	**0.005**	**0.058**	**0.714**	**0.028**	**0.015**
Purity	**Representation**	**NMF**	**LSTM**	**K-means**	**NCut**	**STSC**	**Hierarchical**	**EgoNetSplitter**	**MNMF**	**SymmNMF**	**GEMSEC**
Wine	Original	0.719	0.785	0.966	**0.635**	0.399	0.399	0.927	0.497	0.817	0.567
	RSSR	**0.899**	**0.978**	**0.972**	0.629	**0.404**	**0.629**	**0.944**	**0.685**	**0.865**	**0.579**
Iris	Original	0.687	0.667	0.831	0.847	**0.413**	0.687	0.920	0.533	0.880	0.513
	RSSR	**0.707**	**0.873**	**0.867**	**0.887**	0.393	**0.740**	**0.974**	**0.633**	**0.907**	**0.520**
Glass	Original	0.631	0.640	0.603	0.379	0.463	0.379	0.781	0.549	0.745	0.580
	RSSR	**0.673**	**0.696**	**0.738**	**0.477**	**0.463**	**0.486**	**0.790**	**0.673**	**0.832**	**0.664**
Breast Cancer	Original	0.627	0.638	0.634	0.847	0.627	0.543	**0.809**	0.533	0.663	0.628
	RSSR	**0.647**	**0.641**	**0.645**	**0.887**	**0.629**	**0.627**	0.714	**0.627**	**0.675**	**0.638**

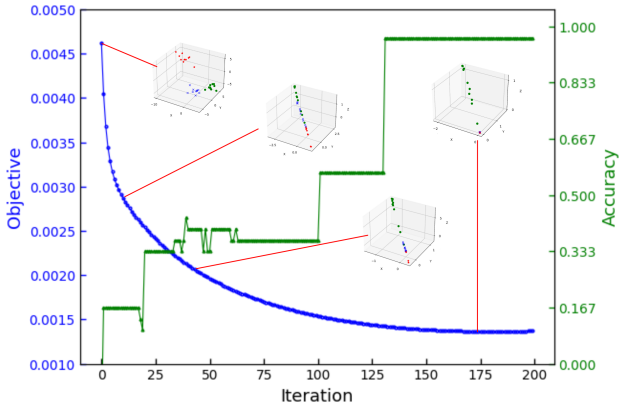

Fig. 2. Objective function with the accuracy on the noisy synthetic dataset

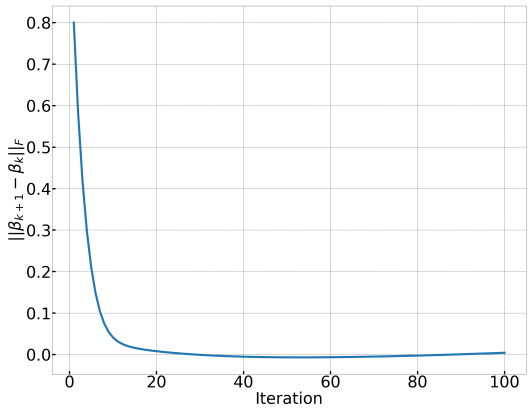

Fig. 3. $||\beta_{k+1} - \beta_k||_F$ plotted as RSSR iterates on the noisy synthetic dataset

consisting of three centroids in a 20-dimensional space, each surrounded by 30 data points. Three centroids are normal distributed vectors with a dimension of 20 are constructed. Each centroid has a different mean value. The goal is to do the clustering for these three centroids of vectors. In addition, noises (mean is set as 0, standard deviation is set as 0.5) are added to the distributed vectors to avoid trivial testing. Experiments are run 100 times with different random seeds. Three standard metrics for clustering are deployed for comparing the clustering performances among various clustering methods.

