Journal of Industrial and
Management Optimization

https://www.aimspress.com/journal/jimo

JIMO, 22(1): 44–69.
DOI:10.3934/jimo.2026003
Received: 11 September 2025
Revised: 29 October 2025
Accepted: 06 November 2025
Published: 14 November 2025

Research article

Group sparsity-based fusion regularized clustering: New model and convergent algorithm

Xiangru Xing[1], **Lingchen Kong**[1], **Xin Wang**[1,*] **and Xianchao Xiu**[2]

[1] School of Mathematics and Statistics, Beijing Jiaotong University, Beijing, 100044, China
[2] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, 200444, China

* **Correspondence:** Email: xinwang2@bjtu.edu.cn.

Abstract: Fusion regularized clustering has gained considerable attention for its ability to perform clustering without prior knowledge of the number of clusters. However, its performance deteriorates in high-dimensional settings due to redundant or noisy features. To address this issue, we propose a novel group sparsity-based fusion regularized clustering model, termed GSFRC, which enables both accurate clustering and informative feature selection. Specifically, GSFRC introduces a bi-level group sparsity regularizer that integrates an inter-group sparsity term ($l_{2,p}$-norm) and an intra-group sparsity term (l_q-norm) to capture both global and local feature structures, where the parameters p and q are chosen from the interval $[0, 1)$. In theory, we establish the necessary Karush–Kuhn–Tucker (KKT) optimality conditions for the resulting nonconvex and non-Lipschitz optimization problem. In the algorithm, we develop an efficient scheme based on the alternating direction method of multipliers (ADMM), and we provide a rigorous global convergence analysis under mild conditions. Extensive experiments demonstrate that the proposed method achieves superior clustering accuracy, stronger feature selection capability, and higher robustness compared with existing fusion regularized clustering approaches.

Keywords: fusion regularized clustering; group sparsity; feature selection; alternating direction method of multipliers
Mathematics Subject Classification: 90C06, 90C26, 90C90

1. Introduction

Clustering serves as an essential task in machine learning, statistics, and data science, aiming to partition data into meaningful groups based on similarity or underlying structure. Common methods, including K-means clustering [1], hierarchical clustering [2], spectral clustering [3], and subspace clustering [4], are typically sensitive to initialization and require a known number of clusters. In contrast, fusion regularized clustering (FRC) overcomes both limitations and has been widely applied in many fields, such as genomics [5], image processing [6], and finance [7].

The FRC, known as clusterpath, sum-of-norms clustering, and convex clustering, which was first proposed by [8–10]. Given a data matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with n observations and d features, the fundamental FRC aims to solve the following optimization problem:

$$\min_{\boldsymbol{X}} F(\boldsymbol{X}) := \frac{1}{2}\sum_{i=1}^{n} \|\boldsymbol{X}_{i\cdot} - \boldsymbol{A}_{i\cdot}\|_2^2 + \alpha \sum_{i<i'} \omega_\iota \|\boldsymbol{X}_{i\cdot} - \boldsymbol{X}_{i'\cdot}\|_2, \tag{1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is a centroid matrix and $\alpha > 0$ is a tuning parameter. The fusion weight ω_ι is conventionally defined for a positive constant ϕ as

$$\omega_\iota = \begin{cases} \exp(-\phi\|\boldsymbol{A}_{i\cdot} - \boldsymbol{A}_{i'\cdot}\|_2^2), & \text{if } (i,i') \in \epsilon \\ 0, & \text{otherwise} \end{cases},$$

where $\epsilon = \cup_{i=1}^{n}\{(i,i') \mid \boldsymbol{A}_{i'\cdot}$ is among $\boldsymbol{A}_{i\cdot}$'s k-nearest neighbors, $1 \le i < i' \le n\}$. When two observations $\boldsymbol{A}_{i\cdot}$ and $\boldsymbol{A}_{i'\cdot}$ are connected by an edge (i,i'), the l_2-norm penalty fuses $\boldsymbol{X}_{i\cdot}$ and $\boldsymbol{X}_{i'\cdot}$ together and is thus termed the fusion penalty or fusion regularization. By tuning α for (1.1), the number of clusters can be determined automatically without prespecification.

Although extensive research has been established on the exact recovery guarantees, statistical properties, and optimization algorithms of FRC [11–15], its clustering performance may degrade in the presence of substantial noise, particularly in high-dimensional settings (where $d \gg n$). To address the curse of dimensionality, conventional dimension reduction techniques like projection or matrix factorization map data to lower-dimensional representations, at the expense of feature transparency as original features are transformed into latent components. Additionally, the critical yet challenging task of factor ranking in practice can significantly influence the final results [5]. In comparison, lasso regularization method preserves the original features while automatically inducing sparsity, thereby enhancing model interpretability. Building on this principle, Wang et al. [16] introduced sparse convex clustering via an inter-group lasso penalty, i.e.,

$$\min_{\boldsymbol{X}} F(\boldsymbol{X}) + \gamma \sum_{j=1}^{d} \beta_j \|\boldsymbol{X}_{\cdot j}\|_2, \tag{1.2}$$

where $\gamma > 0$ controls feature sparsity and β_j weights feature importance. Feature selection is achieved by appropriately increasing γ, which forces the l_2-norm of noise features to zero. Meanwhile, the resulting cluster centroid $\tilde{\boldsymbol{X}}$ is essentially a low-dimensional representation of the original feature space.

Besides, intra-group sparsity is ubiquitous in real-world applications [17–23]. For example, a biological pathway may be associated with the development of a specific cancer, but not every gene within the pathway is necessarily involved or activated [17]. To capture this bi-level sparsity structure, Chen et al. [14] enhanced feature selection by combining inter-group lasso and intra-group lasso penalties, which is formulated as

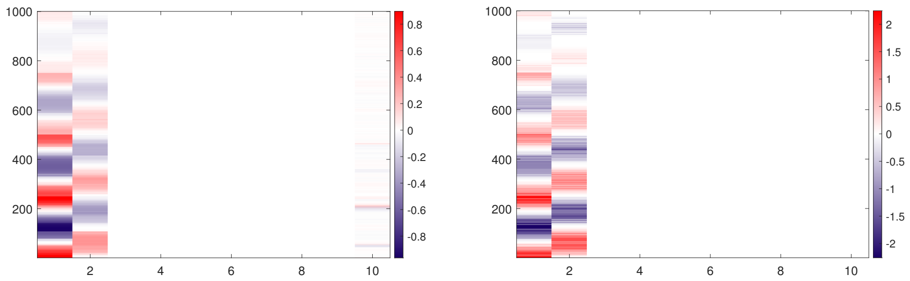
$$\min_{\boldsymbol{X}} F(\boldsymbol{X}) + \gamma \sum_{j=1}^{d} ((1-\eta)\beta_j \|\boldsymbol{X}_{\cdot j}\|_2 + \eta \|\boldsymbol{X}_{\cdot j}\|_1), \tag{1.3}$$

where l_1-norm induces element-wise sparsity by selecting local information from individual features and tuning parameter $\eta \in [0, 1]$ balances two sparsity terms. Currently, this group sparsity or double sparsity structure has demonstrated excellent performance in unsupervised learning [19, 20], bioinformatics [21], and wireless communications [22, 23]. In fact, the l_1-norm acts as a convex approximation of the l_q-norm ($q \in [0, 1)$), often introducing bias in variable selection. Numerous studies have shown that the non-convex l_q-norm, by virtue of its distinctive geometric properties, promotes stronger sparsity and more effective variable selection by aggressively shrinking noise while preserving salient features [24–28]. However, no existing FRC work has explored the group sparsity combined by both $l_{2,p}$-norm and l_q-norm, where both p and q lie in $[0, 1)$.

Motivated by the above idea, we propose the following group sparsity-based fusion regularized clustering model, i.e., GSFRC,

$$\min_{\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{A}\|_F^2 + \alpha\sum_{\iota\in\epsilon}\omega_\iota\|(\boldsymbol{D}\boldsymbol{X})_{\iota\cdot}\|_2 + \gamma((1-\eta)\|\boldsymbol{X}\boldsymbol{\beta}\|_{2,p}^p + \eta\|\boldsymbol{X}\|_q^q), \tag{1.4}$$

where the pairwise difference $(\boldsymbol{D}\boldsymbol{X})_{\iota\cdot} = \boldsymbol{X}_{i\cdot} - \boldsymbol{X}_{i'\cdot}$ for edge $\iota = (i, i') \in \epsilon$ and directed difference matrix $\boldsymbol{D} \in \mathbb{R}^{|\epsilon|\times n}$ such that the first two terms correspond to the matrix form of $F(\boldsymbol{X})$ in (1.1). For $p \in [0, 1)$ and $q \in [0, 1)$, $\|\boldsymbol{X}\boldsymbol{\beta}\|_{2,p}^p = \sum_{j=1}^d \|(\boldsymbol{X}\boldsymbol{\beta})_{\cdot j}\|_2^p$ and $\|\boldsymbol{X}\|_q^q = \sum_{i=1}^n \sum_{j=1}^d |\boldsymbol{X}_{ij}|^q$ are the so-called $l_{2,p}$-norm and l_q-norm, respectively. Unlike the models mentioned above, our group sparsity incorporates both inter-group sparsity (via $l_{2,p}$-norm) and intra-group sparsity (via l_q-norm). The diagonal weighting matrix $\boldsymbol{\beta} = \mathrm{Diag}(\beta_1,\beta_2, \beta_3, \cdots ,\beta_d)$ with user-specified positive entries β_j for $j \in [d]$, adaptively adjusts feature importance. See Algorithm 2 in Supplementary Material (Section I) for the specific adjustment scheme. Each term in (1.4) has a distinct role. The first term performs data fitting to bring the centroids $\boldsymbol{X}_{i\cdot}$ closer to their respective observations $\boldsymbol{A}_{i\cdot}$. The second term induces clustering by enforcing shared centroids for edge-connected observations. Under tuning parameters α and sparse weights ω_ι, this term drives $\boldsymbol{A}_{i\cdot}$ and $\boldsymbol{A}_{i'\cdot}$ to belong to the same cluster when $\mathbf{X}_{i\cdot} = \mathbf{X}_{i'\cdot}$ holds. The last two terms enable feature selection. The roles of parameters α, γ, and η are explained above, while their tuning effects are validated in Section 5.



(a) Centroid $\tilde{\mathbf{X}}$ from $l_{2,1} + l_1$

(b) Centroid $\tilde{\mathbf{X}}$ from $l_{2,p} + l_q$

Figure 1. Visualization of the resulting centroid matrix $\tilde{\mathbf{X}}$ by sparse group lasso and our proposed group sparsity on Dartboard1. Subfigure (1a) identifies 3 features, while Subfigure (1b) identifies 2 features with enhanced sparsity.

Similar to the sparse group lasso in (1.3), we employ the $l_{2,p}$-norm and l_q-norm to characterize feature-wise and element-wise sparsity, respectively. Nevertheless, the feature subsets selected by our group-sparsity method differ from those of sparse group lasso, as demonstrated in Figure 1. This difference influences the clustering results to some extent. Furthermore, the non-convexity and non-Lipschitzness of the group-sparsity term pose significant challenges to the convergence analysis of the optimization algorithm.

The primary contributions of this paper are fourfold:

1) We propose a novel GSFRC model to achieve automatic clustering without predefining the number of clusters while simultaneously enabling feature selection. It effectively overcomes the feature selection bias inherent in traditional lasso-type methods, significantly improving the accuracy and robustness.

2) We derive a KKT necessary optimality condition for the resulting non-convex and non-Lipschitz problem by using the spatial decomposition technique, which provides theoretical support for convergence analysis.

3) We design a 2-block ADMM algorithm that admits closed-form solutions for each subproblem. The global convergence of the algorithm is theoretically guaranteed under reasonable conditions.

4) We conduct extensive numerical experiments on six datasets to evaluate the clustering performance and feature selection capability of GSFRC. Besides, ablation studies are performed, and the model's stability, parameter sensitivity, convergence behavior, and runtime comparison are discussed to further demonstrate the effectiveness of GSFRC.

The paper is structured as follows: Section 2 introduces the notations and preliminary knowledge. The optimality conditions of GSFRC are detailed in Section 3. Section 4 devises the optimization algorithm with convergence and complexity analysis. Experimental results and discussions are demonstrated in Section 5, followed by conclusions in Section 6.

2. Preliminaries

This section begins with a summary of all notations and key terminology, followed by a presentation of the closed-form solutions for the proximal operators of specific l_q-norms.

2.1. Notations and terminology

Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ be a matrix. We denote its i-th row, j-th column, and (i, j)-th element by $\boldsymbol{A}_{i\cdot}$, $\boldsymbol{A}_{\cdot j}$, and $\boldsymbol{A}_{ij}$, respectively. The Frobenius norm of matrix $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_F = \left(\sum_{i=1}^n \sum_{j=1}^d \boldsymbol{A}_{ij}^2\right)^{1/2}$. Analogously, the l_2-norm of vector $\boldsymbol{A}_{i\cdot}$ is defined as $\|\boldsymbol{A}_{i\cdot}\|_2 = \left(\sum_{j=1}^d \boldsymbol{A}_{ij}^2\right)^{1/2}$. For a set $\Omega \subset \mathbb{R}^{n\times d}$, let $\mathrm{dist}(\boldsymbol{A}, \Omega) = \inf_{\boldsymbol{B}\in\Omega} \|\boldsymbol{B} - \boldsymbol{A}\|_F$. Furthermore, we use $[d]$ for the set $\{1, 2, \ldots, d\}$. Unless otherwise specified, $\boldsymbol{I}$ represents the identity matrix of the appropriate dimension.

