Noname manuscript No.
(will be inserted by the editor)

Faster Nonconvex Low-rank Matrix Learning for Image Low-level and High-level Vision: A Unified Framework

Hengmin Zhang · Jian Yang · Jianjun Qian
Chen Gong · Xin Ning · Zhiyuan Zha · Bihan Wen*

Received: date / Accepted: date

Abstract This study introduces a unified approach to tackle challenges in both low-level and high-level vision tasks for image processing. The framework integrates faster nonconvex low-rank matrix computations and continuity techniques to yield efficient and high-quality results. In addressing real-world image complexities like noise, variations, and missing data, the framework exploits the intrinsic low-rank structure of the data and incorporates specific residual measurements. The optimization problem for low-rank matrix learning is effectively solved using the nonconvex Proximal Block Coordinate Descent (PBCD) algorithm, resulting in nearly unbiased estimators. Rigorous theoretical analysis ensures both local and global convergence. The PBCD algorithm updates blocks of variables iteratively with closed-form solutions, adeptly handling nonconvexity and promoting faster convergence. Notably, the framework incorporates the randomized singular value decomposition (RSVD) technique and introduces a continuous strategy for adaptive model parameter updates. These strategic choices reduce computational complexity while maintaining result quality. They offer fine-tuned control over the desired rank of the learned matrix and enhance robustness in a straightforward manner. Furthermore, the versatility of the proposed nonconvex PBCD algorithm extends to handling problems with multiple variables, as supported by theoretical analysis. Experimental evaluations, spanning various image low-level and high-level vision tasks such as inpainting, classification, and clustering, validate the effectiveness and efficiency of our framework across diverse databases. The source code is available at https://github.com/ZhangHengMin/FNPBCD_LR.

In a nutshell, our framework provides a unified solution to tackle both low-level and high-level vision tasks in images. By combining fast nonconvex low-rank matrix learning with adaptive parameter updates, we achieve efficient computation, yielding high-quality results that demonstrate robustness against various types of noise. The evaluations further endorse the reliability and applicability of our proposed framework.

Keywords Faster low-rank matrix learning · Nonconvex PBCD algorithm · Nonconvex relaxation functions · Randomized SVD · Image low-level and high-level vision

H. Zhang, Z. Zha, and B. Wen are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798 (email: zhanghengmin@126.com; {hengmin.zhang, zhiyuan.zha, bihan.wen}@ntu.edu.sg).
J. Yang, J. Qian, and C. Gong are with the PCA Lab, and also with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, P.R. China (email: {csjyang, csjqian, chen.gong}@njust.edu.cn).
X. Ning is with the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China (email: ningxin@semi.ac.cn).
(∗**Corresponding author: Bihan Wen.**)

1 Introduction

In recent years, there has been a significant surge of interest in the realms of low-rank matrix learning, sparse coding, and group sparsity techniques. These methods have proven to be powerful tools for addressing a wide array of low-level and high-level vision tasks within the realm of image processing and analysis. These tasks include but are not limited to image inpainting, classification, clustering, and more, as highlighted in various studies [1–10]. These approaches leverage the inherent low-rank structure present in data matrices, al-

lowing for the capture of underlying patterns and extraction of informative representations. The combination of low-rank matrix learning with (group) sparse coding [11–15] results in the desired low-rank matrix. This effectively models residuals, capturing crucial information for tasks. Moreover, the integration of group sparsity enhances the discriminative power and robustness of the acquired representations. The joint optimization of low-rank matrix learning, sparse coding, and group sparsity provides accurate and efficient solutions for various low-level and high-level vision tasks. These tasks include inpainting of missing image regions [16–18], image classification [19–23], and grouping and clustering of similar images [24–27]. Naturally, the application of these techniques tends to yield improved performance, enhanced visual outcomes, and greater overall reliability, making them valuable tools in the areas of image processing and analysis.

1.1 Related Works

Building on recent advancements, this study delves into the investigation of generalized and unconstrained minimization concerning learnable low-rank matrices. The problem is formulated as follows:

$$\min_{\mathbf{X},\mathbf{E}} \Phi_{\lambda,\mu}(\mathbf{X},\mathbf{E}) = h_\lambda(\mathbf{X}) + f(\mathbf{E}) + g_\mu(\mathbf{D};\mathbf{X},\mathbf{E}), \quad (1)$$

in which, $\lambda > 0$ and $\mu > 0$ represent the regularization and penalty parameters, respectively. The data matrix is denoted as $\mathbf{D}$, and the loss function $f(\mathbf{E})$ evaluates the model's fitting to the data, considering various types of noise [10, 11, 19, 28–32]. Simultaneously, the regularization term $h_\lambda(\mathbf{X})$ replaces the rank function, rank($\mathbf{X}$), which equals the number of non-zero singular values of matrix $\mathbf{X}$, with convex or nonconvex functions. The nonconvex nature of the rank function presents a common challenge in finding a global minimizer for solving Problem (1). To address this, a variety of measurements can be employed for both $h_\lambda(\mathbf{X})$ and $f(\mathbf{E})$, enabling the consideration of different low-rank structures and noise types. This versatility is illustrated in **Fig**. 1, where the X-axis and Y-axis represent row and column labels, leading to concrete problem formulations related to the rank function. More specifically, our focus centers on recent research addressing Problem (1), which revolves around the decomposition of matrices into low-rank components along with additive matrices or their linear combinations, as discussed in [9]. This method provides a well-suited framework for the assessment and ranking of diverse algorithms, each characterized by its unique decomposition strategies, loss functions, optimization problems, and solutions.

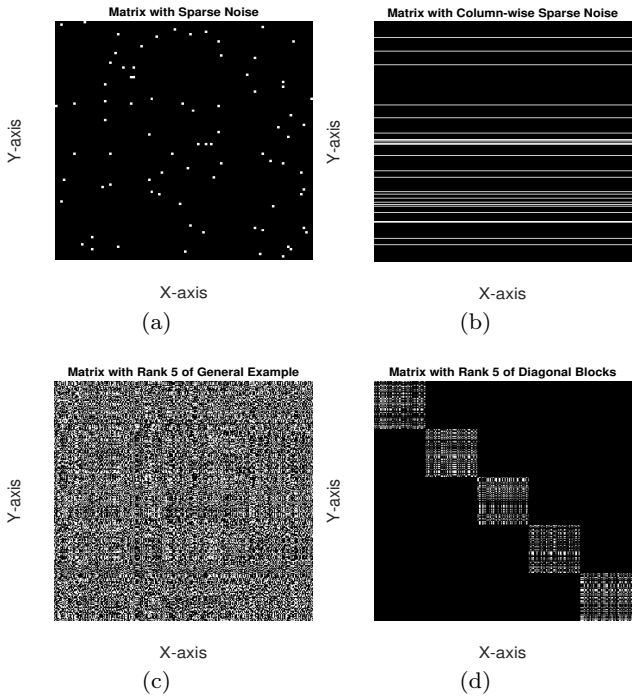


Fig. 1 Visual comparisons of two common types of noise and two common types of low-rank representation structures using objective metrics.

As a result, this adaptability empowers the creation of learnable models that can be customized to suit specific application tasks.