In Fig. 2, our objective function monotonically decreases with each iteration, indicating the convergence of our algorithm. For the purpose of visualization, we embed the data into a 3-dimensional space (r = 3), where the x, y, and z axes correspond to the first, second, and third columns of the original matrix X. In addition, the accuracy stabilizes after 100 iterations and the accuracy reaches 0.96 once stabilized. From the figure, it is evident that the clustering performance on the simplex representation improves as the algorithm iterates as the accuracy gets higher. This experiment not only demonstrates the effectiveness of our proposed method but also empirically confirms the sub-linear convergence of our objective function and the convergence of the solution. As shown in Fig. 3, the $\boldsymbol{\beta}$ variance $||\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k||_F$ converges to 0 in less than 10 iterations and stabilizes in 40 iterations.

B. Real Data Experiment

Table I summarizes the datasets used in the real experiment section. We extracted these datasets from the UCI machine learning repository. In this section, we compare our proposed new method with the NMF [14], LTSM [15], K-means [16], Normalized cut (NCut) [17], Self-Tuning Spectral Clustering (STSC) [18], DBSCAN [19], EgoNetSplitter [20], MNMF [21], SymmNMF [22], GEMSEM [23] methods. Results will be compared including clustering Accuracy (Acc.), Normalized Mutual Information (NMI) and Purity.

The parameter setting for the real dataset experiment is as follows: We split each real dataset into a training and a test set with a proportion of 80% and 20% upon the total number

of samples. During each iteration step, our model learns an updated similarity matrix $\mathbf{S}$ which will be applied to be used as the simplex representation for the dataset. Once the $\mathbf{W}$ (recall that $\mathbf{W} = \frac{1}{2}(\mathbf{S} + \mathbf{S}^T)$) has been learned, we run different clustering methods in the test set with the original representation $\mathbf{X}$ and the simplex representation $\mathbf{XW}$ and compare these two based on the three metrics: Acc, NMI, and purity. We use a 5-fold cross-validation within the training set to identify the best hyperparameters. The validation set in this inner cross-validation is chosen to be 20% of the training set.

C. Original vs Representations

We run several methods for clustering after the new representation is learned. In Table II, each experiment repeats 50 times, and the average is taken. The clustering result is then reported with the original representation versus the simplex representation. We see that RSSR always achieves higher clustering accuracy, normalized mutual information, and purity in analysis. We also compare our method with other methods on large datasets as Table III demonstrates, where since the dataset is large, we split the whole data into small batches due to the high computation cost of the counterpart algorithms and report the average result. Although our method may not consistently outperform others on small datasets, the results remain promising and generally reliable. Overall, the experiments demonstrate strong performance across all datasets, highlighting its potential for broader applications.

TABLE III
CLUSTERING PERFORMANCE OF DIFFERENT METHODS ON LARGE REAL-WORLD DATA SETS, AVERAGED OVER 50 RUNS.

ACC	**Representation**	**NMF**	**LSTM**	**K-means**	**NCut**	**STSC**
Albalone	Original	0.409	0.145	0.431	0.368	0.362
	RSSR	**0.483**	**0.164**	**0.561**	**0.382**	**0.463**
AT&T	Original	0.217	0.253	0.264	**0.323**	**0.412**
	RSSR	**0.376**	**0.266**	**0.298**	0.312	0.380
USPS	Original	0.165	0.088	0.173	**0.171**	0.170
	RSSR	**0.173**	**0.164**	**0.179**	0.165	**0.182**
NMI	**Representation**	**NMF**	**LSTM**	**K-means**	**NCut**	**STSC**
Albalone	Original	0.168	0.323	0.175	**0.088**	0.104
	RSSR	**0.213**	**0.353**	**0.181**	0.062	**0.115**
AT&T	Original	0.484	0.443	0.464	0.524	**0.556**
	RSSR	**0.570**	**0.518**	**0.581**	**0.546**	0.437
USPS	Original	0.267	0.253	0.257	0.281	0.113
	RSSR	**0.294**	**0.330**	**0.324**	**0.302**	**0.196**
Purity	**Representation**	**NMF**	**LSTM**	**K-means**	**NCut**	**STSC**
Albalone	Original	0.326	0.124	0.358	**0.422**	0.475
	RSSR	**0.334**	**0.141**	**0.374**	0.343	**0.484**
AT&T	Original	0.333	0.363	0.383	0.378	**0.345**
	RSSR	**0.415**	**0.428**	**0.411**	**0.407**	0.326
USPS	Original	0.145	0.160	0.137	0.147	0.096
	RSSR	**0.164**	**0.257**	**0.231**	**0.177**	**0.107**

VI. CONCLUSION

In this work, our proposed Robust Simplex Sparse Representation methodology, coupled with the Proximal Alternating Linearized Minimization algorithm, addresses challenges in constructing graphical representations for high-dimensional data. This approach reduces computation, improves robustness to outliers, and outperforms state-of-the-art clustering methods on diverse datasets. Our research provides valuable insights into data representation challenges for big data.