The Kurdyka–Łojasiewicz (KŁ) property plays an important role in proving the strong convergence of ADMM algorithms for non-convex optimization problems. Before introducing the uniform KŁ property, we first establish the necessary notation.

Let $\bar{\boldsymbol{\wp}}$ be the set of concave functions $\wp : [0, \nu) \to \mathbb{R}_+$ that satisfy (i) $\wp(0) = 0$; (ii) $\wp$ is continuously differentiable on $(0, \nu)$ and continuous at 0; (iii) $\wp'(s) > 0, \forall s \in (0, \nu)$. For example, $\wp(s) = cs^{1-\bar{\theta}}$ with $c > 0$, $s \geq 0$, and $\bar{\theta} \in [0, 1)$ [29,30].

Definition 2.1. ([31]) Suppose that $\Phi(\cdot):\mathbb{R}^{n\times d}\to\mathbb{R}$ is a proper lower semicontinuous function, Φ^* is a constant, and Ω is a compact set. If $\Phi(\boldsymbol{A})\equiv\Phi^*$ for $\forall\boldsymbol{A}\in\Omega$ and satisfies the KŁ property at each point of Ω. Then, there exist $\varepsilon>0,\nu>0$, and $\wp\in\bar{\boldsymbol{\wp}}$ such that

$$\wp'(\Phi(\boldsymbol{A})-\Phi^*)\mathrm{dist}(0,\partial\Phi(\boldsymbol{A}))\geq 1,$$

for $\forall\boldsymbol{A}\in\{\mathbb{R}^{n\times d}\mid\mathrm{dist}(\boldsymbol{A},\Omega)<\varepsilon$ and $\Phi^*<\Phi(\boldsymbol{A})<\Phi^*+\nu\}$. Moreover, if $\Phi(\cdot)$ satisfies the KŁ property at each point of $\mathrm{dom}(\partial\Phi(\cdot))$, then $\Phi(\cdot)$ is called a KŁ function.

2.2. Proximal operators

For a proper closed function $\Phi(\cdot):\mathbb{R}^{n\times d}\to\mathbb{R}$, its proximal operator $\mathrm{Prox}_\Phi(\cdot)$ is defined by

$$\mathrm{Prox}_\Phi(\boldsymbol{A})=\arg\min_{\boldsymbol{B}\in\mathbb{R}^{n\times d}}\Big\{\Phi(\boldsymbol{B})+\frac{1}{2}\|\boldsymbol{B}-\boldsymbol{A}\|_F^2\Big\}.$$

Lemma 2.2. *([32]) If $\Phi:\mathbb{R}^{n\times d}\to\mathbb{R}$ is proper closed, convex, and separable,*

$$\Phi(\mathbf{A})=\sum_{i=1}^{n}\Phi_i(\boldsymbol{A}_{i\cdot}),$$

with Φ_i being proper closed and convex univariate functions, then

$$\mathrm{Prox}_\Phi(\mathbf{A})=\big[\,\mathrm{Prox}_{\Phi_i}(\boldsymbol{A}_{i\cdot})\big]_{i=1}^{n}.$$

It follows from the calculus rule in Lemma 2.2 and [32] that the proximal operator of $\lambda\|\cdot\|_{2,1}$ can be computed as

$$\begin{aligned}\mathrm{Prox}_{\lambda\|\cdot\|_{2,1}}(\boldsymbol{A})&=\big[\,\mathrm{Prox}_{\lambda\|\cdot\|_2}(\boldsymbol{A}_{i\cdot})\big]_{i=1}^{n}\\&=\Big[(1-\frac{\lambda}{\max\{\|\boldsymbol{A}_{i\cdot}\|_2,\lambda\}})\boldsymbol{A}_{i\cdot}\Big]_{i=1}^{n},\end{aligned}$$

where $\lambda>0$ and $\lambda\|\boldsymbol{A}\|_{2,1}=\lambda\sum_{i=1}^{n}\|\boldsymbol{A}_{i\cdot}\|_2$.

Moreover, since both the $l_{2,p}$-norm and l_q-norm are intrinsically linear combinations of $|\cdot|^s$ for $s\in[0,1)$, we subsequently present only the proximal operator for $|\cdot|^s$.

For given scalars $\lambda>0$, $a\in\mathbb{R}$, and $s\in[0,1)$,

$$\begin{aligned}\mathrm{Prox}_{\lambda|\cdot|^s}(a)&=\arg\min_{b\in\mathbb{R}}\frac{1}{2}(b-a)^2+\lambda|b|^s\\&=\begin{cases}\{0\}, & |a|<\kappa(\lambda,s)\\ \{0,\mathrm{sgn}(a)f(\lambda,s)\}, & |a|=\kappa(\lambda,s)\\ \{\mathrm{sgn}(a)\varpi_s(|a|)\}, & |a|>\kappa(\lambda,s)\end{cases},\end{aligned}\tag{2.1}$$

where

$$\kappa(\lambda,s)=(2-s)\lambda^{\frac{1}{2-s}}(2(1-s))^{\frac{1-s}{s-2}},\ f(\lambda,s)=(2\lambda(1-s))^{\frac{1}{2-s}},$$

$$\varpi_s(a) \in \left\{b \mid b - a + \lambda s \operatorname{sgn}(b) b^{s-1} = 0, b > 0\right\}, \ \operatorname{sgn}(a) = \begin{cases} 1, & \text{if } a > 0 \\ 0, & \text{if } a = 0 \\ -1, & \text{if } a < 0 \end{cases}.$$

When s takes values of 0, 1/2, or 2/3, closed-form solutions for the corresponding proximal operator can be derived from (2.1). When s takes other values in $[0, 1)$, the efficient algorithms proposed in [33, 34] can be employed. See [33] for further details.

3. Optimality conditions

For better describing optimality conditions, we recast (1.4) as the following equivalent constrained problem:

$$\begin{aligned} &\min_{\boldsymbol{X},\boldsymbol{E},\boldsymbol{F}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{A}\|_F^2 + g(\boldsymbol{E}) + r(\boldsymbol{F}) \\ &\text{s.t.} \quad \boldsymbol{D}\boldsymbol{X} = \boldsymbol{E}, \ \boldsymbol{X} = \boldsymbol{F}, \end{aligned} \tag{3.1}$$

where $g(\boldsymbol{E}) = \alpha \sum_{\iota \in \epsilon} \omega_\iota \|(\boldsymbol{E})_{\iota \cdot}\|_2$ and $r(\boldsymbol{F}) = \gamma((1 - \eta)\|\boldsymbol{F}\boldsymbol{\beta}\|_{2,p}^p + \eta\|\boldsymbol{F}\|_q^q)$.

The Lagrangian function of (3.1), with multipliers $\boldsymbol{P} \in \mathbb{R}^{|\epsilon| \times d}$ and $\boldsymbol{G} \in \mathbb{R}^{n \times d}$, is given by

$$\begin{aligned} &L(\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}; \boldsymbol{P}, \boldsymbol{G}) \\ &= \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{A}\|_F^2 + g(\boldsymbol{E}) + r(\boldsymbol{F}) + \langle \boldsymbol{P}, \boldsymbol{D}\boldsymbol{X} - \boldsymbol{E} \rangle + \langle \boldsymbol{G}, \boldsymbol{X} - \boldsymbol{F} \rangle. \end{aligned}$$

Definition 3.1. A point $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{E}}, \tilde{\boldsymbol{F}})$ is said to be a KKT point of (3.1) provided that there exist matrices $\tilde{\boldsymbol{P}} \in \mathbb{R}^{|\epsilon| \times d}$ and $\tilde{\boldsymbol{G}} \in \mathbb{R}^{n \times d}$ such that

$$\begin{cases} 0 = \tilde{\boldsymbol{X}} - \boldsymbol{A} + \boldsymbol{D}^\top \tilde{\boldsymbol{P}} + \tilde{\boldsymbol{G}}, \\ 0 \in \partial g(\tilde{\boldsymbol{E}}) - \tilde{\boldsymbol{P}}, \\ 0 \in \partial r_1(\tilde{\boldsymbol{F}}) + \partial r_2(\tilde{\boldsymbol{F}}) - \tilde{\boldsymbol{G}}, \\ 0 = \boldsymbol{D}\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{E}}, \\ 0 = \tilde{\boldsymbol{X}} - \tilde{\boldsymbol{F}}, \end{cases}$$

where $r_1(\tilde{\boldsymbol{F}}) = \gamma(1 - \eta)\|\tilde{\boldsymbol{F}}\boldsymbol{\beta}\|_{2,p}^p$ and $r_2(\tilde{\boldsymbol{F}}) = \gamma\eta\|\tilde{\boldsymbol{F}}\|_q^q$.

Note that, in the derivation of KKT conditions, $\partial(r_1(\tilde{\boldsymbol{F}}) + r_2(\tilde{\boldsymbol{F}})) \neq \partial r_1(\tilde{\boldsymbol{F}}) + \partial r_2(\tilde{\boldsymbol{F}})$ for general non-differentiable functions. However, the specific structures of $r_1(\cdot)$ and $r_2(\cdot)$ in this paper admit a particular separability rule for the subdifferential. The following theorem elaborates this rule and further establishes the necessity of the KKT conditions for optimality.

Theorem 3.2. *If* $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{E}}, \tilde{\boldsymbol{F}})$ *is a local minimizer of* (3.1)*, then it is also a KKT point of* (3.1).

Proof. $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{E}}, \tilde{\boldsymbol{F}})$ being a local minimizer of (3.1) implies that $\tilde{\boldsymbol{X}}$ is a local minimizer of (1.4), $\boldsymbol{D}\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{E}}$, as well as $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{F}}$. According to Fermat's rule [35, Theorem 10.1], the optimality condition of (1.4) is

$$0 \in \tilde{\boldsymbol{X}} - \boldsymbol{A} + \partial(g(\tilde{\boldsymbol{X}}) + r(\tilde{\boldsymbol{X}})),$$

where $g(\tilde{\boldsymbol{X}}) = \alpha \sum_{\iota\in\epsilon} w_\iota \|(\tilde{\boldsymbol{X}})_{\iota\cdot}\|_2$ and $r(\tilde{\boldsymbol{X}}) = r_1(\tilde{\boldsymbol{X}}) + r_2(\tilde{\boldsymbol{X}})$.

Due to [35, Definition 9.1 and Exercise 10.10] and the strict continuity of $g(\cdot)$, we have

$$0 \in \tilde{\boldsymbol{X}} - \boldsymbol{A} + \boldsymbol{D}^\top \partial g(\tilde{\boldsymbol{X}}) + \partial r(\tilde{\boldsymbol{X}}). \tag{3.2}$$

Further, we demonstrate the separability of $\partial r(\cdot)$ with respect to $r_1(\cdot)$ and $r_2(\cdot)$.

Define

$$\begin{aligned} I_1 &= \{j \mid \|\tilde{\boldsymbol{X}}_{\cdot j}\|_2 \neq 0, |\mathrm{supp}(\tilde{\boldsymbol{X}}_{\cdot j})| = n\}, \\ I_2 &= \{j \mid \|\tilde{\boldsymbol{X}}_{\cdot j}\|_2 \neq 0, |\mathrm{supp}(\tilde{\boldsymbol{X}}_{\cdot j})| < n\}, \\ I_3 &= \{j \mid \|\tilde{\boldsymbol{X}}_{\cdot j}\|_2 = 0\}, \end{aligned}$$

where $\mathrm{supp}(\tilde{\boldsymbol{X}}_{\cdot j})$ is the index set of nonzero elements in $\tilde{\boldsymbol{X}}_{\cdot j}$, and $|\mathrm{supp}(\tilde{\boldsymbol{X}}_{\cdot j})|$ is the cardinality for the set $\mathrm{supp}(\tilde{\boldsymbol{X}}_{\cdot j})$.

On account of the separability of $r(\tilde{\boldsymbol{X}})$ with respect to features, it follows from [35, Proposition 10.5] that the limiting subdifferentials can be separated into

$$\partial r(\tilde{\boldsymbol{X}}) = \partial r(\tilde{\boldsymbol{X}}_{\cdot 1}) \times \partial r(\tilde{\boldsymbol{X}}_{\cdot 2}) \times \cdots \times \partial r(\tilde{\boldsymbol{X}}_{\cdot d}),$$

and hence $\partial r(\tilde{\boldsymbol{F}}) = \partial r(\tilde{\boldsymbol{X}}_{I_1}) \times \partial r(\tilde{\boldsymbol{X}}_{I_2}) \times \partial r(\tilde{\boldsymbol{X}}_{I_3})$. Next, we analyze the properties of $r(\cdot)$ for these three partitioned matrices: $\tilde{\boldsymbol{X}}_{I_1}$, $\tilde{\boldsymbol{X}}_{I_2}$, and $\tilde{\boldsymbol{X}}_{I_3}$.