To circumvent the complexities associated with handling constraints, the penalty function $g_\mu(\mathbf{D};\mathbf{X},\mathbf{E})$ commonly employs the squared Frobenius norm. This transformation allows the problem to be recast into an unconstrained form, diverging from the typical constrained setup optimized using the alternating direction method of multipliers (ADMM) with Lagrange multipliers, as explored in [33–36]. Addressing the nonconvex nature and its associated challenges in Problem (1), particularly related to the rank function, various existing methods have been developed. These methods can be categorized based on different perspectives, each offering unique advantages. In the following, we introduce three widely adopted low-rank matrix learning methods, highlighting their distinctive viewpoints.

- Robust matrix completion (RMC) [39–41] is a technique devised to retrieve missing or corrupted entries within a matrix, capitalizing on its low-rank structure. The primary goal of this approach is to identify a low-rank matrix $\mathbf{X}$ that minimizes the following constrained optimization problem:

$$\min_{\mathbf{X},\mathbf{E}} \lambda\|\mathbf{X}\|_{S_p}^p + \|\mathcal{P}_\Omega(\mathbf{E})\|_{\ell_p}^p \text{ s.t. } \mathbf{D} = \mathbf{X} + \mathbf{E}, \quad (2)$$

Table 1 Several nonconvex relaxation function formulas of the ℓ_0-norm and their proximal operators used for Problem (1).

//	ℓ_p-norm [20]	MCP [37]	SCAD [38]
$h_\lambda(\cdot)$	$\lambda\|t\|^p$.	$\begin{cases} \lambda\|t\| - \frac{t^2}{2p}, & \|t\| \le p\lambda, \\ \frac{1}{2}p\lambda^2, & \text{otherwise.} \end{cases}$	$\begin{cases} \text{sign}(t)\max(\|t\|-\lambda, 0), & \|t\| \le 2\lambda, \\ \frac{(p-1)t - \text{sign}(t)\lambda p}{p-2}, & 2\lambda < \|t\| \le p\lambda, \\ t, & \text{otherwise.} \end{cases}$
$\text{Prox}_\lambda^h(\cdot)$	$\begin{cases} 0, & \|t\| \le \nu_1, \\ \text{argmin}_{t\in\{0,t_1\}} h(t), & t > \nu_1, \\ \text{argmin}_{t\in\{0,t_2\}} h(t), & \text{otherwise.} \end{cases}$	$\begin{cases} 0, & \|t\| \le \lambda, \\ t, & \|t\| > p\lambda, \\ \text{sign}(t)\frac{p(\|t\|-\lambda)}{p-1}, & \text{otherwise.} \end{cases}$	$\begin{cases} \lambda\|t\|, & \|t\| \le \lambda, \\ \frac{1}{2}\lambda^2(p+1), & \|t\| > p\lambda, \\ \frac{\lambda p\|t\| - \frac{1}{2}(t^2+\lambda^2)}{p-1}, & \text{otherwise.} \end{cases}$

where Ω represents the set of indices corresponding to the observed entries in the data matrix. The projected operator $\mathcal{P}_\Omega(\mathbf{D})$ is defined as follows: $\mathcal{P}_\Omega(\mathbf{D}_{ij}) = \mathbf{D}_{ij}$ if $(i,j) \in \Omega$, and 0 otherwise. The formulation usually incorporates the Schatten p-norm and ℓ_p-norm with $p > 0$ to gauge the low-rankness of the matrix $\mathbf{X}$ and the residuals in $\mathbf{E}$, respectively. As indicated in (2), the primary objective of RMC and its various extensions is to leverage the assumption of low-rankness for recovering missing entries while mitigating the impact of outliers or corrupted data.

– Robust matrix regression (RMR) [22, 42–44] stands out as a powerful methodology that blends the identification of structural noises using low-rank matrices with the assessment of representation coefficients through sparsity and collaboration. However, a notable drawback of RMR is its oversight of the relationship between testing and training samples, denoted as $\mathbf{D}$ and $\mathbf{A}$, respectively. To overcome this limitation, we here introduce a modified matrix regression formulation:

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{E}} \ \lambda\|\mathbf{X}\|_{S_p}^p + \sum_{i=1}^n \|\mathcal{M}(\mathbf{E}_i)\|_{S_p}^p \\ &\text{s.t. } \mathbf{D} = \mathbf{A}\mathbf{X} + \mathbf{E}, \end{aligned} \tag{3}$$

where $\mathbf{E}_i$ represents the column vector of the residual matrix $\mathbf{E}$, and $\mathcal{M}(\cdot)$ serves to transform a vector into a matrix representation. This transformation enables the incorporation of low-rank structures in both the error measurements and representation coefficients, eliminating the need for relying on the independent and identically distributed assumption. The utilization of the Schatten-p norm enhances the classification framework's flexibility and robustness, offering various p-values that effectively handle and capture underlying low-rank properties.

– Low-rank representation (LRR) [11, 25, 45] is an unsupervised method employed to model high-dimensional data by assuming its underlying low-rank structure, accounting for column noise and variations. Given a collection of data points $\{\mathbf{D}_1, \ldots, \mathbf{D}_n\}$, where each $\mathbf{D}_i$ signifies a high-dimensional vector, the primary goal is to identify a low-rank matrix $\mathbf{X}$ through the following problem:

$$\min_{\mathbf{X},\mathbf{E}} \ \lambda\|\mathbf{X}\|_{S_p}^p + \|\mathbf{E}\|_{\ell_{2,p}}^p \text{ s.t. } \mathbf{D} = \mathbf{A}\mathbf{X} + \mathbf{E}. \tag{4}$$

In this formulation, $\|\mathbf{X}\|_{\ell_{2,p}}^p$ represents the $\ell_{2,p}$-norm of $\mathbf{E}$, and $\mathbf{A}$ is a dictionary matrix comprising a set of basis vectors or original data points. The central objective of LRR is to identify a low-rank matrix $\mathbf{X}$ capable of effectively reconstructing the data points using the provided dictionary.

Crucially, Problem (1) serves as a cohesive framework that unifies various low-rank learning methods, encompassing Problems (2)-(4), and incorporating alternative measurements for the objective terms like minimax concave penalty (MCP) [37] and smoothly clipped absolute deviation (SCAD) [38], as illustrated in **Table** 1. The inclusion of penalty function techniques allows the transformation of these problems into unconstrained formulations, effectively imposing constraints on the data fidelity term by capturing the discrepancy between the observed data and the model's predictions [31, 46, 47]. Examining these methods through the lens of Problem (1) provides a deeper exploration of their strengths, limitations, and interconnections, facilitating a detailed analysis of their characteristics and relationships. This perspective proves invaluable for comprehensively understanding and analyzing these methods, as elaborated in the following discussion.

– The main challenge presented by the nonconvex and nonsmooth rank function in Problem (1) has led to

the adoption of convex nuclear norm substitution [48, 49]. The nuclear norm encourages low-rankness by penalizing the sum of singular values, facilitating efficient optimization algorithms based on factorization strategies, such as $\min_{\mathbf{X}=\mathbf{U}\mathbf{V}^\top} \frac{1}{2}(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)$ [50]. However, the nuclear norm tends to underestimate the true rank and may not provide an unbiased estimator in most cases.