REFERENCES

[1] Z. Peng, W. Huang, M. Luo, Q. Zheng, Y. Rong, T. Xu, and J. Huang, "Graph representation learning via graphical mutual information maximization," in *Proceedings of The Web Conference*, 2020, pp. 259–270.

[2] J. Laaksonen and E. Oja, "Classification with learning k-nearest neighbors," in *Proceedings of international conference on neural networks (ICNN'96)*, vol. 3. IEEE, 1996, pp. 1480–1483.

[3] L. Kozma, "k nearest neighbors algorithm (knn)," *Helsinki University of Technology*, vol. 32, 2008.

[4] Y. Xu, V. Olman, and D. Xu, "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees," *Bioinformatics*, vol. 18, no. 4, pp. 536–545, 2002.

[5] M. H. Rohban and H. R. Rabiee, "Supervised neighborhood graph construction for semi-supervised classification," *Pattern Recognition*, vol. 45, no. 4, pp. 1363–1372, 2012.

[6] G. Wen, "Robust self-tuning spectral clustering," *Neurocomputing*, vol. 391, pp. 243–248, 2020.

[7] J. J. Yanez-Borjas, J. M. Machorro-Lopez, D. Camarena-Martinez, M. Valtierra-Rodriguez, J. P. Amezquita-Sanchez, F. J. Carrion-Viramontes, and J. A. Quintana-Rodriguez, "A new damage index based on statistical features, pca, and mahalanobis distance for detecting and locating cables loss in a cable-stayed bridge," *International Journal of Structural Stability and Dynamics*, vol. 21, no. 09, p. 2150127, 2021.

[8] Y. Liu, Q. Gao, J. Han, and S. Wang, "Euler sparse representation for image classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[9] A. Shrivastava, V. M. Patel, and R. Chellappa, "Multiple kernel learning for sparse representation-based classification," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3013–3024, 2014.

[10] K. Lu, Z. Ding, and S. Ge, "Sparse-representation-based graph embedding for traffic sign recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1515–1524, 2012.

[11] A. Kyrillidis, S. Becker, V. Cevher, and C. Koch, "Sparse projections onto the simplex," in *International Conference on Machine Learning*. PMLR, 2013, pp. 235–243.

[12] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[13] T. Pock and S. Sabach, "Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems," *SIAM journal on imaging sciences*, vol. 9, no. 4, pp. 1756–1787, 2016.

[14] D. Guillamet, M. Bressan, and J. Vitria, "A weighted non-negative matrix factorization for local representations," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1. IEEE, 2001, pp. I–I.

[15] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker Diarization with LSTM," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[16] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[17] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[18] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," *Advances in neural information processing systems*, vol. 17, 2004.

[19] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[20] A. Epasto, S. Lattanzi, and R. Paes Leme, "Ego-splitting framework: From non-overlapping to overlapping clusters," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 145–154.

[21] X. Wang, P. Cui, J. Wang, J. Pei, W. Zhu, and S. Yang, "Community preserving network embedding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[22] D. Kuang, C. Ding, and H. Park, "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of the 2012 SIAM international conference on data mining*. SIAM, 2012, pp. 106–117.

[23] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, "Gemsec: Graph embedding with self clustering," in *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, 2019, pp. 65–72.

Supplementary Material

A. Derivation details

In this section, we provide the derivation of the solution for each variable discussed in Section. We take the derivative of the equation Eq. (12) to Eq. (14) with respect to each parameter and set it to zero to earn the update for each parameter.