For all $j \in I_1$, $r(\boldsymbol{X}_{\cdot j})$ is strictly differentiable. Thus, we obtain from [35, Exercise 10.10] that

$$\partial r(\tilde{\boldsymbol{X}}_{\cdot j}) = \nabla r_1(\tilde{\boldsymbol{X}}_{\cdot j}) + \nabla r_2(\tilde{\boldsymbol{X}}_{\cdot j}),\ \ j \in I_1. \tag{3.3}$$

For all $j \in I_2$, $\|\boldsymbol{X}_{\cdot j}\|_{2,p}^p$ is strictly differentiable; we know that

$$\partial r(\tilde{\boldsymbol{X}}_{\cdot j}) = \nabla r_1(\tilde{\boldsymbol{X}}_{\cdot j}) + \partial r_2(\tilde{\boldsymbol{X}}_{\cdot j}). \tag{3.4}$$

For all $j \in I_3$, $\partial r_1(\tilde{\boldsymbol{X}}_{\cdot j}) = \partial r_2(\tilde{\boldsymbol{X}}_{\cdot j}) = \mathbb{R}^n$, we have

$$\partial r(\tilde{\boldsymbol{X}}_{\cdot j}) \subset \partial r_1(\tilde{\boldsymbol{X}}_{\cdot j}) + \partial r_2(\tilde{\boldsymbol{X}}_{\cdot j}) = \mathbb{R}^n. \tag{3.5}$$

Combined with (3.2)–(3.5), we conclude that

$$0 \in \tilde{\boldsymbol{X}} - \boldsymbol{A} + \boldsymbol{D}^\top \partial g(\tilde{\boldsymbol{X}}) + \partial r_1(\tilde{\boldsymbol{X}}) + \partial r_2(\tilde{\boldsymbol{X}}).$$

It follows from Definition 3.1 and the lower semicontinuity of $g(\cdot)$ and $r(\cdot)$ that there exist matrices $\tilde{\boldsymbol{P}} \in \partial g(\tilde{\boldsymbol{E}})$ and $\tilde{\boldsymbol{G}} \in \partial r(\tilde{\boldsymbol{F}})$ such that $0 = \tilde{\boldsymbol{X}} - \boldsymbol{A} + \boldsymbol{D}^\top \tilde{\boldsymbol{P}} + \tilde{\boldsymbol{G}}$ holds. Therefore, $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{E}}, \tilde{\boldsymbol{F}})$ is a KKT point. □

4. Optimization algorithm

This section presents an ADMM algorithm for solving (3.1), along with an analysis of its global convergence and computational complexity.

4.1. Algorithm design

For ease of solving, we introduce a slack variable $\bar{\boldsymbol{F}} \in \mathbb{R}^{n\times d}$ such that (3.1) can be equivalently transformed to

$$\begin{aligned} &\min_{\boldsymbol{X},\boldsymbol{E},\boldsymbol{F},\bar{\boldsymbol{F}}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{A}\|_F^2 + g(\boldsymbol{E}) + r_1(\boldsymbol{F}) + r_2(\bar{\boldsymbol{F}}) \\ &\text{s.t.} \quad \boldsymbol{DX}=\boldsymbol{E},\ \boldsymbol{X}=\boldsymbol{F},\ \boldsymbol{X}=\bar{\boldsymbol{F}}. \end{aligned} \tag{4.1}$$

Given $\tau > 0$, the augmented Lagrangian function of (4.1) is

$$\begin{aligned} &\mathcal{L}_\tau(\boldsymbol{X},\boldsymbol{E},\boldsymbol{F},\bar{\boldsymbol{F}};\boldsymbol{P},\boldsymbol{G},\bar{\boldsymbol{G}}) \\ &= \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{A}\|_F^2 + g(\boldsymbol{E}) + r_1(\boldsymbol{F}) + r_2(\bar{\boldsymbol{F}}) + \langle \boldsymbol{P}, \boldsymbol{DX}-\boldsymbol{E}\rangle + \frac{\tau}{2}\|\boldsymbol{DX}-\boldsymbol{E}\|_F^2 \\ &+ \langle \boldsymbol{G}, \boldsymbol{X}-\boldsymbol{F}\rangle + \frac{\tau}{2}\|\boldsymbol{X}-\boldsymbol{F}\|_F^2 + \langle \bar{\boldsymbol{G}}, \boldsymbol{X}-\bar{\boldsymbol{F}}\rangle + \frac{\tau}{2}\|\boldsymbol{X}-\bar{\boldsymbol{F}}\|_F^2, \end{aligned} \tag{4.2}$$

where multiplier $\bar{\boldsymbol{G}} \in \mathbb{R}^{n\times d}$.

As stated in [36], the convergence of multi-block ADMM cannot be guaranteed. Fortunately, (1.4) admits a parallel 2-block ADMM implementation through the following variable splitting

$$\boldsymbol{\Upsilon}_1 = (\boldsymbol{E},\boldsymbol{F},\bar{\boldsymbol{F}}) \text{ and } \boldsymbol{\Upsilon}_2 = (\boldsymbol{P},\boldsymbol{G},\bar{\boldsymbol{G}}).$$

Then, we alternately minimize the primal variables in $\boldsymbol{\Upsilon}_1$ and $\boldsymbol{X}$, followed by multiplier $\boldsymbol{\Upsilon}_2$ updates. The ordering of $\boldsymbol{E}$, $\boldsymbol{F}$, and $\bar{\boldsymbol{F}}$ within block $\boldsymbol{\Upsilon}_1$ is arbitrary and does not affect ADMM accuracy.

At the t-th iteration with the current point $(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{\Upsilon}_2^{(t)})$, the next iterate $(\boldsymbol{\Upsilon}_1^{(t+1)}, \boldsymbol{X}^{(t+1)}, \boldsymbol{\Upsilon}_2^{(t+1)})$ is generated below.

1) Update $\boldsymbol{E}^{(t+1)}$. The $\boldsymbol{E}$-subproblem is updated via

$$\min_{\boldsymbol{E}} \ g(\boldsymbol{E}) + \frac{\tau}{2}\|\boldsymbol{DX}^{(t)} - \boldsymbol{E} + \frac{1}{\tau}\boldsymbol{P}^{(t)}\|_F^2,$$

with a closed-form solution

$$\boldsymbol{E}^{(t+1)} \in \text{Prox}_{g/\tau}(\boldsymbol{DX}^{(t)} + \frac{1}{\tau}\boldsymbol{P}^{(t)}). \tag{4.3}$$

2) Update $\boldsymbol{F}^{(t+1)}$. The $\boldsymbol{F}$-subproblem can be formulated using (4.1) and (4.2) as

$$\min_{\boldsymbol{F}} \ r_1(\boldsymbol{F}) + \frac{\tau}{2}\|\boldsymbol{X}^{(t)} - \boldsymbol{F} + \frac{1}{\tau}\boldsymbol{G}^{(t)}\|_F^2,$$

yielding the solution

$$\boldsymbol{F}^{(t+1)} \in \text{Prox}_{r_1/\tau}(\boldsymbol{X}^{(t)} + \frac{1}{\tau}\boldsymbol{G}^{(t)}). \tag{4.4}$$

3) Update $\bar{\boldsymbol{F}}^{(t+1)}$. The optimization problem (4.1) concerning $\bar{\boldsymbol{F}}$ can be simplified to

$$\min_{\bar{\boldsymbol{F}}} \ r_2(\bar{\boldsymbol{F}}) + \frac{\tau}{2}\|\boldsymbol{X}^{(t)} - \bar{\boldsymbol{F}} + \frac{1}{\tau}\bar{\boldsymbol{G}}^{(t)}\|_F^2.$$

Algorithm 1 ADMM for solving (3.1)

```
Initialization: Choose $(\boldsymbol{\Upsilon}_1^{(0)}, \boldsymbol{X}^{(0)}, \boldsymbol{\Upsilon}_2^{(0)})$ and $\tau > 0$
while not converged do
    Step 1. Compute $\boldsymbol{\Upsilon}_1^{(t+1)}$:
        Update $\boldsymbol{E}^{(t+1)}$ by (4.3)
        Update $\boldsymbol{F}^{(t+1)}$ by (4.4)
        Update $\bar{\boldsymbol{F}}^{(t+1)}$ by (4.5)
    Step 2. Compute $\boldsymbol{X}^{(t+1)}$:
        Update $\boldsymbol{X}^{(t+1)}$ by (4.6)
    Step 3. Compute $\boldsymbol{\Upsilon}_2^{(t+1)}$:
        Update $\boldsymbol{P}^{(t+1)}, \boldsymbol{G}^{(t+1)}, \bar{\boldsymbol{G}}^{(t+1)}$ by (4.7)
end while
Output: $(\boldsymbol{\Upsilon}_1^{(t+1)}, \boldsymbol{X}^{(t+1)}, \boldsymbol{\Upsilon}_2^{(t+1)})$
```

The corresponding solution takes the form

$$\bar{\boldsymbol{F}}^{(t+1)} \in \mathrm{Prox}_{r_2/\tau}(\boldsymbol{X}^{(t)} + \frac{1}{\tau}\bar{\boldsymbol{G}}^{(t)}). \tag{4.5}$$

4) Update $\mathbf{X}^{(t+1)}$. We derive the closed-form solution from the $\boldsymbol{X}$-subproblem

$$\begin{aligned}\min_{\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{A}\|_F^2 + \frac{\tau}{2}\|\boldsymbol{D}\boldsymbol{X} - \boldsymbol{E}^{(t+1)} + \frac{1}{\tau}\boldsymbol{P}^{(t)}\|_F^2 \\ + \frac{\tau}{2}\|\boldsymbol{X} - \boldsymbol{F}^{(t+1)} + \frac{1}{\tau}\boldsymbol{G}^{(t)}\|_F^2 + \frac{\tau}{2}\|\boldsymbol{X} - \bar{\boldsymbol{F}}^{(t+1)} + \frac{1}{\tau}\bar{\boldsymbol{G}}^{(t)}\|_F^2\end{aligned}$$

as

$$\begin{aligned}\boldsymbol{X}^{(t+1)} = ((2\tau + 1)\boldsymbol{I} + \tau\boldsymbol{D}^\top\boldsymbol{D})^{-1}(\boldsymbol{A} + \tau\boldsymbol{D}^\top\boldsymbol{E}^{(t+1)} - \boldsymbol{D}^\top\boldsymbol{P}^{(t)} \\ + \tau\boldsymbol{F}^{(t+1)} - \boldsymbol{G}^{(t)} + \tau\bar{\boldsymbol{F}}^{(t+1)} - \bar{\boldsymbol{G}}^{(t)}).\end{aligned} \tag{4.6}$$

5) Update multipliers. The multipliers in (4.2) are updated by

$$\begin{aligned}\boldsymbol{P}^{(t+1)} &= \boldsymbol{P}^{(t)} + \tau(\boldsymbol{D}\boldsymbol{X}^{(t+1)} - \boldsymbol{E}^{(t+1)}), \\ \boldsymbol{G}^{(t+1)} &= \boldsymbol{G}^{(t)} + \tau(\boldsymbol{X}^{(t+1)} - \boldsymbol{F}^{(t+1)}), \\ \bar{\boldsymbol{G}}^{(t+1)} &= \bar{\boldsymbol{G}}^{(t)} + \tau(\boldsymbol{X}^{(t+1)} - \bar{\boldsymbol{F}}^{(t+1)}).\end{aligned} \tag{4.7}$$

For the specific forms of the solutions to (4.3)–(4.5), the reader is referred to Subsection 2.2. The complete iterative steps are summarized in Algorithm 1.

4.2. Convergence analysis

To analyze convergence, we construct a potential function inspired by [37] as

$$\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1, \boldsymbol{X}; \boldsymbol{\Upsilon}_2, \varepsilon^{(t)}) = \mathcal{L}_\tau(\boldsymbol{\Upsilon}_1, \boldsymbol{X}; \boldsymbol{\Upsilon}_2) + \varepsilon^{(t)}, \tag{4.8}$$

where $\{\varepsilon^{(t)}\}_{t=1}^{+\infty}$ is a positive and monotonically decreasing sequence satisfying $\varepsilon^{(t)} \to 0$ and $\sum_{t=1}^{+\infty} \sqrt{\varepsilon^{(t)} - \varepsilon^{(t+1)}} < +\infty$ as $t \to +\infty$. A common and concrete choice for such a perturbation sequence is $\varepsilon^{(t)} = c\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F^2$ with $c > 0$ [31].

It is well-known that potential functions are commonly constructed for handling non-convex and non-smooth optimization problems [31,38]. By introducing a controllable perturbation term $\varepsilon^{(t)}$ to the augmented Lagrangian function $\mathcal{L}_\tau(\cdot)$, we transform the convergence analysis of the original problem (4.1) into that of a potential function exhibiting monotonic descent and satisfying the KŁ property. Note that $\mathcal{L}_\tau(\cdot)$ is not required to be monotonically decreasing. Unlike conventional non-convex non-smooth optimization problems [31,38–40], the coefficient matrix $\boldsymbol{D}$ in our **X**-subproblem is neither row-full-rank nor column-full-rank. Thus, the following assumption is necessary to bound the incremental change of multipliers.

Assumption 4.1. The sequence of multipliers $\{\boldsymbol{\Upsilon}_2^{(t)}\}_{t=1}^{+\infty}$ is bounded and satisfies

$$\|\boldsymbol{\Upsilon}_2^{(t+1)} - \boldsymbol{\Upsilon}_2^{(t)}\|_F^2 \leq \theta\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F^2 + \tau(\varepsilon^{(t)} - \varepsilon^{(t+1)}),$$

where $\theta \in [0, \frac{2\tau^2+\tau}{2})$.