To address this limitation, researchers have proposed alternative norms, including nonconvex truncated or weighted nuclear norms and Schatten p-norms [25, 51–53]. These norms act as substitutes for the nuclear norm and have demonstrated promising results across various applications. They introduce additional parameters, offering more flexibility in controlling the rank approximation and resulting in more accurate rank estimations. Notably, the Schatten p-norm has undergone extensive study with factorization formulas for specific values of p, including $p \in \{2/3, 1/2\}$ [25, 40]. Additionally, some relaxations of the ℓ_0-norm[1], such as the ℓ_p-norm [20], MCP [37], and SCAD [38] of **Table** 1, have been explored for singular values. These extensions can directly promote sparsity in the spectrum of singular values [21, 53], encouraging the preservation of a small number of dominant singular values while suppressing the remaining ones.

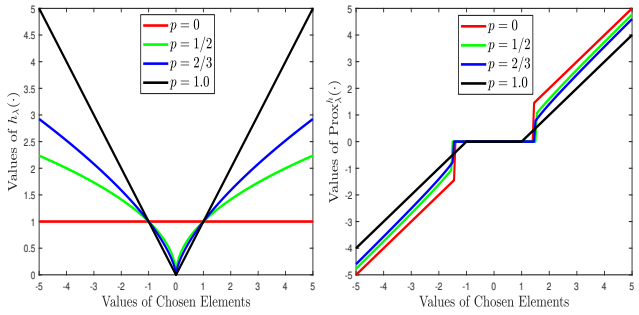
– The choice of an appropriate loss function for Problem (1) is pivotal in quantifying the dissimilarity between observed data and model predictions, especially when dealing with diverse types of noise. In practical applications, both convex and nonconvex norms are employed in loss functions to capture the specific characteristics of the noise distribution [22, 42, 54, 55]. For instance, the ℓ_1-norm is commonly used to model noise following a Laplace distribution, while the ℓ_2-norm is suitable for noise following a Gaussian distribution. The $\ell_{2,1}$-norm proves effective in the presence of column sparse noise, and the total variation norm is widely employed to promote piecewise smoothness and preserve edges in images. It's essential to note that nonconvex norms, such as the ℓ_p-norm, $\ell_{2,p}$-norm, and Schatten p-norm with $0 < p < 1$, have also demonstrated efficacy in handling specific noise patterns.

  Incorporating these nonconvex norms and extending them to measure specific loss functions becomes crucial for achieving more accurate and robust modeling, particularly when dealing with various noise styles, as exemplified in Problems (2)-(4) together with **Fig**. 1 (a)-(c). Leveraging these nonconvex norms has empirically shown significant improvements in image low-level and high-level vision tasks.

– The inclusion of penalty function strategies in Problem (1) plays a crucial role in crafting effective and efficient optimization algorithms, such as proximal alternating linearized minimization (PALM) [56] and scalable proximal Jacobian iteration method (SPJIM) [47]. These strategies prove particularly adept at addressing various types of noise or a combination thereof. The primary aim of these penalty functions is to impose constraints on model parameters or solutions, serving objectives like variable reduction, model simplification, computational complexity reduction, and ensuring convergence properties [25, 56, 57]. Additionally, penalty functions allow for the integration of prior knowledge or assumptions about the underlying data structure, striking a balance between algorithm design and an accurate representation of noise characteristics. Furthermore, the avoidance of introducing some auxiliary variables, achieved through the penalty function $g_\mu(\mathbf{D};\mathbf{X},\mathbf{E})$ derived from the equality constraints in Problems (2)-(4), is often employed as a strategy to enhance the optimization process. This approach facilitates problem formulation and enables efficient solution methods. Our work stands out from existing algorithms like ADMM, PALM, and SPJIM by introducing a unified optimization scheme tailored for low-rank matrix learning problems and applications in image low-level and high-level vision.

To gain a deeper understanding of (group) sparse coding and low-rank matrix learning within the context of Problem (1), we extend our exploration to encompass various nonconvex relaxations, as outlined in **Table** 1, rather than limiting our focus to a specific function. Our aim is to develop faster optimization algorithms adept at efficiently addressing Problems (2)-(4). In these algorithms, we exclusively employ the Schatten p-norm for low-rank matrix regularization, incorporating the ℓ_p-norm, $\ell_{2,p}$-norm, and Schatten p-norm for residual matrix regularization. This comprehensive iterative framework finds applications in diverse image low-level and high-level vision tasks. Moreover, the utilization of the Schatten-p norm enhances correlation and adaptability for structural analysis. In our approach, variable updates are accomplished through the application of proximal operators, a common practice in iterative procedures [35, 58]. To streamline this process, we provide

[1] The parameter choices are as follows: ℓ_p-norm with $0 < p < 1$, MCP with $p > 1$, and SCAD with $p > 2$. The proximal operators for ℓ_p-norm are defined by $\nu_1 = \nu+\lambda p|\nu|^{p-1}$, where $\nu = [\lambda p(1-p)]^{\frac{1}{2-p}}$. Additionally, t_1 and t_2 are the roots of $h(s) = (s-t)+\lambda p|s|^{p-1}\text{sign}(s) = 0$ at $\nu < s < t$ and $t < s < -\nu$, respectively.



(a) (b)

Fig. 2 Visual comparisons of the plotted nonconvex functions under $p \in \{0, 1/2, 2/3, 1.0\}$ and their proximal operators related to ℓ_p-norm in (a) and (b).

generalized proximal operators applicable in all three cases studied in this work:

Proposition 1 *Let $f(\cdot)$ be a proper and lower semi-continuous function defined on the entries or singular values (e.g., $\boldsymbol{e}_i$) of a vector or the column vectors (e.g., $\boldsymbol{E}_{:,i}$) of a matrix. The $Prox^f_\mu(\cdot)$ can be obtained by*

$$\begin{cases} argmin_{\boldsymbol{e}} f(\boldsymbol{e}) + \dfrac{\mu}{2}\|\boldsymbol{e}-\boldsymbol{t}\|_2^2, & (5) \\ argmin_{\boldsymbol{E}} \displaystyle\sum_i f(\|\boldsymbol{E}_{:,i}\|_2) + \dfrac{\mu}{2}\|\boldsymbol{E}-\boldsymbol{T}\|_F^2, & (6) \end{cases}$$

Here, $f(\cdot)$ represents a nonconvex relaxation of the ℓ_0-norm, $\ell_{2,0}$-norm, or rank function. The proximal operators for both (5) and (6) are denoted as follows:

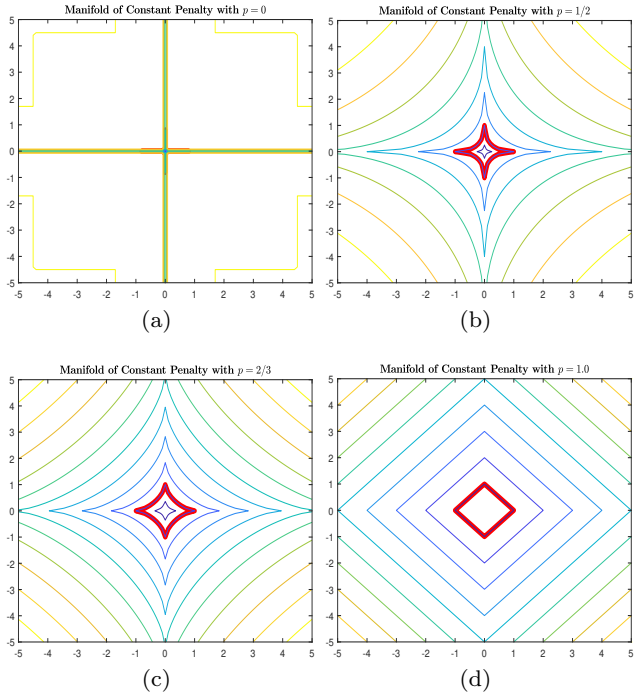
$$\begin{cases} Prox^f_\mu(\boldsymbol{t}) = \left[Prox^f_\mu(\boldsymbol{t}_1), ..., Prox^f_\mu(\boldsymbol{t}_n)\right], & (7) \\ Prox^f_\mu(\boldsymbol{T}) = \left[Prox^f_\mu(\boldsymbol{T}_{:,1}), ..., Prox^f_\mu(\boldsymbol{T}_{:,n})\right], & (8) \end{cases}$$

where we have

$$Prox^f_\mu(\boldsymbol{T}_{:,i}) = \begin{cases} \boldsymbol{0}, & \boldsymbol{T}_{:,i} = \boldsymbol{0}, \\ \frac{Prox^h_\lambda(\|\boldsymbol{T}_{:,i}\|_2)}{\|\boldsymbol{T}_{:,i}\|_2}\boldsymbol{T}_{:,i}, & otherwise. \end{cases} \quad (9)$$

This proposition outlines a scalable and manageable strategy for computing the proximal operator associated with various non-convex functions $f(\cdot)$ in Problems (2)-(4). To the best of our knowledge, the proximal operator plays a crucial role in promoting (group) sparsity and low-rankness through the functions listed in **Table** 1 and their extensions, all of which can be computed using **Proposition** 1. This computation approach proves highly efficient for solving a variety of problems and achieving desirable properties.

To highlight the distinctions, we present visualizations of the curves corresponding to the listed functions: specifically, the ℓ_p-norm under $p \in \{0, 1/2, 2/3, 1.0\}$ and their respective proximal operators in **Fig**. 2 (a) and (b), as well as the manifold of constant penalties for a symmetric 2×2 matrix in **Fig**. 3 (a)-(d). *Our primary focus centers on four distinct p-values, with an emphasis on the ℓ_p-norm due to its widespread usage in sparse coding and low-rank matrix recovery.* These visualizations provide valuable insights into the variations and characteristics of the nonconvex functions. When obtaining closed-form or analytic solutions proves challenging, researchers have extensively explored the weighted formulation of nonconvex functions, as previously studied [52, 53, 59]. This approach offers a flexible way to address complex optimization problems and achieve desirable properties, even when explicit solutions are not readily available, by introducing appropriate weightings based on the computable supergradient.



(a) (b) (c) (d)

Fig. 3 Visual comparisons of the nonconvex ℓ_p-function under $p \in \{0, 1/2, 2/3, 1.0\}$ through the manifold of constant penalties for a symmetric 2×2 matrix in (a)-(d).

1.2 Main Contributions

Examining the landscape of low-rank matrix recovery, it is clear that the field confronts various common limitations that impact diverse aspects, including evaluation efficacy, computational efficiency, robustness, interpretability, and generalization. Then, we categorize these challenges into three main directions, closely aligning with the primary contributions of this work. The following succinctly summarizes these areas.

- **Improvement for Computational Efficiency**: To address the limitations inherent in existing iterative algorithms for low-rank matrix learning problems (1)-(4), we propose the adoption of a continuation strategy and RSVD to augment the computational efficiency of the nonconvex PBCD algorithm. Firstly, the continuation strategy involves gradually increasing the penalty parameter value in each iteration until it reaches a predetermined maximum threshold. This incremental approach promotes faster convergence and reduces the computational burden by capitalizing on closed-form solutions for updating each variable. Secondly, the RSVD technique enables the factorization of small-scale matrices, leading to a more efficient solution with fewer auxiliary variables compared to traditional factorization methods. By incorporating these strategies, we can significantly reduce computation time and improve overall feasibility, making the approach more suitable for addressing large-scale problems encountered in image low-level and high-level vision tasks. This combination contributes to the enhanced computational efficiency and scalability of the proposed approach, enabling its successful applications.
- **Robustness to Noise and Variation**: Real-world data is often subjected to multiple sources of noise and variations, each characterized by specific features captured in the function $f(\mathbf{E})$. These factors exert a significant impact on the performance of image low-level and high-level vision tasks, necessitating the development of noise models and suitable residual functions based on the ℓ_p-norm, $\ell_{2,p}$-norm, and Schatten p-norm with $0 < p < 1$ in Problems (2)-(4). By carefully considering these factors, we can accurately assess and mitigate the effects caused by various noise sources, as illustrated in **Fig**. 1. Moreover, the integration of data-driven strategies enhances robustness by adapting to diverse data distributions and effectively handling outliers. To further validate the effectiveness of the proposed methods, this work focuses on conducting a comprehensive robustness analysis and validation across diverse data. This related analysis provides empirical evidence for handling various noise sources and variations, ensuring the practical applicability and reliability of our methodology.
- **Interpretability and Generalization**: Ensuring the interpretability and generalization capabilities of learnable low-rank models derived from Problem (1) is crucial for their practical utility in various low-level and high-level vision tasks. In this work, our focus is on incorporating both the rank of the coefficient matrix $\mathbf{X}$ and the representation error matrix $\mathbf{E}$ to uncover underlying patterns and structures. By utilizing the nonconvex and specific function $g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E})$ in the objective terms, our aim is to provide meaningful insights and interpretable representations. This approach not only yields accurate predictions or reconstructions but also offers understandable illustrations, enabling further analysis and informed decision-making to some extent. Moreover, the developed methods go beyond achieving accurate results and focus on strong generalization capabilities. This is accomplished through an extended iteration procedure of the nonconvex PBCD algorithm and the utilization of RSVD for learning low-rank models in Problems (2)-(4). To validate the superiority of the proposed methods, comprehensive experimental validations and comparative analyses are conducted, showcasing the better evaluation performance.

Building upon the statements above, we have highlighted the real challenges that substantiate the author's work from two distinct perspectives:

- Tackling these challenges is vital for the progression of the field. The incorporation of nearly unbiased estimators takes center stage as a crucial effort, playing a pivotal role in improving recovery and reconstruction capabilities. Simultaneously, the introduction of a unified framework for addressing multiple rank-relaxed problems emphasizes its importance, promising wide applicability across diverse scenarios. Moreover, the creation of an efficient and effective nonconvex optimization algorithm, coupled with the assurance of both local and global convergence guarantees, emerges as a critical pursuit. This not only ensures the practicality of our methods but also establishes a robust theoretical foundation.
- Attaining nearly unbiased estimators involves surmounting statistical complexities and ensuring a high level of accuracy in the estimation processes, adding an additional layer of intricacy to the overall task. The creation of a unified framework requires a comprehensive understanding of diverse rank-relaxed problems, demanding sophisticated theoretical and computational frameworks. Similarly, developing a nonconvex optimization algorithm that ensures convergence guarantees is a challenge, involving navigating the complex landscape of optimization theory and reconciling local and global convergence properties. The difficulty lies in striking a delicate balance between recovery precision, computational efficiency, and theoretical robustness, amplifying the complexity of these endeavors in pushing the boundaries of current low-rank matrix learning approaches.