Step 1: Updating $\boldsymbol{\beta}_{i(k+1)}$

$$\begin{aligned}\nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)}) &= \frac{\partial\mathcal{J}(\boldsymbol{\beta}_{i(k)},\lambda_i,\boldsymbol{\mu}_i)}{\partial(\boldsymbol{\beta}_{i(k)})} \\ &= 2\mathbf{X}_{-i}^T(\mathbf{X}_{-i}\boldsymbol{\beta}_{i(k)}-\mathbf{x}_i)-\lambda_i\mathbf{1}-\boldsymbol{\mu}_i\end{aligned} \tag{66}$$

Set derivative Eq. (12) to 0:

$$\begin{aligned}&\frac{\partial(\langle\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)},\nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})\rangle+\frac{\gamma_1}{2}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)}\|_2^2)}{\partial(\boldsymbol{\beta}_i)} \\ =&\frac{\partial(\langle\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)},\nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})\rangle)}{\partial(\boldsymbol{\beta}_i)}+\frac{\partial(\frac{\gamma_1}{2}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)}\|_2^2)}{\partial(\boldsymbol{\beta}_i)} \\ =&\frac{\partial(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)})}{\partial(\boldsymbol{\beta}_i)}\nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})+\frac{\partial(\nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})}{\partial(\boldsymbol{\beta}_i)}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)}) \\ &+\frac{\partial(\frac{\gamma_1}{2}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)}\|_2^2)}{\partial(\boldsymbol{\beta}_i)} \\ =&\mathbf{I}\nabla\mathcal{J}(\boldsymbol{\beta}_{i(k)})+0+\gamma_1(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)}) \\ =&0\end{aligned} \tag{67}$$

Plug in Eq. (66), this gives us:

$$2\mathbf{X}_{-i}^T(\mathbf{X}_{-i}\boldsymbol{\beta}_{i(k)}-\mathbf{x}_i)-\lambda_i\mathbf{1}-\boldsymbol{\mu}_i+\gamma_1(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i(k)})=0 \tag{68}$$

Which gives the updated equation (15).

Step 2: Updating $\lambda_{i(k+1)}$

$$\nabla\mathcal{J}(\lambda_{i(k)})=\frac{\partial\mathcal{J}(\boldsymbol{\beta}_i,\lambda_{i(k)},\boldsymbol{\mu}_i)}{\partial(\lambda_{i(k)})}=1-\boldsymbol{\beta}_i^T\mathbf{1} \tag{69}$$

Set derivative Eq. (13) to 0:

$$\begin{aligned}&\frac{\partial(\langle\lambda_i-\lambda_{i(k)},\nabla\mathcal{J}(\lambda_{i(k)})\rangle+\frac{\gamma_2}{2}(\lambda_i-\lambda_{i(k)})^2)}{\partial\lambda_i} \\ =&\frac{\partial(\langle\lambda_i-\lambda_{i(k)},\nabla\mathcal{J}(\lambda_{i(k)})\rangle)}{\partial\lambda_i}+\frac{\partial(\frac{\gamma_2}{2}(\lambda_i-\lambda_{i(k)})^2)}{\partial\lambda_i} \\ =&\frac{\partial(\lambda_i-\lambda_{i(k)})}{\partial\lambda_i}\nabla\mathcal{J}(\lambda_{i(k)})+\frac{\partial\nabla\mathcal{J}(\lambda_{i(k)})}{\partial\lambda_i} \\ &+\frac{\partial(\frac{\gamma_2}{2}(\lambda_i-\lambda_{i(k)})^2}{\partial\lambda_i} \\ =&\nabla\mathcal{J}(\lambda_{i(k)})+0+\gamma_2(\lambda_i-\lambda_{i(k)}) \\ =&1-\boldsymbol{\beta}_{i(k)}^T\mathbf{1}+\gamma_2(\lambda_i-\lambda_{i(k)})=0\end{aligned} \tag{70}$$

Which gives us the updated equation (16).