The reasonableness of Assumption 4.1 is readily verifiable, as demonstrated in Subsection 5.4. Next, we give the descent property of the potential function $\hat{\mathcal{L}}_\tau(\cdot)$ and the boundedness of variables $(\boldsymbol{\Upsilon}_1, \boldsymbol{X}; \boldsymbol{\Upsilon}_2)$ that allows the convergence of $\{\hat{\mathcal{L}}_\tau(\cdot)\}$ and the vanishing of $\left\|(\boldsymbol{\Upsilon}_1^{(t+1)}, \boldsymbol{X}^{(t+1)}; \boldsymbol{\Upsilon}_2^{(t+1)}) - (\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)})\right\|_F$. For ease of reading, the detailed proof of the following two lemmas can be found in the Supplementary Material (Section II and Section III).

Lemma 4.2. *Let $\{(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{\Upsilon}_2^{(t)})\}_{t=1}^{\infty}$ be the sequence generated by Algorithm 1. If Assumption 4.1 holds, then*

1) $\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t+1)}, \boldsymbol{X}^{(t+1)}; \boldsymbol{\Upsilon}_2^{(t+1)}, \varepsilon^{(t+1)}) < \hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)})$;

2) $\mathrm{dist}(0, \partial\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)})) \leq \tilde{w}(\|\boldsymbol{X}^{(t)} - \boldsymbol{X}^{(t-1)}\|_F + \sqrt{\varepsilon^{(t-1)} - \varepsilon^{(t)}})$, where $\tilde{w} = \max\{2\tau + \tau\sqrt{\lambda_{max}(\boldsymbol{D}^\top\boldsymbol{D})} + 1 + \frac{\sqrt{\theta}}{\tau} + \sqrt{\theta}, \frac{\sqrt{\tau}}{\tau} + \sqrt{\tau}\}$ and $\lambda_{max}(\boldsymbol{D}^\top\boldsymbol{D})$ denotes the largest eigenvalue of matrix $\boldsymbol{D}^\top\boldsymbol{D}$.

Lemma 4.2 establishes both the descent property and the subgradient boundedness of $\hat{\mathcal{L}}_\tau(\cdot)$, which lays the groundwork for the convergence proof of $\{(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{\Upsilon}_2^{(t)})\}$.

Lemma 4.3. *Under Assumption 4.1, the following results hold:*

1) The sequences $\{\boldsymbol{\Upsilon}_1^{(t)}\}_{t=1}^{+\infty}$ and $\{\boldsymbol{X}^{(t)}\}_{t=1}^{+\infty}$ are bounded;

2) $\lim\limits_{t\to+\infty} \left\|(\boldsymbol{\Upsilon}_1^{(t+1)}, \boldsymbol{X}^{(t+1)}; \boldsymbol{\Upsilon}_2^{(t+1)}) - (\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)})\right\|_F = 0.$

From [41], it is evident that $\hat{\mathcal{L}}_\tau(\cdot)$ is a KŁ function, thus the global convergence of $\{\boldsymbol{\Upsilon}^{(t)}\}_{t=1}^{+\infty}$ can be established.

Theorem 4.4. *Suppose that Assumption 4.1 is satisfied. Then,*

1) Any accumulation point $\tilde{\boldsymbol{\Upsilon}}$ of the generated sequence $\{(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{\Upsilon}_2^{(t)})\}_{t=1}^{+\infty}$ is a KKT point of (3.1)*;*

2) Because $\hat{\mathcal{L}}_\tau(\cdot)$ is a KŁ function, the sequence $\{(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{\Upsilon}_2^{(t)})\}_{t=1}^{+\infty}$ converges to a KKT point of (3.1).

Proof. 1) Lemma 4.3 suggests $(\boldsymbol{\Upsilon}_1^{(t_j+1)}, \boldsymbol{X}^{(t_j+1)}; \boldsymbol{\Upsilon}_2^{(t_j+1)}) \to \tilde{\boldsymbol{\Upsilon}}$ as $j \to +\infty$. By virtue of the lower semi-continuity of $\mathcal{L}_\tau(\cdot)$, we get

$$\begin{aligned}
&\liminf_{j\to+\infty} \mathcal{L}_\tau(\boldsymbol{\Upsilon}_1^{(t_j+1)}, \boldsymbol{X}^{(t_j+1)}; \boldsymbol{\Upsilon}_2^{(t_j+1)}) \geq \mathcal{L}_\tau(\tilde{\boldsymbol{\Upsilon}}) \\
&= \frac{1}{2}\|\tilde{\boldsymbol{X}} - \boldsymbol{A}\|_F^2 + g(\tilde{\boldsymbol{E}}) + r_1(\tilde{\boldsymbol{F}}) + r_2(\tilde{\bar{\boldsymbol{F}}}) \\
&+ \frac{\tau}{2}\|\boldsymbol{D}\tilde{\boldsymbol{X}} - \boldsymbol{E}^{(t)} + \frac{1}{\tau}\tilde{\boldsymbol{P}}\|_F^2 + \frac{\tau}{2}\|\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{F}} + \frac{1}{\tau}\tilde{\boldsymbol{G}}\|_F^2 \\
&+ \frac{\tau}{2}\|\tilde{\boldsymbol{X}} - \tilde{\bar{\boldsymbol{F}}} + \frac{1}{\tau}\tilde{\bar{\boldsymbol{G}}}\|_F^2 - \frac{1}{2\tau}\|\tilde{\boldsymbol{\Upsilon}}_2\|_F^2.
\end{aligned} \tag{4.9}$$

Since $\boldsymbol{\Upsilon}_1^{(t_j+1)}$ is the minimizer of the $\boldsymbol{\Upsilon}_1$-subproblem, it holds

$$\begin{aligned}
&\mathcal{L}_\tau(\boldsymbol{\Upsilon}_1^{(t_j+1)}, \boldsymbol{X}^{(t_j)}; \boldsymbol{\Upsilon}_2^{(t_j)}) \leq \mathcal{L}_\tau(\tilde{\boldsymbol{\Upsilon}}_1, \boldsymbol{X}^{(t_j)}; \boldsymbol{\Upsilon}_2^{(t_j)}) \\
&= \frac{1}{2}\|\boldsymbol{X}^{(t_j)} - \boldsymbol{A}\|_F^2 + g(\tilde{\boldsymbol{E}}) + r_1(\tilde{\boldsymbol{F}}) + r_2(\tilde{\bar{\boldsymbol{F}}}) \\
&+ \frac{\tau}{2}\|\boldsymbol{D}\boldsymbol{X}^{(t_j)} - \boldsymbol{E}^{(t)} + \frac{1}{\tau}\boldsymbol{P}^{(t_j)}\|_F^2 + \frac{\tau}{2}\|\boldsymbol{X}^{(t_j)} - \tilde{\boldsymbol{F}} + \frac{1}{\tau}\boldsymbol{G}^{(t_j)}\|_F^2 \\
&+ \frac{\tau}{2}\|\boldsymbol{X}^{(t_j)} - \tilde{\bar{\boldsymbol{F}}} + \frac{1}{\tau}\bar{\boldsymbol{G}}^{(t_j)}\|_F^2 - \frac{1}{2\tau}\|\boldsymbol{\Upsilon}_2^{(t_j)}\|_F^2.
\end{aligned}$$

Based on the continuity of $\|\cdot\|_F^2$, we find that

$$\limsup_{j\to+\infty} \mathcal{L}_\tau(\boldsymbol{\Upsilon}_1^{(t_j+1)}, \boldsymbol{X}^{(t_j)}; \boldsymbol{\Upsilon}_2^{(t_j)}) \leq \mathcal{L}_\tau(\tilde{\boldsymbol{\Upsilon}}). \tag{4.10}$$

Considering both (4.9) and (4.10), we can deduce

$$\begin{aligned}
&\lim_{j\to+\infty} \mathcal{L}_\tau(\boldsymbol{\Upsilon}_1^{(t_j+1)}, \boldsymbol{X}^{(t_j)}; \boldsymbol{\Upsilon}_2^{(t_j)}) \\
&= \frac{1}{2}\|\tilde{\boldsymbol{X}} - \boldsymbol{A}\|_F^2 + g(\tilde{\boldsymbol{E}}) + r_1(\tilde{\boldsymbol{F}}) + r_2(\tilde{\bar{\boldsymbol{F}}}) \\
&+ \frac{\tau}{2}\|\boldsymbol{D}\tilde{\boldsymbol{X}} - \boldsymbol{E}^{(t)} + \frac{1}{\tau}\tilde{\boldsymbol{P}}\|_F^2 + \frac{\tau}{2}\|\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{F}} + \frac{1}{\tau}\tilde{\boldsymbol{G}}\|_F^2 \\
&+ \frac{\tau}{2}\|\tilde{\boldsymbol{X}} - \tilde{\bar{\boldsymbol{F}}} + \frac{1}{\tau}\tilde{\bar{\boldsymbol{G}}}\|_F^2 - \frac{1}{2\tau}\|\tilde{\boldsymbol{\Upsilon}}_2\|_F^2.
\end{aligned}$$

Further, it is easy to derive from the continuity of $g(\cdot)$ that

$$\lim_{j\to+\infty} r_1(\boldsymbol{F}^{(t_j+1)}) + r_2(\bar{\boldsymbol{F}}^{(t_j+1)}) = r_1(\tilde{\boldsymbol{F}}) + r_2(\tilde{\bar{\boldsymbol{F}}}). \tag{4.11}$$

Therefore, taking the limit in (6.5) of the Supplementary Material along subsequence $\{(\boldsymbol{\Upsilon}_1^{(t_j)}, \boldsymbol{X}^{(t_j)}, \boldsymbol{\Upsilon}_2^{(t_j)})\}$ for $j \to +\infty$, and using (4.11) together with Lemma 4.3, we obtain

$$\begin{cases}
0 = \tilde{\boldsymbol{X}} - \boldsymbol{A} + \boldsymbol{D}^\top \tilde{\boldsymbol{P}} + \tilde{\boldsymbol{G}}, \\
0 \in \partial g(\tilde{\boldsymbol{E}}) - \tilde{\boldsymbol{P}}, \\
0 \in \partial(r_1(\tilde{\boldsymbol{F}}) + r_2(\tilde{\boldsymbol{F}})) - \tilde{\boldsymbol{G}}, \\
0 = \boldsymbol{D}\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{E}}, \\
0 = \tilde{\boldsymbol{X}} - \tilde{\boldsymbol{F}},
\end{cases}$$

which implies that $\tilde{\boldsymbol{\Upsilon}} = (\tilde{\boldsymbol{E}}, \tilde{\boldsymbol{F}}, \tilde{\tilde{\boldsymbol{F}}}, \tilde{\boldsymbol{X}}; \tilde{\boldsymbol{P}}, \tilde{\boldsymbol{G}}, \tilde{\tilde{\boldsymbol{G}}})$ is a KKT point in the sense of Definition 3.1.

2) On account of the Assumption 4.1 and Lemma 4.2, we know that $\{\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)})\}_{t=1}^{+\infty}$ generated by the proposed Algorithm 1 is bounded from below and nonincreasing. So, there exists a constant $\tilde{l}$ such that

$$\lim_{t\to+\infty} \hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)}) = \lim_{t\to+\infty} \mathcal{L}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}) = \tilde{l}. \tag{4.12}$$

Let $\boldsymbol{\Gamma}$ denote the set of cluster points. Take $\forall \tilde{\boldsymbol{\Upsilon}} \in \boldsymbol{\Gamma}$ and consider a converged subsequence satisfying (6.9) in the Supplementary Material, it follows from (4.9), (4.10), and (4.12) that

$$\begin{aligned}
\tilde{l} &= \liminf_{j\to+\infty} \hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t_j+1)}, \boldsymbol{X}^{(t_j)}; \boldsymbol{\Upsilon}_2^{(t_j)}, \varepsilon^{(t_j)}) \geq \hat{\mathcal{L}}_\tau(\tilde{\boldsymbol{\Upsilon}}_1, \tilde{\boldsymbol{X}}; \tilde{\boldsymbol{\Upsilon}}_2),\\
\tilde{l} &= \limsup_{j\to+\infty} \hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t_j+1)}, \boldsymbol{X}^{(t_j)}; \boldsymbol{\Upsilon}_2^{(t_j)}, \varepsilon^{(t_j)}) \leq \hat{\mathcal{L}}_\tau(\tilde{\boldsymbol{\Upsilon}}_1, \tilde{\boldsymbol{X}}; \tilde{\boldsymbol{\Upsilon}}_2),
\end{aligned}$$

which shows that $\hat{\mathcal{L}}_\tau(\tilde{\boldsymbol{\Upsilon}}_1, \tilde{\boldsymbol{X}}; \tilde{\boldsymbol{\Upsilon}}_2) = \tilde{l}$. Because of the arbitrariness of $\tilde{\boldsymbol{\Upsilon}}$ in set $\boldsymbol{\Gamma}$, we further infer that $\hat{\mathcal{L}}_\tau(\cdot)$ is constant on $\boldsymbol{\Gamma}$. And, for given $l_1 > 0$, there exists $\mathrm{t}_1 > 0$ such that $\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)}) < \tilde{l} + l_1$ for $\forall t > \mathrm{t}_1$. Simultaneously, $\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)}) > \tilde{l}$ for $\forall t > 1$. In view of the definition of $\boldsymbol{\Gamma}$, there exist $\mathrm{t}_2 > 0$ and $\rho > 0$ such that $\mathrm{dist}((\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}), \boldsymbol{\Gamma}) < \rho$ for $\forall t > \mathrm{t}_2$. Therefore, for KL function $\hat{\mathcal{L}}_\tau(\cdot)$, it follows Definition 2.1 that for $\forall t > \bar{\mathrm{t}}$, there exists $\wp \in \bar{\wp}$ such that

$$\wp'(\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)}) - \tilde{l}\,)\mathrm{dist}(0, \hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)})) \geq 1,$$

where $\bar{\mathrm{t}} = \max\{\mathrm{t}_1, \mathrm{t}_2\}$.