1.3 Outline of this work

The structure of this work unfolds as follows: In Section 2, we introduce the unified formulation of the low-rank matrix learning problem and present efficient and effective optimization algorithms. This section encompasses a wide array of crucial techniques, offering detailed algorithm designs for the proposed approaches. It particularly emphasizes the relevance of low-rank matrices in addressing various image low-level and high-level vision tasks. Moving on to Section 3, our focus shifts to the analysis of the nonconvex PBCD algorithm, providing provable convergence guarantees at both local and global levels. This section delves into the theoretical aspects of the algorithm, enhancing our understanding of its workings. In Section 4, we extend the proposed nonconvex PBCD algorithm to handle problems with multiple variables, supplementing the explanation with illustrative examples to aid comprehension. Section 5 is dedicated to presenting experimental results on image inpainting, classification, and clustering tasks. We include ablation analysis to showcase the effects from several aspects, demonstrating the superiority of our approach in terms of accuracy and efficiency. These experimental results further offer empirical evidence supporting our claims. Finally, in Section 6, we conclude this work by summarizing the key findings and contributions. We also engage in discussions on future research areas for further improvements, highlighting potential directions for exploration.

2 Nonconvex PBCD Algorithm

To simplify the resolution of the nonconvex low-rank matrix learning Problem (1), we suggest an approach that involves introducing auxiliary variables $\hat{\mathbf{X}}$ and $\hat{\mathbf{E}}$. This reformulation allows us to express the problem without constraints, leading to the expression:

$$\begin{aligned}&\min_{\mathbf{X},\mathbf{E},\hat{\mathbf{X}},\hat{\mathbf{E}}} \Phi_{\lambda,\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}(\mathbf{X},\mathbf{E},\hat{\mathbf{X}},\hat{\mathbf{E}})\\ &= h_\lambda(\mathbf{X}) + f(\mathbf{E}) + g_\mu(\mathbf{D};\mathbf{X},\mathbf{E})\\ &\quad + \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}-\hat{\mathbf{X}}\|_F^2 + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}-\hat{\mathbf{E}}\|_F^2, \end{aligned} \tag{10}$$

where we establish the following assumptions to guide the proposed approach:

(***A1***) Both $h_\lambda(\cdot)$ and $f(\cdot)$ are proper, closed, nonnegative, and lower semi-continuous functions.

(***A2***) The function $g_\mu(\mathbf{D};\cdot,\cdot)$ is a C^1 function.

As described in Problems (2)-(4), the functions $h_\lambda(\cdot)$ and $f(\cdot)$ play a role in inducing desired structures. While convex norms like the nuclear norm, ℓ_1-norm, and $\ell_{2,1}$-norm are commonly used, we refrain from considering them in this work due to their tendency to introduce biased estimators. The stated assumptions (***A1***) and (***A2***) have broad applicability in the realm of nonconvex optimization problems, where the objective function often combines nonconvex and convex differentiable terms. Unlike existing algorithms for such problems, we integrate proximal terms with two parameters, $\phi_{\mathbf{X}} > 0$ and $\phi_{\mathbf{E}} > 0$, and address Problem (10) instead of directly approaching Problem (1). This specific choice stems from our objective to devise a nonconvex PBCD algorithm that efficiently solves subproblems using the closed-form solutions and minimizes the computational load associated with variable computations. Most importantly, we need to provide a more detailed explanation and justification for choosing PBCD over ADMM, focusing on two key aspects:

- Unlike the PBCD algorithm, the ADMM algorithm often requires updates for a larger number of variables when dealing with constrained problems. Proving convergence properties, especially in terms of global guarantees, is generally more challenging [45, 57]. Moreover, attaining superior efficiency through acceleration techniques poses inherent challenges. These considerations have propelled our quest for a specialized solver crafted to handle nonconvex optimization problems, particularly those associated with low-rank matrices.
- In contrast, our aim was to introduce an effective and efficient implementation of the nonconvex algorithm, complemented by dual acceleration techniques [44]. This tailored approach, namely the PBCD algorithm, is crafted to address unified nonconvex low-rank matrix learning problems without constraints. The novelty is supported by thorough and provable convergence results, enhancements in computational efficiency, and rigorous validations.

In this work, our proposed approach exhibits distinctive iterative characteristics that demonstrate effectiveness and efficiency when compared to widely used methods, including ADMM and its variants [18, 34–36], commonly applied to constrained problems. Additionally, we benchmark against PALM [56] and SPJIM [47], which primarily target unconstrained problems. Drawing insights from these pertinent minimization problems and optimization algorithms, we address two key considerations in the development of our approach:

- **Convergence to Problem (1)**: As the proximal parameters $\phi_{\mathbf{X}}$ and $\phi_{\mathbf{E}}$ approach zero, the solution of Problem (10) converges to the solution of the original Problem (1). A crucial perspective to consider is that theoretical analysis provides additional support for this assertion. It suggests that the approxi-

mate solution derived through the optimization process closely approximates the solution of the original problem, affirming convergence to the desired outcome. This analytical insight also reinforces our confidence in the effectiveness and reliability of the proposed methodology.

– **Higher Efficiency in Optimization**: By incorporating proximal terms, a continuation strategy, and RSVD into the iteration procedures of the PBCD algorithm, we can obtain closed-form solutions with a reduced number of variables. This strategy substantially boosts computational efficiency by minimizing the computational load at each iteration. As a result, it accelerates convergence and markedly enhances the overall efficiency of the optimization process, particularly for applications in image low-level and high-level visual tasks.

To address the aforementioned concerns, we furnish theoretical guarantees that validate the efficacy of our approach in capturing the desired solution and illuminate the relationship between Problems (1) and (10). Additionally, we introduce an optimized algorithm tailored specifically for Problem (10). This algorithm harnesses advanced techniques and capitalizes on the distinctive features of the problem to bolster computational efficiency, enhance convergence towards a high-quality solution, and ensure the practical feasibility of Problems (2)-(4) through a unified approach.

Remark 1 *Contrasting convex scenarios, nonconvex instances introduce a nearly unbiased estimator, as rigorously established in [60]. Delving into their convergence properties, nonconvex variants are associated with verifiable challenges [61]. These challenges span the sufficient decrease property, global subsequential convergence, and global convergence of the entire sequence. In contrast, convex formulations relying on nuclear norm minimization problems, being convex, typically exhibit the local convergence property, inherently implying global convergence [18] with theoretical assurances.*