Step 3: Updating $\boldsymbol{\mu}_{i(k+1)}$

$$\nabla\mathcal{J}(\boldsymbol{\mu}_{i(k)})=\frac{\partial\mathcal{J}(\boldsymbol{\beta}_i,\lambda_i,\boldsymbol{\mu}_{i(k)})}{\partial(\boldsymbol{\mu}_{i(k)})}=-\boldsymbol{\beta}_i \tag{71}$$

Set derivative Eq. (14) to 0:

$$\begin{aligned}&\frac{\partial(\langle\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i(k)},\nabla\mathcal{J}(\boldsymbol{\mu}_{i(k)})\rangle+\frac{\gamma_3}{2}\|(\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i(k)})\|_2^2)}{\partial\boldsymbol{\mu}_i} \\ =&\frac{\partial(\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i(k)})}{\partial\boldsymbol{\mu}_i}\nabla\mathcal{J}(\boldsymbol{\mu}_{i(k)})+\frac{\partial(\nabla\mathcal{J}(\boldsymbol{\mu}_{i(k)}))}{\partial\boldsymbol{\mu}_i}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i(k)}) \\ &+\frac{\partial(\frac{\gamma_3}{2}\|(\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i(k)})\|_2^2)}{\partial\boldsymbol{\mu}_i} \\ =&\mathbf{I}\nabla\mathcal{J}(\boldsymbol{\mu}_{i(k)})+0+\gamma_3(\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i(k)}) \\ =&-\boldsymbol{\beta}_i+\gamma_3(\boldsymbol{\mu}_i-\boldsymbol{\mu}_{i(k)})=0\end{aligned} \tag{72}$$

Which gives us the updated equation (17).

B. Convergence Analysis Related Lemma

Lemma VI.1. *[Uniform KL Property] The objective function satisfies the KL property. There exists* $\delta_0>0$, $\theta_{KL}\in[0,1), C_{KL}>0$ *such that as long as* $\operatorname{dist}((S),\mathbb{C}(S_0))\leq\delta_0$*:*

$$|f(S)-\bar{f}|^{\theta_{KL}}\leq C_{KL}\operatorname{dist}(0,\partial f(S)), \tag{73}$$

with $\bar{f}$ *representing the limiting function value defined in assertion (i) of Theorem* (IV.2)*.*

Proof. First, we recognize that the union $\bigcup_k B\left(\overline{S_k},\delta_k\right)$ forms an open cover of $\mathbb{C}(S_0)$ with $\overline{S_k}$ denoting all points in $\mathbb{C}(S_0)$ and δ_i to be chosen so that the KL property of f at $\overline{S_i}\in\mathbb{C}(S_0)$ holds:

$$|f(S)-\bar{f}|^{\theta_k}\leq C_i\operatorname{dist}(0,\partial f(S)),\forall(S)\in B\left(\overline{S_k},\delta_k\right), \tag{74}$$

in which we have used all $f\left(\overline{S_k}\right)=\bar{f}$ by assertion (iii) of Theorem (IV.2). Then since $\mathbb{C}(S_0)$ is compact, it has a finite subcover $\bigcup_{k=1}^p B\left(\overline{S_{s_i}},\delta_{s_i}\right)$ for some positive integer p. Considering all these facts, we have $\forall S\in\bigcup_{k=1}^p B\left(\overline{S_{s_i}},\delta_{s_i}\right)$:

$$|f(S)-\bar{f}|^{\theta_{KL}}\leq C_{KL}\operatorname{dist}(0,\partial f(S)), \tag{75}$$

with $\theta_{KL}=\max_{k=1}^p\{\theta_{s_k}\}$ and $C_{KL}=\max_{k=1}^p\{C_{s_k}\}$. In the end, since $\bigcup_{k=1}^p B\left(\overline{S_{s_i}},\delta_{s_i}\right)$ is an open cover of $\mathbb{C}(S_0)$, there should exist a sufficiently small number δ_0, so that

$$\{(S):\operatorname{dist}(S,\mathbb{C}(S_0))\leq\delta_0\}\subset\bigcup_{k=1}^p B\left(\overline{S_k},\delta_{s_k}\right). \tag{76}$$

Thus Equation (75) holds when $\operatorname{dist}(S,\mathbb{C}(S_0))\leq\delta_0$. □