Let

$$\Delta_{t,t+1} = \wp(\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)}) - \tilde{l}\,) - \wp(\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t+1)}, \boldsymbol{X}^{(t+1)}; \boldsymbol{\Upsilon}_2^{(t+1)}, \varepsilon^{(t+1)}) - \tilde{l}\,).$$

Using Lemma 4.2 and the concavity of $\wp(\cdot)$, we get

$$\begin{aligned}
&\tilde{w}(\|\mathbf{X}^{(t)} - \mathbf{X}^{(t-1)}\|_F + \sqrt{\varepsilon^{(t-1)} - \varepsilon^{(t)}})\Delta_{t,t+1}\\
&\geq \mathrm{dist}(0, \partial\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)})) \times \wp'(\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)}) - \tilde{l}\,)\\
&\times (\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)}, \varepsilon^{(t)}) - \hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(t+1)}, \boldsymbol{X}^{(t+1)}; \boldsymbol{\Upsilon}_2^{(t+1)}, \varepsilon^{(t+1)}))\\
&\geq \bar{w}\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F^2,
\end{aligned} \tag{4.13}$$

from which we can see

$$\begin{aligned}
2\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F &= 2\sqrt{\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F^2}\\
&\leq 2\sqrt{\frac{\tilde{w}}{\bar{w}}(\|\mathbf{X}^{(t)} - \mathbf{X}^{(t-1)}\|_F + \sqrt{\varepsilon^{(t-1)} - \varepsilon^{(t)}})\Delta_{t,t+1}}\\
&\leq \|\mathbf{X}^{(t)} - \mathbf{X}^{(t-1)}\|_F + \sqrt{\varepsilon^{(t-1)} - \varepsilon^{(t)}} + \frac{\tilde{w}}{\bar{w}}\Delta_{t,t+1},
\end{aligned}$$

where the first "≤" holds via (4.13) and the second "≤" holds via $2\sqrt{ab} \leq a + b$ when $a \geq 0$ and $b \geq 0$.

Then,

$$\begin{aligned}
&\sum_{t=\bar{\mathrm{t}}}^{+\infty} \|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F \\
&\leq \|\mathbf{X}^{(\bar{\mathrm{t}})} - \mathbf{X}^{(\bar{\mathrm{t}}-1)}\|_F + \sum_{t=\bar{\mathrm{t}}}^{+\infty} \sqrt{\varepsilon^{(t-1)} - \varepsilon^{(t)}} + \frac{\tilde{w}}{\bar{w}}\wp(\hat{\mathcal{L}}_\tau(\boldsymbol{\Upsilon}_1^{(\bar{\mathrm{t}})}, \boldsymbol{X}^{(\bar{\mathrm{t}})}; \boldsymbol{\Upsilon}_2^{(\bar{\mathrm{t}})}, \varepsilon^{(\bar{\mathrm{t}})}) - \tilde{l}\,) \\
&< +\infty.
\end{aligned} \tag{4.14}$$

Besides, it follows from Assumption 4.1, (4.7), (4.14), and $a^2 + b^2 \leq (a+b)^2$ $(a \geq 0, b \geq 0)$ that

$$\begin{aligned}
&\sum_{t=\bar{\mathrm{t}}}^{+\infty} \|\boldsymbol{\Upsilon}_2^{(t+1)} - \boldsymbol{\Upsilon}_2^{(t)}\|_F \\
&\leq \sum_{t=\bar{\mathrm{t}}}^{+\infty} (\,\sqrt{\theta}\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F + \sqrt{\tau(\varepsilon^{(t)} - \varepsilon^{(t+1)})}) \\
&< +\infty
\end{aligned} \tag{4.15}$$

and

$$\begin{aligned}
&\sum_{t=\bar{\mathrm{t}}}^{+\infty} \|\boldsymbol{\Upsilon}_1^{(t+1)} - \boldsymbol{\Upsilon}_1^{(t)}\|_F \\
&\leq \sum_{t=\bar{\mathrm{t}}}^{+\infty} (\frac{1}{\tau}\|\boldsymbol{\Upsilon}_2^{(t+1)} - \boldsymbol{\Upsilon}_2^{(t)}\|_F + \frac{1}{\tau}\|\boldsymbol{\Upsilon}_2^{(t)} - \boldsymbol{\Upsilon}_2^{(t-1)}\|_F \\
&+ \|\boldsymbol{D}\boldsymbol{X}^{(t+1)} - \boldsymbol{D}\boldsymbol{X}^{(t)}\|_F + 2\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F) \\
&\leq \frac{2}{\tau}\sum_{t=\bar{\mathrm{t}}}^{+\infty} \|\boldsymbol{\Upsilon}_2^{(t+1)} - \boldsymbol{\Upsilon}_2^{(t)}\|_F + (2 + \sqrt{\lambda_{max}(\boldsymbol{D}^\top \boldsymbol{D})}) \sum_{t=\bar{\mathrm{t}}}^{+\infty} \|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F \\
&+ \frac{1}{\tau}\|\boldsymbol{\Upsilon_2}^{(\bar{\mathrm{t}})} - \boldsymbol{\Upsilon_2}^{(\bar{\mathrm{t}}-1)}\|_F \\
&< +\infty.
\end{aligned} \tag{4.16}$$

By combining (4.14)–(4.16) for the fixed $\bar{\mathrm{t}}$, we deduce

$$\begin{aligned}
&\sum_{t=\bar{\mathrm{t}}}^{+\infty} \left\|(\boldsymbol{\Upsilon}_1^{(t+1)}, \boldsymbol{X}^{(t+1)}; \boldsymbol{\Upsilon}_2^{(t+1)}) - (\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}; \boldsymbol{\Upsilon}_2^{(t)})\right\|_F \\
&\leq \sum_{t=\bar{\mathrm{t}}}^{+\infty} (\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F + \|\boldsymbol{\Upsilon}_2^{(t+1)} - \boldsymbol{\Upsilon}_2^{(t)}\|_F + \|\boldsymbol{\Upsilon}_1^{(t+1)} - \boldsymbol{\Upsilon}_1^{(t)}\|_F) \\
&< +\,\infty,
\end{aligned}$$

which implies that $\{(\boldsymbol{\Upsilon}_1^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{\Upsilon}_2^{(t)})\}_{t=1}^{+\infty}$ is a Cauchy sequence and hence converges to a KKT point. □

4.3. Complexity analysis

The complexity primarily stems from the update of original variables and multipliers. As previously defined, n, d, $|\epsilon|$, and k denote the number of observations, features, edges connecting observations, and nearest neighbors per observation, respectively. The specific choice of k in numerical experiments leads to $|\epsilon| > n$.

- Time Complexity: The update process of the original variables focuses on $\boldsymbol{E}$, $\boldsymbol{F}$, $\bar{\boldsymbol{F}}$, and $\boldsymbol{X}$, with corresponding complexities of $O(nd)$, $O(nd)$, $O(nd)$, $O(|\epsilon|d)$, and $O(|\epsilon|d + n^3 + n^2d)$. Here, for updating $\boldsymbol{E}$, $\boldsymbol{F}$, and $\bar{\boldsymbol{F}}$, the time cost is dominated by matrix addition and subtraction. Updating $\boldsymbol{E}$ requires computing $\boldsymbol{DX} = [\boldsymbol{D}_{ii'}(\boldsymbol{X}_{i\cdot} - \boldsymbol{X}_{i'\cdot})]_{(i,i')\in\epsilon} \in \mathbb{R}^{|\epsilon|\times d}$, which involves $|\epsilon|$ vector subtractions. The cost of updating $\boldsymbol{X}$ is highly dependent on computing matrix inverse $((2\tau + 1)\boldsymbol{I} + \boldsymbol{D}^\top\boldsymbol{D})^{-1}$ and matrix multiplication. Similarly, for the multipliers $\boldsymbol{P}$, $\boldsymbol{G}$, and $\bar{\boldsymbol{G}}$, the costs are $O(nd)$, $O(nd)$, $O(nd)$, and $O(|\epsilon|d)$. Overall, the time complexity is $O(T(|\epsilon|d + n^3 + n^2d))$, where T represents the iteration count of Algorithm 1.

- Space Complexity: The dominant space cost comes from storing matrices $\boldsymbol{E} \in \mathbb{R}^{|\epsilon|\times d}$, $\boldsymbol{D} \in \mathbb{R}^{|\epsilon|\times n}$, and $\boldsymbol{P} \in \mathbb{R}^{|\epsilon|\times d}$ during the updates of $\boldsymbol{E}$, $\boldsymbol{U}$, and $\boldsymbol{P}$. Thus, the space complexity is $O(|\epsilon|n + |\epsilon|d)$.

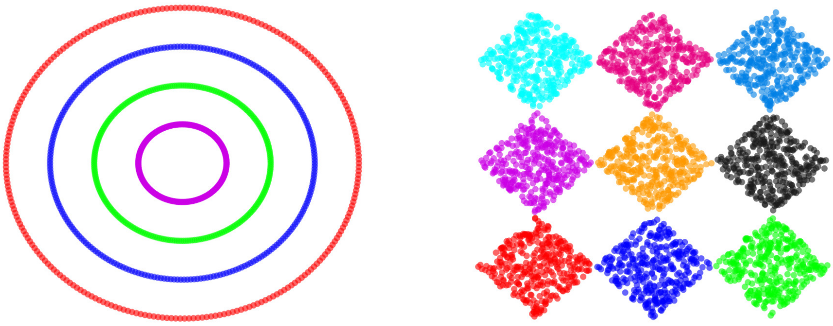
We would like to point out that compared to the clustering method Gecco+ [5], our approach has the same computational complexity.

5. Numerical experiments

For performance evaluation, we compare the proposed GSFRC with conventional FRC approaches on both simulated and real-world datasets. Subsection 5.1 outlines the experimental setup, including dataset description, compared methods, parameter configurations, and evaluation metrics. Subsection 5.2 presents numerical results on simulated and real datasets. Subsection 5.3 provides ablation studies, and Subsection 5.4 analyzes model stability, parameter sensitivity, and convergence behavior.

Table 1. Details of synthetic and real-world datasets.

Type	Datasets	Clusters	Features	Observations
Synthetic datasets	Dartboard1	4	10	1000
	Diamond9	9	30	3000
Real-world datasets	Authors	4	69	841
	Lung-discrete	7	325	73
	GLIOMA	4	4434	50
	Brain	5	5597	42



(a) Dartboard1 (b) Diamond9

Figure 2. Visualization of simulated datasets without noise.

5.1. Experimental setup

1) Dataset description. We evaluate the performance of GSFRC using six benchmark datasets listed in Table 1. Here, Dartboard1 and Diamond9 [20] are simulated datasets where the first two features are informative and the remaining ones consist of Gaussian noise. The noise-free versions of these datasets are visualized in Figure 2. The remaining four are real-world datasets, including the Authors dataset [5] for textual analysis and three biomedical datasets (Lung-discrete [42], GLIOMA [42], and Brain [14]) for high-dimensional biological pattern recognition.

2) Compared methods and parameter configurations. To demonstrate the advantages of GSFRC, we perform comprehensive methodological comparisons.

- K-means: K-means clustering. It is a baseline method using MATLAB's built-in *kmeans* function.
- CC [12]: Convex fusion regularized clustering in (1.1), which is solved via a semismooth Newton-based augmented Lagrangian method.
- SCC [16]: Sparse convex fusion regularized clustering with inter-group lasso penalty in (1.2), using the ADMM for optimization.
- SGLCC [14]: Convex fusion regularized clustering with sparse group lasso penalty in (1.3), solved via the ADMM.
- ERC [43]: Fusion regularized clustering with adaptive spurious connection, solved by a semismooth Newton-based alternating minimization algorithm.