2.1 Relations between (1) and (10)

By considering Problems (1) and (10) along with the nonnegativity property of the proximal terms and model parameters, we observe that the inequality

$$\Phi_{\lambda,\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}(\mathbf{X},\mathbf{E},\hat{\mathbf{X}},\hat{\mathbf{E}}) \geq \Phi_{\lambda,\mu}(\mathbf{X},\mathbf{E}) \tag{11}$$

holds easily, and equality holds only if $\phi_{\mathbf{X}} = \phi_{\mathbf{E}} = 0$. This implies that the objective function value of Problem (10) is always greater than or equal to the objective function value of Problem (1). Therefore, solving Problem (10) provides an upper bound for the optimal objective value of Problem (1). Additionally, for any given $\epsilon > 0$, by carefully choosing the proximal parameters $\phi_{\mathbf{X}}$ and $\phi_{\mathbf{E}}$, we can ensure that

$$\Phi_{\lambda,\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}(\mathbf{X},\mathbf{E},\hat{\mathbf{X}},\hat{\mathbf{E}}) \leq \Phi_{\lambda,\mu}(\mathbf{X},\mathbf{E}) + \epsilon \tag{12}$$

also holds. As ϵ approaches 0, the solution obtained by solving Problem (10) serves as an approximation to the solution of Problem (1). This approximate solution can be used as a surrogate solution based on the Majorization-Minimization (MM) strategy, as studied in [62–64]. The goal of this analysis is to establish the relationships between the various objective values and approximation solutions involved. Here, let $(\mathbf{X}_o^*, \mathbf{E}_o^*)$ and $(\mathbf{X}^*, \mathbf{E}^*)$ be the optimal solutions to Problems (1) and (10), respectively. Using the aforementioned inequalities (11) and (12), we can deduce that

$$\begin{aligned} 0 &\leq \Phi_{\lambda,\mu}(\mathbf{X}^*,\mathbf{E}^*) - \Phi_{\lambda,\mu}(\mathbf{X}_o^*,\mathbf{E}_o^*) \\ &\leq \Phi_{\lambda,\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}(\mathbf{X}^*,\mathbf{E}^*,\hat{\mathbf{X}},\hat{\mathbf{E}}) \\ &\quad - \Phi_{\lambda,\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}(\mathbf{X}_o^*,\mathbf{E}_o^*,\hat{\mathbf{X}},\hat{\mathbf{E}}) + \epsilon \leq \epsilon. \end{aligned} \tag{13}$$

This inequality demonstrates that the optimal solution $(\mathbf{X}^*, \mathbf{E}^*)$ obtained from solving Problem (10) is ϵ-optimal for the optimization Problem (1). In other words, solving Problem (10) not only provides an upper bound for Problem (1) but also guides the approximation of the optimal solution of Problem (1).

2.2 Iteration Procedures for Problem (10)

To optimize Problem (10), we utilize the nonconvex PBCD algorithm, which iteratively minimizes the objective function by updating blocks of variables while keeping others fixed, following a specific update scheme. The iterations can be summarized as follows:

$$\begin{aligned} \mathbf{X}_{k+1} = \operatorname{argmin}_{\mathbf{X}}\ & h_\lambda(\mathbf{X}) + g_\mu(\mathbf{D};\mathbf{X},\mathbf{E}_k) \\ & + \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}-\mathbf{X}_k\|_F^2, \end{aligned} \tag{14}$$

$$\begin{aligned} \mathbf{E}_{k+1} = \operatorname{argmin}_{\mathbf{E}}\ & f(\mathbf{E}) + g_\mu(\mathbf{D};\mathbf{X}_{k+1},\mathbf{E}) \\ & + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}-\mathbf{E}_k\|_F^2, \end{aligned} \tag{15}$$

where $\mathbf{X}_k$ and $\mathbf{E}_k$ represent the variables at the k-th iteration. Then, we have a direct consequence from (14) and (15), represented as

$$\begin{aligned} & h_\lambda(\mathbf{X}_{k+1}) + g_\mu(\mathbf{D};\mathbf{X}_{k+1},\mathbf{E}_k) + \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k+1}-\mathbf{X}_k\|_F^2 \\ & \leq h_\lambda(\mathbf{X}) + g_\mu(\mathbf{D};\mathbf{X},\mathbf{E}_k) + \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}-\mathbf{X}_k\|_F^2; \end{aligned} \tag{16}$$

$$f(\mathbf{E}_{k+1}) + g_\mu(\mathbf{D}; \mathbf{X}_{k+1}, \mathbf{E}_{k+1}) + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2$$
$$\leq f(\mathbf{E}) + g_\mu(\mathbf{D}; \mathbf{X}_{k+1}, \mathbf{E}) + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E} - \mathbf{E}_k\|_F^2. \quad (17)$$

The iteration procedures in this algorithm are intentionally designed to be simple and easily understandable, setting it apart from some first-order optimization algorithms. To expedite computations at the $(k+1)$-th iteration, we rely on both (16) and (17), incorporating proximity operators or analytic solvers as outlined in **Table** 1. Our primary focus is the efficient optimization of (14) and (15), achieved through **Algorithm** 1, with d being the fixed or known rank number and $\mu_0 = \hat{\mu}/\|\mathbf{D}\|_2$. Moreover, specific measurement criteria for $h_\lambda(\mathbf{X})$ and $f(\mathbf{E})$ are determined by considering the correlation between different low-rank structures and noise styles. These proper measurements play a crucial role in optimizing the objective function by incorporating the data fidelity term, especially in relation to the definition of the penalty function. For RMC, the penalty function is defined as $g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E}) = \frac{\mu}{2}\|\mathbf{X} + \mathbf{E} - \mathbf{D}\|_F^2$. On the other hand, both LRR and RMR employ the penalty function $g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E}) = \frac{\mu}{2}\|\mathbf{A}\mathbf{X} + \mathbf{E} - \mathbf{D}\|_F^2$. This diversity in constructing the matrix $\mathbf{A}$ leads to the exploration of various studies focusing on both single and multiple subspace models, as well as supervised and unsupervised learning tasks.

2.2.1 Solving (14) for Learnable Low-rank Matrix

To obtain closed-form solutions, we approach the problem by employing specific formulations of $g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E})$ tailored to three distinct minimization problems (2)-(4). These formulations are derived based on different low-rank structures and residual measurements. The solutions are presented as follows:

– For RMC, the closed-form solution is derived by

$$\mathbf{X}_{k+1} = \text{argmin}_{\mathbf{X}}\ h_\lambda(\mathbf{X}) + \frac{\phi_{\mathbf{X}} + \mu}{2}\|\mathbf{X} - \frac{\phi_{\mathbf{X}}\mathbf{X}_k + \mu(\mathbf{D} - \mathbf{E}_k)}{\phi_{\mathbf{X}} + \mu}\|_F^2. \quad (18)$$

– For LRR and RMR, the closed-form solution can also be achieved as (18), while we need to linearize the quadratic term for $g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E}_k)$ at $\mathbf{X}_k$ and adding a proximal term, it can be denoted as

$$\begin{aligned}&\frac{\mu}{2}\|\mathbf{A}\mathbf{X} + \mathbf{E}_k - \mathbf{D}\|_F^2 \\ &= \frac{\beta_{\mathbf{X}}\mu}{2}\|\mathbf{X} - \mathbf{X}_k\|_F^2 + \frac{\mu}{2}\|\mathbf{A}\mathbf{X}_k + \mathbf{E}_k - \mathbf{D}\|_F^2 \\ &\quad + \langle \mathbf{X} - \mathbf{X}_k, \mu\mathbf{A}^\top(\mathbf{A}\mathbf{X}_k + \mathbf{E}_k - \mathbf{D})\rangle \\ &= \frac{\beta_{\mathbf{X}}\mu}{2}\|\mathbf{X} - \hat{g}_{\mu,\beta_{\mathbf{X}}}(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_k)\|_F^2 + C, \end{aligned} \quad (19)$$