Unless specified otherwise, the reported K-means results are averaged across 30 executions. The other methods are not stochastic. Besides, we use Algorithm 2 to tune ω_ι and $\boldsymbol{\beta}$, the parameter adjustment method from [5] to tune α and γ, and a grid search to select η from $\{2^{-8}, 2^{-7}, \ldots, 2^{0}\}$ and τ from $\{0.1, 0.2, \ldots, 1.5\}$. To ensure a fair comparison, the parameters ϕ and neighborhood size k used in ω_ι for all FRC methods are selected from the sets $[10^{-5}, 10^{-3}, 0, 1, 2, 5]$ and $\{3, 4, \ldots, 13\}$, respectively. As suggested in [33], the values of p and q for GSFRC are chosen from the set $\{0, 1/2, 2/3\}$. Based on Definition 3.1 and (4.1), the iterative process of Algorithm 1 terminates when the generated sequence $\{(\boldsymbol{\Upsilon}_1^{t+1}, \boldsymbol{X}^{t+1}, \boldsymbol{\Upsilon}_2^{t+1})\}$ satisfies

$$\max\{\text{error}_1^{t+1},\ \text{error}_2^{t+1},\ \text{error}_3^{t+1},\ \text{error}_4^{t+1},\ \text{error}_5^{t+1}\} < 10^{-3}$$

or when the number of iterations reaches 500, where

$$\begin{aligned}
\text{error}_1^{t+1} &= \frac{\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^t\|_F}{\|\boldsymbol{X}^t\|_F + 1}, \ \text{error}_2^{t+1} = \frac{\|\boldsymbol{E}^{t+1} - \boldsymbol{E}^t\|_F}{\|\boldsymbol{E}^t\|_F + 1},\\
\text{error}_3^{t+1} &= \frac{\|\boldsymbol{F}^{t+1} - \boldsymbol{F}^t\|_F}{\|\boldsymbol{F}^t\|_F + 1}, \ \text{error}_4^{t+1} = \frac{\|\bar{\boldsymbol{F}}^{t+1} - \bar{\boldsymbol{F}}^t\|_F}{\|\bar{\boldsymbol{F}}^t\|_F + 1},\\
\text{error}_5^{t+1} &= \frac{\|\boldsymbol{X}^{t+1} - \boldsymbol{A} + \boldsymbol{D}^\top \boldsymbol{P}^{t+1} + \boldsymbol{G}^{t+1} + \bar{\boldsymbol{G}}^{t+1}\|_F}{\|\boldsymbol{A}\|_F + 1}.
\end{aligned}$$

3) Evaluation metrics. To assess the accuracy of clustering, four indicators are chosen, namely accuracy (ACC) [44], normalized mutual information (NMI) [45], adjusted rand index (ARI) [5], and Fowlkes–Mallows index (FMI) [46]. Moreover, given that the effective features in the simulated datasets are known, we utilize false negative rate (FNR) and false positive rate (FPR) to evaluate the error rate of feature selection within the simulated datasets [16]. All metrics range from 0 to 1 except ARI, which ranges from -1 to 1. Generally, the higher the ARI and FMI, the smaller the FNR and FPR, and the better performance of the method.

5.2. Experimental results

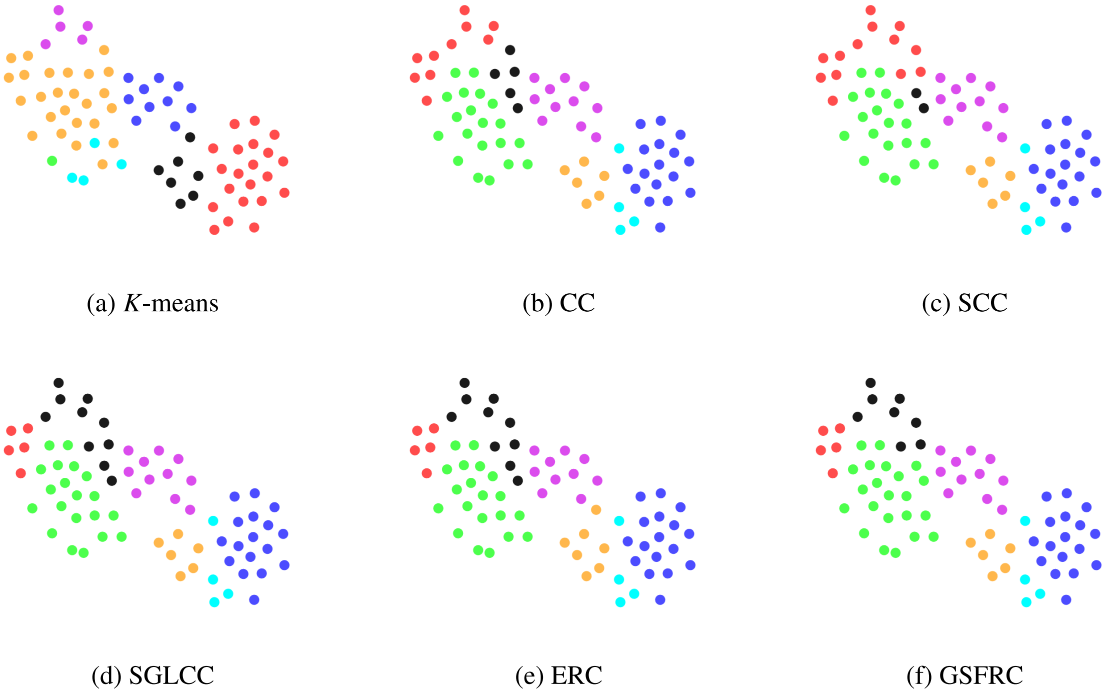
1) Simulated results. In this experiment, comparative evaluations of different methods are conducted on two simulated datasets. Table 2 demonstrates that our proposed GSFRC achieves the highest clustering accuracy and the lowest feature selection error rate among all compared approaches, with the best and second-best performances highlighted in boldface and underlined, respectively.

Table 2. Comparisons of ACC, NMI, FNR, and FPR for different clustering methods on simulated datasets.

Datasets	Methods	K-means	CC	SCC	SGLCC	ERC	GSFRC
Dartboard1	ACC ↑	26.00	25.70	**100.00**	56.50	25.70	**100.00**
	NMI ↑	0.05	0.62	**100.00**	46.27	0.62	**100.00**
	FNR ↓	100.00	100.00	**0.00**	**0.00**	100.00	**0.00**
	FPR ↓	**0.00**	**0.00**	80.00	33.33	**0.00**	**0.00**
Diamond9	ACC ↑	42.37	36.57	88.87	99.33	11.40	**99.37**
	NMI ↑	45.53	34.01	92.80	98.55	0.45	**98.63**
	FNR ↓	100.00	100.00	**0.00**	**0.00**	100.00	**0.00**
	FPR ↓	**0.00**	**0.00**	**0.00**	50.00	**0.00**	**0.00**

Table 3. Comparisons of ACC, NMI, ARI, and FMI for different clustering methods on real-world datasets.

Datasets	Methods	*K*-means	CC	SCC	SGLCC	ERC	GSFRC
Authors	ACC ↑	77.29	99.52	99.76	99.52	98.93	**99.88**
	NMI ↑	75.68	97.60	98.83	97.68	95.11	**99.35**
	ARI ↑	74.88	98.81	99.19	98.58	96.75	**99.60**
	FMI ↑	82.41	99.10	99.45	99.02	97.76	**99.72**
Lung-discrete	ACC ↑	65.21	82.19	79.45	86.30	86.30	**89.04**
	NMI ↑	62.58	76.46	77.48	78.04	78.12	**81.95**
	ARI ↑	47.78	71.17	71.75	73.01	73.45	**80.26**
	FMI ↑	57.03	76.22	76.84	77.74	78.09	**83.82**
GLIOMA	ACC ↑	52.00	60.00	60.00	60.00	58.00	**63.00**
	NMI ↑	43.79	**52.42**	46.03	46.03	46.69	48.86
	ARI ↑	30.52	39.45	38.74	38.74	36.70	**39.65**
	FMI ↑	50.33	55.08	58.46	58.46	56.45	**62.39**
Brain	ACC ↑	54.76	78.57	80.95	73.81	78.57	**81.32**
	NMI ↑	42.77	68.31	69.23	65.42	68.31	**76.47**
	ARI ↑	31.68	60.09	64.36	56.36	60.09	**69.47**
	FMI ↑	52.16	69.04	72.75	66.47	69.04	**77.44**



(a) *K*-means

(b) CC

(c) SCC

(d) SGLCC

(e) ERC

(f) GSFRC

Figure 3. Visual comparison of clustering results from 6 methods on Lung-discrete.

From Table 2, GSFRC achieves and maintains the highest accuracy, with SCC or SGLCC being the second-best performer judging by the values of ACC and NMI. FNR and FPR evaluate feature selection performance by quantifying missed true features and incorrectly selected noise, respectively. As shown in Table 2, SCC and SGLCC achieve higher FNR or FPR than GSFRC. This phenomenon suggests that l_1-norm is inherently limited compared to l_q-norm for feature selection, because l_1-norm selects more noise components. Moreover, the FNR=0 and FPR=1 of K-means, CC, and ERC reflect that these methods lack feature selection capability, so that all features are considered to be valid features.

4) Real-world results. As shown in Table 3, GSFRC achieves the highest clustering accuracy. Compared to the four fusion-regularized clustering methods (CC, SCC, SGLCC, and ERC), GSFRC improves the average ARI by approximately 1.27%, 7.9%, 1.24%, and 9.25%, respectively. Besides, the parameter values corresponding to the optimal accuracy achieved by each method are documented in Table 6 of the Supplementary Material (Section IV).

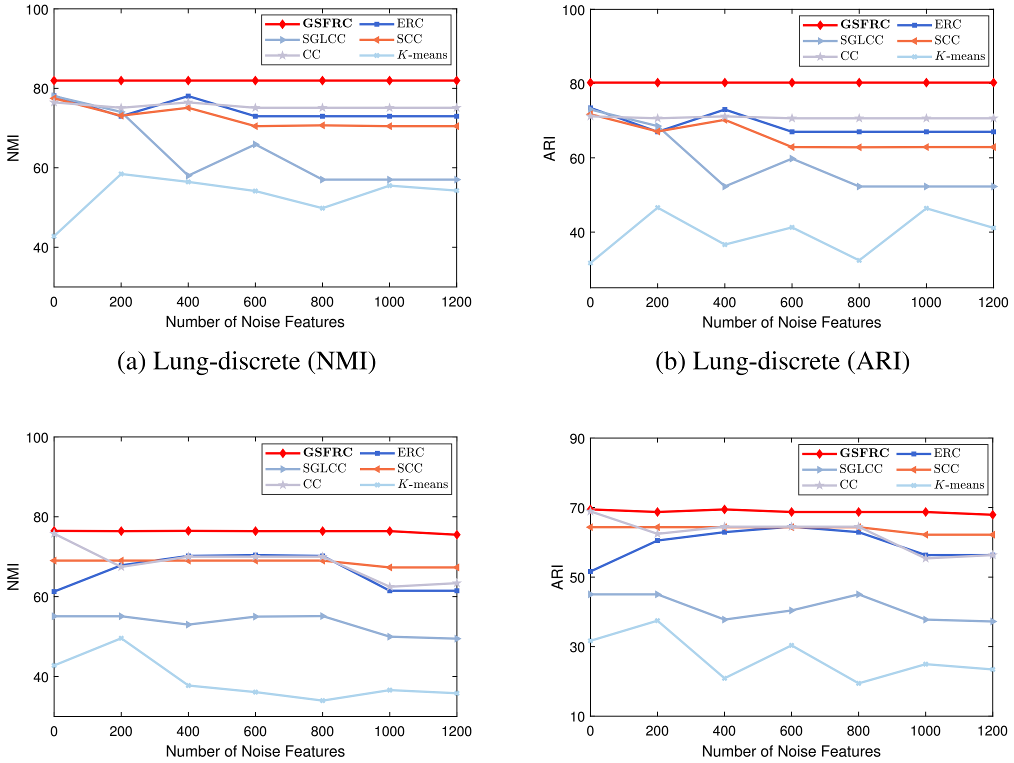
For a more intuitive comparison, Figure 3 visualizes the clustering results of all six methods on the Lung-discrete dataset using a t-SNE embedding, where each color represents a distinct cluster. As shown in Figure 3, cluster separation varies across different methods. GSFRC shows clearer inter-cluster boundaries because it achieves the highest accuracy.

Table 4. Ablation study of clustering performance and feature selection capability.

Datasets	$l_{2,p}$	l_q	ACC ↑	NMI ↑	ARI ↑	FMI ↑
Authors	✓	✓	**99.88**	**99.35**	**99.60**	**99.72**
	×	✓	65.16	59.47	49.65	73.22
	✓	×	**99.88**	**99.35**	**99.60**	**99.72**
	×	×	99.76	98.84	99.19	99.45
Lung-discrete	✓	✓	**89.04**	**81.95**	**80.26**	**83.82**
	×	✓	**89.04**	**81.95**	**80.26**	**83.82**
	✓	×	86.03	78.04	73.01	77.74
	×	×	86.03	78.04	73.01	77.74
GLIOMA	✓	✓	**63.00**	**48.86**	**39.65**	**62.39**
	×	✓	62.00	47.66	38.96	58.07
	✓	×	60.00	46.03	38.73	58.46
	×	×	58.00	46.69	36.70	56.45
Brain	✓	✓	**81.32**	**76.47**	**69.47**	**77.44**
	×	✓	73.81	61.58	55.03	65.86
	✓	×	80.95	69.23	64.36	72.75
	×	×	57.14	51.65	38.95	59.67

5.3. Ablation studies

To further demonstrate the effectiveness of the group sparsity strategy ($l_{2,p}$ + l_q), we conduct comprehensive ablation studies on the Authors, Lung-discrete, GLIOMA, and Brain datasets. As shown in Table 4, we compare the clustering performance of our proposed method against three of its variants by systematically removing the inter-group sparsity term ($l_{2,p}$), the intra-group sparsity term (l_q), or both. Quantitative results in Table 4, evaluated using ACC, NMI, ARI, and FMI, show that our method consistently outperforms or performs comparably to the variants. Both the $l_{2,p}$ and l_q components distinctly enhance clustering accuracy, demonstrating that the sparse feature patterns they produce are more conducive to effective clustering.



(a) Lung-discrete (NMI)

(b) Lung-discrete (ARI)

(c) Brain (NMI)

(d) Brain (ARI)

Figure 4. Performance comparison (NMI (%) and ARI (%)) from 6 methods on Lung-discrete and Brain with an increasing number of noise features.