Algorithm 1 Nonconvex PBCD with RSVD

Input: $\mathbf{D}$, $\mathbf{X}_0$, $\mathbf{E}_0$
Parameter: λ, p, r, ρ, μ_0 and $k = 0$
Output: $\mathbf{X}^* \leftarrow \mathbf{X}_{k+1}$

```
while not converged do
    if facing RMC then
        update X_{k+1} by combining (18) with (22);
    else
        update X_{k+1} by combining (21) with (22);
    end if
    update E_{k+1} by combining (23) with Proposition 1;
end while
```

where we set $\beta_{\mathbf{X}} = 1.05 \times \max(\text{eig}(\mathbf{A}^\top\mathbf{A}))$ from [36, 43, 47, 56] commonly used as a proximal parameter and C is a constant unrelated to $\mathbf{X}$, and for notational simplicity, we define

$$\hat{g}_{\mu,\beta_{\mathbf{X}}}(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_k) = \mathbf{X}_k - \frac{1}{\beta_{\mathbf{X}}}\mathbf{A}^\top(\mathbf{A}\mathbf{X}_k + \mathbf{E}_k - \mathbf{D}). \quad (20)$$

Combining (14) with both (19) and (20), we can easily derive the following expression:

$$\mathbf{X}_{k+1} = \text{argmin}_{\mathbf{X}}\ h_\lambda(\mathbf{X}) + \frac{\phi_{\mathbf{X}} + \beta_{\mathbf{X}}\mu}{2} \times \|\mathbf{X} - \frac{\phi_{\mathbf{X}}\mathbf{X}_k + \beta_{\mathbf{X}}\mu\hat{g}_{\mu,\beta_{\mathbf{X}}}(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_k)}{\phi_{\mathbf{X}} + \beta_{\mathbf{X}}\mu}\|_F^2. \quad (21)$$

The primary challenge in addressing the subproblems (18) and (21) is the computation of proximal operators, which encompasses rank relaxations and often involves performing the full SVD. However, direct computation of all singular values can be computationally demanding, especially for large-scale matrices. To expedite the update of $\mathbf{X}_{k+1}$ and enhance efficiency, we introduce an acceleration technique known as RSVD. This strategy alleviates computational complexity, leading to faster updates of $\mathbf{X}_{k+1}$. Formally, we describe both the proximal operator and RSVD as follows:

Proposition 2 *Assume that rank*$(\mathbf{X}) = r < d \leq min\ (m, n)$, $\mathbf{Q} \in \mathbb{R}^{m\times d}$ *is an orthogonal matrix satisfying* $\mathbf{Q}^\top\mathbf{Q} = \boldsymbol{I}$, *and* $\mathbf{Y} \in \mathbb{R}^{d\times n}$ *such that* $\mathbf{X} = \mathbf{Q}\mathbf{Y}$. *Then, the proximal operator* $Prox_\lambda^h(\mathbf{Q}^\top\mathbf{Z})$ *is the solution to*

$$Prox_\lambda^h(\mathbf{Q}^\top\mathbf{Z}) = argmin_{\mathbf{Y}} h_\lambda(\mathbf{Y}) + \frac{1}{2}\|\mathbf{Y} - \mathbf{Q}^\top\mathbf{Z}\|_F^2, \quad (22)$$

where we have $h_\lambda(\mathbf{Y}) = \sum_{i=1}^d h_\lambda(\sigma_i(\mathbf{Y}))$, $\|\mathbf{X} - \mathbf{Z}\|_F^2 = \|\mathbf{Y} - \mathbf{Q}^\top\mathbf{Z}\|_F^2$, *and the matrices* $\mathbf{X}$ *and* $\mathbf{Y}$ *have the same singular values. Thus, by finding the optimal* $\mathbf{Y}^* = Prox_\lambda^h(\mathbf{Q}^\top\mathbf{Z})$, *we have the desired matrix* $\mathbf{X}^* = \mathbf{Q}\mathbf{Y}^*$.

It is well-known that the RSVD is an efficient technique that involves randomly projecting a large-scale matrix onto a lower-dimensional subspace. Subsequently,

it computes an approximate singular value decomposition (SVD) using the orthogonal basis matrix $\mathbf{Q}$ and the QR decomposition associated with that subspace to learn $\mathbf{X}$, as outlined in **Proposition** 2. The detailed computation process is provided in **Algorithm** 2, where $\hat{\mathbf{Y}}$ represents the given large-scale matrix. By leveraging RSVD, the computational complexity is significantly reduced, especially when $d \leq \min(m, n)$, as it requires fewer singular values to be calculated, resulting in faster running times. This scalable and tractable formulation underscores the importance of the proximal operator in finding the optimal solution and approximating the matrix using generalized singular value thresholding (GSVT) [17, 65]. Algorithms employing such proximal operators, extended from Equation (22), have gained popularity in addressing large-scale low-rank matrix problems due to their superior performance compared to existing techniques like the power method [66] and PROPACK [67].

Algorithm 2 RSVD for Problems (18) and (21)

Input: large-scale matrix $\hat{\mathbf{Y}}$
Parameter: trunction rank number r;
power of iteration s; oversmpling parameter t;
Output: $\mathbf{Y}^*$

$ny \leftarrow \text{size}(\hat{\mathbf{Y}}, 2)$
$\mathbf{P} \leftarrow \text{randn}(ny, r + t)$
$\mathbf{Z} \leftarrow \hat{\mathbf{Y}}\mathbf{P}$
while $k = 1$ to s **do**
  $\mathbf{Z} \leftarrow \hat{\mathbf{Y}}(\hat{\mathbf{Y}}^\top \mathbf{Z})$
end while
$[\mathbf{Z}_1, \mathbf{Z}_2] \leftarrow \text{qr}(\mathbf{Z}, 0)$
$\bar{\mathbf{Y}} \leftarrow \mathbf{Z}_1^\top \hat{\mathbf{Y}}$;
$[\mathbf{U}_{\bar{\mathbf{Y}}}, \mathbf{S}, \mathbf{V}] \leftarrow \text{SVD}(\bar{\mathbf{Y}}, \text{'econ'})$;
$\mathbf{U} \leftarrow \mathbf{Q}\mathbf{U}_{\bar{\mathbf{Y}}}$;
$\mathbf{ss} = \text{diag}(\mathbf{S})$;
$\mathbf{vv} = \mathbf{GSVT}(\mathbf{ss}, p, \lambda)$;
$\text{indx} = \text{find}(\mathbf{vv})$;
$\mathbf{Y}^* = \mathbf{U}(:, \text{indx})\text{diag}(\mathbf{vv}(\text{indx}))\mathbf{V}(:, \text{indx})^\top$;

2.2.2 Solving (14) for Learnable Residual Matrix

The utilization of proximal operators plays a crucial role in effectively solving minimization problems that involve the relaxations of the ℓ_0-norm, $\ell_{2,0}$-norm, and rank function. These operators serve as powerful tools for promoting (group) sparsity and low-rankness, allowing us to approximate optimization problems and facilitate efficient computation. By leveraging these operators, we can derive the following update for $\mathbf{E}$:

$$\mathbf{E}_{k+1} = \text{argmin}_{\mathbf{E}} \ f(\mathbf{E}) + \frac{\phi_{\mathbf{E}} + \mu}{2} \|\mathbf{E} - \frac{\phi_{\mathbf{E}}\mathbf{E}_k + \mu(\mathbf{D} - \mathbf{A}\mathbf{X}_{k+1})}{\phi_{\mathbf{E}} + \mu}\|_F^2, \quad (23)$$

where it is important to note that certain methods, such as RMC, use a fixed matrix $\mathbf{A}$ set to $\mathbf{I}$. However, methods like LRR and RMR derive the matrix $\mathbf{A}$ from clustering and training samples, as extensively studied in [35, 40, 47]. This difference in constructing the matrix $\mathbf{A}$ highlights the diversity of techniques used in the field of existing low-rank matrix learning problems.