5.4. Discussions

1) Stability analysis. In this study, Gaussian noise, generated by randn(size($\boldsymbol{X}$)), is added to the Lung-discrete and Brain datasets to simulate data corruption. The clustering performance of GSFRC and five other competing methods was compared using line charts. As illustrated in Figure 4, GSFRC consistently maintains stable and superior performance as the number of noisy features increases, confirming the robustness and effectiveness of the proposed approach.

2) Parameter analysis. We conduct a sensitivity analysis of the hyperparameters α, γ, and η using the Brain dataset. Figure 5 illustrates the variation in the structure of $\tilde{\boldsymbol{X}}$ with respect to these parameters. The dimensions of $\tilde{\boldsymbol{X}}$ reflect both the clustering structure and the feature selection behavior. The following observations are made:

- With fixed γ and η, the number of clusters decreases as α increases.
- With fixed α and η, the number of selected features decreases as γ increases.
- With fixed α and γ, all informative features are selected when $\eta = 0$; as η increases, the method progressively selects more locally informative features.

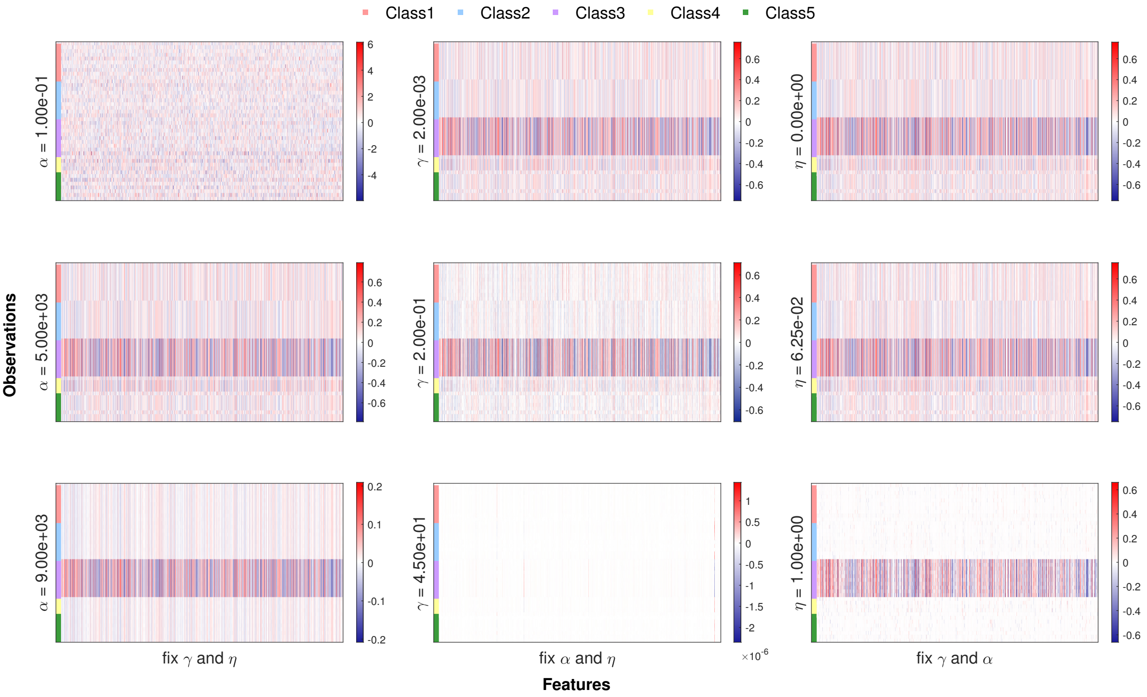
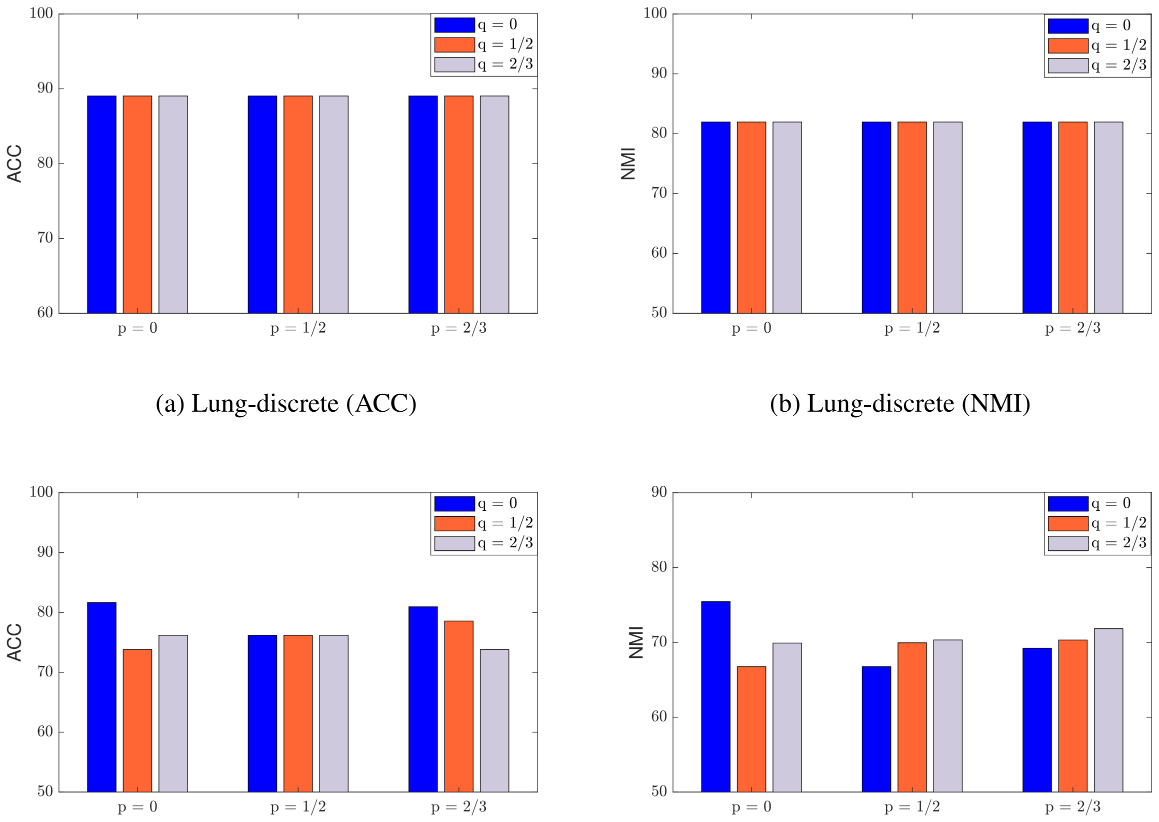


Figure 5. Clustering path α and regularization path γ, η of our method on Brain. For the subplots of each dataset, from left to right, we characterize the responses of $\tilde{X}$ by fixing γ and η and adjusting α, fixing α and η and adjusting γ, and fixing γ and α and adjusting η, respectively.

Subsequently, we investigate the impact of p and q on the clustering results, evaluated through ACC and NMI. Here, we set p and q to their commonly adopted values, namely 0, 1/2, and 2/3. As shown in Figure 6, the effects of p and q vary significantly across different datasets. On the Lung-discrete dataset (Subfigures 6a and 6b), GSFRC exhibits relatively low sensitivity to these parameters, with both ACC and NMI remaining stable under different configurations. In contrast, on the Brain dataset (Subfigures 6c and 6d), the choice of p and q exerts a more pronounced effect on model performance. The highest clustering agreement with ground-truth labels is achieved when $p = 0$ and $q = 0$, yielding optimal values in both ACC and NMI. A secondary favorable configuration occurs when $p = 2/3$ and $q = 0$, which leads to relatively high ACC, whereas the pair $p = 2/3$ and $q = 2/3$ leads to competitively high NMI. These observations suggest that the sensitivity to p and q is data-dependent, and their optimal values should be selected according to specific data characteristics.



(a) Lung-discrete (ACC)

(b) Lung-discrete (NMI)

(c) Brain (ACC)

(d) Brain (NMI)

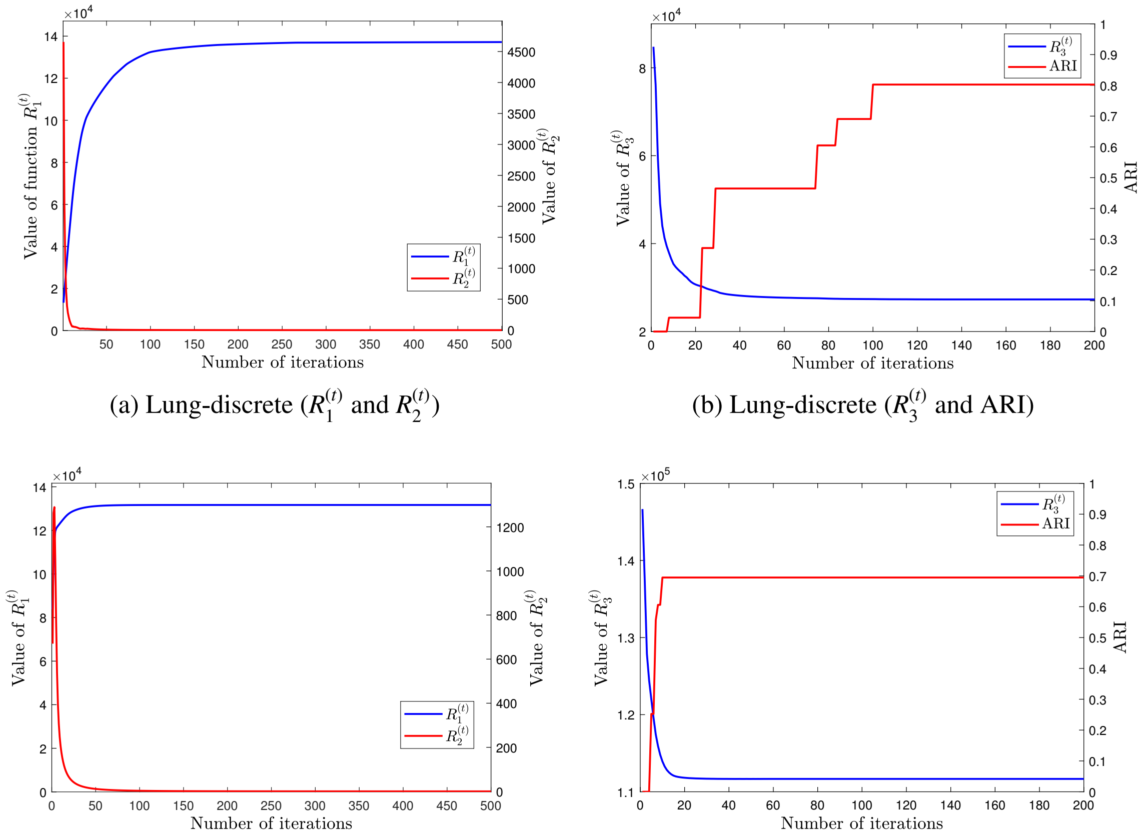
Figure 6. Effects of p and q on Lung-discrete and Brain in terms of ACC (%) and NMI (%).

3) Convergence. Let

$$\begin{aligned} R_1^{(t)} &= \|\boldsymbol{\Upsilon}_2^{(t)}\|_F^2, \\ R_2^{(t)} &= \theta\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F^2 + \tau(\varepsilon^{(t)} - \varepsilon^{(t+1)}) - \|\boldsymbol{\Upsilon}_2^{(t+1)} - \boldsymbol{\Upsilon}_2^{(t)}\|_F^2, \\ R_3^{(t)} &= \frac{1}{2}\|\boldsymbol{X}^{(t)} - \boldsymbol{A}\|_F^2 + \alpha\sum_{\iota\in\epsilon} w_\iota\|(\boldsymbol{D}\boldsymbol{X}^{(t)})_{\iota\cdot}\|_2 + \gamma((1-\eta)\|\boldsymbol{X}^{(t)}\boldsymbol{\beta}\|_{2,p}^p + \eta\|\boldsymbol{X}^{(t)}\|_q^q). \end{aligned}$$

Next, we set $\varepsilon^{(t)} = c\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F^2$ to validate Assumption 4.1, while monitoring the evolution of both objective function value and ARI value with increasing iterations t, where c is a positive constant of situation-dependent magnitude.

As observed in Subfigures 7a and 7c, both $R_1^{(t)}$ and $R_2^{(t)}$ exhibit clear convergence trends with increasing iterations on the Lung and Brain datasets. This empirical observation substantiates the boundedness of the multiplier sequence $\{\boldsymbol{\Upsilon}_2^{(t)}\}_{t=1}^{+\infty}$ and the non-negativity of $R_2^{(t)}$, thereby validating Assumption 4.1. Meanwhile, Subfigures 7b and 7d demonstrate a consistent decrease in the objective function value $R_3^{(t)}$ accompanied by a corresponding increase in ARI, with both reaching stabilization within finite iterations. These coordinated convergence behaviors confirm that Algorithm 1 possesses both optimization convergence and clustering effectiveness.



(a) Lung-discrete ($R_1^{(t)}$ and $R_2^{(t)}$)

(b) Lung-discrete ($R_3^{(t)}$ and ARI)

(c) Brain ($R_1^{(t)}$ and $R_2^{(t)}$)

(d) Brain ($R_3^{(t)}$ and ARI)

Figure 7. Dynamic evolution of $R_1^{(t)}$, $R_2^{(t)}$, $R_3^{(t)}$, and ARI on Lung-discrete and Brain datasets.

4) Runtime comparison. We report the computation time of different FRC methods on real-world datasets. All experiments are conducted using MATLAB (R2024a) on a Windows PC equipped with an Intel Core Ultra 5 125H processor (1.20 GHz) and 32 GB of RAM. As shown in Table 5, GSFRC demonstrates superior efficiency over ERC and is comparable to SCC and SGLCC across all datasets. This indicates that GSFRC achieves more powerful clustering and feature selection capabilities without incurring substantial additional time costs.

Table 5. Runtime results (in seconds) of different FRC methods on real-world datasets.

datasets	CC	SCC	SGLCC	ERC	GSFRC
Authors	0.48	5.73	8.07	41.75	8.09
Lung-discrete	0.14	0.13	0.15	13.51	0.13
GLIOMA	2.90	0.49	0.57	145.63	0.62
Brain	10.46	0.61	0.69	148.78	0.67

6. Conclusions

In this paper, we propose a novel group sparsity-based fusion regularized clustering method (GSFRC) designed to tackle clustering and feature selection problems in data with substantial uninformative features. To the best of our knowledge, this is the first work to incorporate non-convex inter-group sparsity and intra-group sparsity within a fusion regularization framework. Furthermore, we tackle the theoretical challenges of the non-convex and non-Lipschitz model by establishing the KKT necessary optimality conditions and proving the strong convergence of the ADMM algorithm. Extensive experimental results validate the superiority of our approach, particularly on the Lung-discrete dataset, where GSFRC achieves improvements of at least 0.37% in ACC, 7.15% in NMI, 5.11% in ARI, and 4.69% in FMI over the compared methods.

Given the runtime performance of GSFRC, future work will prioritize the development of accelerated computational approaches, such as Newton-type algorithms [13] and adaptive sieving strategies [15], to drastically reduce the runtime. Additionally, we will explore a more rigorous theoretical convergence analysis for the non-convex ADMM, which would also help substantiate the current boundedness assumption.

Author contributions

Xiangru Xing: Writing–original draft, Methodology, Data curation, Visualization, Conceptualization. Linglong Kong: Writing–review, Validation, Supervision, Resources. Xin Wang: Writing–review and editing, Methodology, Supervision, Resources. Xianchao Xiu: Writing–review and editing, Methodology, Formal analysis, Conceptualization, Resources.

Use of Generative-AI tools declaration

During the preparation of this manuscript, the authors utilized DeepSeek and ChatGPT for assistance with proofreading and language polishing. The authors thoroughly reviewed and revised all AI-assisted content and are solely responsible for the final version.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant 12371322 and 12371306, and the Innovative Talent Project of Karamay under Grant XQZX20240078.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. L. Ling, Y. Gu, S. Zhang, J. Wen, A generalized K-means problem for clustering and an ADMM-based K-means algorithm, *J. Ind. Manag. Optim.*, **20** (2024), 2089–2115. https://doi.org/10.3934/jimo.2023157

2. L. Gao, J. Bien, D. Witten, Selective inference for hierarchical clustering, *J. Am. Stat. Assoc.*, **119** (2024), 332–342. https://doi.org/10.1080/01621459.2022.2116331

3. L. Ding, C. Li, D. Jin, S. Ding, Survey of spectral clustering based on graph theory, *Pattern Recogn.*, **151** (2024), 110366. https://doi.org/10.1016/j.patcog.2024.110366

4. Y. Zhu, X. Xiu, W. Liu, C. Yin, Joint sparse subspace clustering via fast $l_{2,0}$-norm constrained optimization, *Expert Syst. Appl.*, **265** (2025), 125845. https://doi.org/10.1016/j.eswa.2024.125845

5. M. Wang, G. I. Allen, Integrative generalized convex clustering optimization and feature selection for mixed multi-view data, *J. Mach. Learn. Res.*, **22** (2021), 2498–2571.

6. Z. Wang, Y. Yuan, J. Ma, T. Zeng, D. Sun, Randomly projected convex clustering model: Motivation, realization, and cluster recovery guarantees, *J. Mach. Learn. Res.*, **26** (2025), 1–57.

7. Q. Feng, C. Chen, L. Liu, A review of convex clustering from multiple perspectives: Models, optimizations, statistical properties, applications, and connections, *IEEE Trans. Neural Netw. Learn. Syst.*, **35** (2024), 13122–13142. https://doi.org/10.1109/TNNLS.2023.3276393

8. K. Pelckmans, J. De Brabanter, J. A. Suykens, B. D. Moor, Convex clustering shrinkage, *in PASCAL Workshop on Statistics and Optimization of Clustering Workshop*, 2005.

9. F. Lindsten, H. Ohlsson, L. Ljung, Clustering using sum-of-norms regularization: With application to particle filter output computation, *in 2011 IEEE Statistical Signal Processing Workshop (SSP)*, 2011, 201–204. https://doi.org/10.1109/SSP.2011.5967659

10. T. D. Hocking, A. Joulin, F. Bach, J. P. Vert, Clusterpath an algorithm for clustering using convex fusion penalties, *in Proc. Int. Conf. Mach. Learn.*, 2011.

11. Z. Wang, X. Liu, Q. Li, A euclidean distance matrix model for convex clustering, *J. Optim. Theory Appl.*, **205** (2025), 1. https://doi.org/10.1007/s10957-025-02616-5

12. D. Sun, K. C. Toh, Y. Yuan, Convex clustering: Model, theoretical guarantee and efficient algorithm, *J. Mach. Learn. Res.*, **22** (2021), 427–458.

13. M. Lin, Y. Zhang, Low rank convex clustering for matrix-valued observations, *arXiv preprint arXiv:2412.17328*.

14. H. Chen, L. Kong, Y. Li, A novel convex clustering method for high-dimensional data using semiproximal ADMM, *Math. Probl. Eng.*, **2020** (2020), 1–12. https://doi.org/10.1155/2020/9216351

15. Y. Yuan, M. Lin, D. Sun, K. C. Toh, Adaptive sieving: A dimension reduction technique for sparse optimization problems, *Math. Program. Comput.*, **17** (2025), 585–616. https://doi.org/10.1007/s12532-025-00282-2

16. B. Wang, Y. Zhang, W. Sun, Y. Fang, Sparse convex clustering, *J. Comput. Graph. Stat.*, **27** (2018), 393–403. https://doi.org/10.1080/10618600.2017.1377081

17. T. Hastie, R. Tibshirani, M. Wainwright, *Statistical learning with sparsity: The lasso and generalizations*, 1 Eds., New York: CRC Press, 2015. https://doi.org/10.1201/b18401

18. A. Parekh, I. W. Selesnick, Improved sparse low-rank matrix estimation, *Signal Process.*, **139** (2017), 62–69. https://doi.org/10.1016/j.sigpro.2017.04.011

19. Y. Zhu, X. Zhang, G. Wen, W. He, D. Cheng, Double sparse-representation feature selection algorithm for classification, *Multimed. Tools Appl.*, **76** (2017), 17525–17539. https://doi.org/10.1007/s11042-016-4121-8

20. X. Xiu, C. Huang, P. Shang, W. Liu, Bi-sparse unsupervised feature selection, *IEEE Trans. Image Process.*, (2025), 1–15. https://doi.org/10.1109/TIP.2025.3620667

21. Y. Hu, J. X. Liu, Y. L. Gao, J. Shang, DSTPCA: Double-sparse constrained tensor principal component analysis method for feature selection, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **18** (2021), 1481–1491. https://doi.org/10.1109/TCBB.2019.2943459

22. Y. Xu, G. Liu, X. Lu, C. Xie, L. Xue, T. Jiang, Double sparse structure-enhanced mmWave NLOS imaging under multiangle relay surface, *IEEE Trans. Signal Process.*, **72** (2024), 5628–5643. https://doi.org/10.1109/TSP.2024.3505938

23. S. Liao, C. Han, T. Guo, B. Li, Subspace Newton method for sparse group l_0 optimization problem, *J. Global Optim.*, **90** (2024), 93–125. https://doi.org/10.1007/s10898-024-01396-y

24. J. Huang, J. L. Horowitz, S. Ma, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *Ann. Stat.*, **36** (2008), 587–613. https://doi.org/10.1214/009053607000000875

25. R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, *IEEE Signal Process. Lett.*, **14** (2007), 707–710. https://doi.org/10.1109/LSP.2007.898300

26. L. Niu, R. Zhou, Y. Tian, Z. Qi, P. Zhang, Nonsmooth penalized clustering via l_p regularized sparse regression, *IEEE Trans. Cybern.*, **47** (2016), 1423–1433. https://doi.org/10.1109/TCYB.2016.2546965

27. H. Chen, L. Kong, Y. Li, Nonconvex clustering via l_0 fusion penalized regression, *Pattern Recogn.*, **128** (2022), 108689. https://doi.org/10.1016/j.patcog.2022.108689

28. X. Gao, Y. Bai, Smoothed hybrid l_p-l_2 model for sparse optimization, *J. Ind. Manag. Optim.*, **21** (2025), 4712–4729. https://doi.org/10.3934/jimo.2025071

29. J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Math. Program.*, **146** (2014), 459–494. https://doi.org/10.1007/s10107-013-0701-9

30. K. Guo, D. Han, T. T. Wu, Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints, *Int. J. Comput. Math.*, **94** (2017), 1653–1669. https://doi.org/10.1080/00207160.2016.1227432

31. H. Zhang, J. Gao, J. Qian, J. Yang, C. Xu, B. Zhang, Linear regression problem relaxations solved by nonconvex ADMM with convergence analysis, *IEEE Trans. Circuits Syst. Video Technol.*, **34** (2024), 828–838. https://doi.org/10.1109/TCSVT.2023.3291821

32. A. Beck, *First-order methods in optimization*, SIAM, 2017. https://doi.org/10.1137/1.9781611974997

33. S. Zhou, X. Xiu, Y. Wang, D. Peng, Revisiting l_q $(0 \leq q < 1)$ norm regularized optimization, *arXiv preprint arXiv:2306.14394*

34. J. Liu, M. Feng, X. Xiu, W. Liu, X. Zeng, Efficient and robust sparse linear discriminant analysis for data classification, *IEEE Trans. Emerg. Top. Comput. Intell.*, **9** (2024), 617–629. https://doi.org/10.1109/TETCI.2024.3403912

35. R. T. Rockafellar, R. J. B. Wets, *Variational analysis*, Springer Science & Business Media, 2009. https://doi.org/10.1007/978-3-642-02431-3

36. C. Chen, B. He, Y. Ye, X. Yuan, The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent, *Math. Program.*, **155** (2016), 57–79. https://doi.org/10.1007/s10107-014-0826-5

37. S. Zhou, G. Y. Li, Federated learning via inexact ADMM, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 9699–9708. https://doi.org/10.1109/TPAMI.2023.3243080

38. T. T. Cui, Y. Z. Dang, Y. Gao, Distributed multi-block partially symmetric Bregman ADMM for nonconvex and nonsmooth sharing problem, *J. Oper. Res. Soc. China*, (2025), 1–27. https://doi.org/10.1007/s40305-024-00579-4

39. M. Chao, Z. Deng, J. Jian, Convergence of linear Bregman ADMM for nonconvex and nonsmooth problems with nonseparable structure, *Complexity*, **2020** (2020), 6237942. https://doi.org/10.1155/2020/6237942

40. J. Yin, C. Tang, J. Jian, Q. Huang, A partial Bregman ADMM with a general relaxation factor for structured nonconvex and nonsmooth optimization, *J. Global Optim.*, **89** (2024), 899–926. https://doi.org/10.1007/s10898-024-01384-2

41. H. Attouch, J. Bolte, B. F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods, *Math. Program.*, **137** (2013), 91–129. https://doi.org/10.1007/s10107-011-0484-9

42. F. Yang, Z. Xu, H. Wang, L. Sun, M. Zhai, J. Zhang, A hybrid feature selection algorithm combining information gain and grouping particle swarm optimization for cancer diagnosis, *PLoS One*, **19** (2024), e0290332. https://doi.org/10.1371/journal.pone.0290332

43. H. Chen, L. Kong, W. Qu, X. Xiu, An enhanced regularized clustering method with adaptive spurious connection detection, *IEEE Signal Process. Lett.*, **30** (2023), 1332–1336. https://doi.org/10.1109/LSP.2023.3316023

44. K. Zhan, F. Nie, J. Wang, Y. Yang, Multiview consensus graph clustering, *IEEE Trans. Image Process.*, **28** (2018), 1261–1270. https://doi.org/10.1109/TIP.2018.2877335

45. S. G. Fang, D. Huang, C. D. Wang, Y. Tang, Joint multi-view unsupervised feature selection and graph learning, *IEEE Trans. Emerg. Top. Comput. Intell.*, **8** (2024), 16–31. https://doi.org/10.1109/TETCI.2023.3306233

46. P. Zhang, X. Liu, J. Xiong, S. Zhou, W. Zhao, E. Zhu, et al., Consensus one-step multi-view subspace clustering, *IEEE Trans. Knowl. Data Eng.*, **34** (2022), 4676–4689. https://doi.org/10.1109/TKDE.2020.3045770

AIMS Press