The update equation incorporates the desired (group) sparsity and low-rank properties, enabling efficient and effective computation of the solution in a standard parallel manner for $\mathbf{E}$ and approximation of the underlying optimization problem. Building upon the use of non-convex functions for coefficient measurements, such as the ℓ_p-norm, MCP, SCAD, and their extensions from **Table** 1 to relax the ℓ_0-norm, $\ell_{2,0}$-norm, and rank function, we employ these functions to measure $f(\mathbf{E})$ and capture various noise styles illustrated in **Fig**. 1. The use of these processing methods enables efficient computation and optimization of the problems, as discussed in [21, 35, 40, 68]. The use of proximal operators serves as a powerful tool for solving these problems and guarantees the existence of closed-form solutions, as stated in **Proposition** 1 for higher computational efficiency.

2.2.3 Speed-up with Continuation Strategy

Different from the RSVD technique aimed at reducing computational complexity at each iteration, **Algorithm** 1 can be further accelerated from a different perspective. Specifically, to enhance the convergence of the alternating scheme with update rules (14) and (15), a continuation strategy for the penalty parameter μ can be employed. This strategy, inspired by [16, 17, 45, 53], involves gradually increasing the penalty parameter during the optimization process. The rate of increase is controlled by a parameter $\rho > 1$, which allows for a decrease in the total number of iterations required, and larger values of ρ result in faster convergence speed while causing lower solution quality. Then, it is easy to achieve

$$\mu_{k+1} = \min(\rho\mu_k, \mu_{\max}), \quad (24)$$

where by adopting this strategy, the algorithm can efficiently navigate the optimization landscape and converge to the desired solution. We set the initial value of μ to a small positive value and increase it at each iteration until it reaches a pre-defined maximum threshold. This continuation strategy promotes faster convergence and reduces the computational burden while still ensuring high-quality solutions. Especially, as μ_k approaches the maximum value, the penalty function $g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E})$ decreases, and the objective problem can be approximately transformed into a constrained problem as described in Problems (2)-(4). To better enhance compre-

Algorithm 3 Algorithm 1 with Continuation

Input: $\mathbf{D}$, $\mathbf{X}_0$, $\mathbf{E}_0$, $\mu_{\max}$
Parameter: λ, p, d, ρ, μ_0 and $k = 0$
Output: $\mathbf{X}^* \leftarrow \mathbf{X}_{k+1}$

1: **while** $\mu < \mu_{\max}$ **do**
2: update $(\mathbf{X}_{k+1}, \mathbf{E}_{k+1})$ by **Algorithms** 1 and 2;
3: update μ_{k+1} by $\rho\mu_k$ with $\rho > 1$;
4: **end while**

hension, we further consolidate the presented iteration procedures into a unified framework, as depicted in **Algorithm** 3. This algorithm efficiently accelerates the convergence speed of iterations for solving the general problem formulation (10).

2.2.4 Analysis of Computational Reduction Techniques

To demonstrate a more in-depth discussion regarding the differences and a thorough comparison of the proposed acceleration strategies, we further highlight both the advantages and limitations, offering a more comprehensive perspective on the main contributions and superior performance of our proposed techniques from two following analysis.

- On the one hand, we introduce a sophisticated approach that integrates a continuation strategy and RSVD to enhance the computational efficiency of the nonconvex PBCD algorithm, as the extension of [17, 61]. This fusion notably improves the computational efficiency and scalability of our proposed approaches, facilitating its successful application across various low-rank matrix learning problems.
- On the other hand, we formulated RSVD based on an optimization algorithm, steering clear of the introduction of excessive variables. This deviates from traditional matrix factorization strategies [8, 25, 40, 46]. Additionally, the adoption of the continuation strategy proves advantageous by limiting the number of iterations, setting it apart from extrapolation and Nesterov's techniques [21, 47, 69], which involve additional variables.

It is essential to engage in a comprehensive discussion to address potential challenges and intricacies associated with implementing and discussing our proposed complexity reduction strategies. Subsequently, we will delve into a detailed explanation of these strategies, considering four key aspects as follows:

- Algorithmic Integration: Integrating RSVD and continuity techniques demands meticulous algorithmic integration. Ensuring seamless collaboration between these components may present challenges in terms of synchronization and proper functioning. Careful consideration and testing are necessary to guarantee the smooth interaction and mutual effectiveness of these algorithmic elements.
- Algorithmic Stability: Addressing the stability of the overall algorithm becomes paramount when combining RSVD and continuity techniques. The potential emergence of instabilities or convergence issues underscores the need for developing strategies that enhance the stability of the algorithm. Robustness in the face of various scenarios and datasets is essential for the algorithm's reliability.
- Acceleration Scalability: Noting that computational complexity reduction strategies scale effectively with larger datasets or more intricate tasks is imperative. Challenges may arise in maintaining efficiency and computational speed as the scale of the problem increases. Therefore, achieving scalability is crucial for the applicability of these strategies to real-world, large-scale scenarios.
- Parameter Tuning: Observing that fine-tuning parameters for RSVD and continuity techniques to attain optimal performance can be a nuanced task. The delicate interplay between these hyper-parameters requires careful consideration to strike the right balance for efficient complexity reduction. Thorough experimentation and validation are essential to identify parameter choices that lead to optimal results.

As a whole, our main objective is to enhance readers' understanding by incorporating these improvements and providing detailed explanations that elaborate on the analysis of computational complexity reduction. This integration is naturally designed to make the central contributions of this work more accessible and comprehensible to the readers.

2.2.5 Stopping Condition

The iteration procedure in **Algorithms** 1 and 3 does not involve an inner loop. We initialize the primal variables and the associated parameters using the technique suggested in [25, 36, 40]. To terminate the algorithms, we compute the stopping criterion as follows:

$$\frac{\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F}{\|\mathbf{X}_k\|_F} < \epsilon, \tag{25}$$

where $0 < \epsilon \ll 1$. Once this criterion is met, implying that the change in $\mathbf{X}$ between consecutive iterations is sufficiently small, the algorithms will be terminated in solving revised Problems (2)-(4), efficiently. Thus, the proposed algorithm effectively achieves the desired low-rank matrices. This successful attainment is a critical step towards solving various problems in the field of low-rank matrix learning and image low-level and high-level vision applications.

2.2.6 Connections Between Low-rank Matrices and Evaluation Criteria for Different Application Tasks

To better address the image low-level and high-level vision tasks of inpainting, classification, and clustering, it becomes crucial to incorporate task-specific post-processing steps for the optimal utilization of the learned low-rank matrices obtained from (14). These customized post-processing steps are designed to meet the distinct requirements of each task, ensuring that the matrices are applied effectively within their respective contexts and in conjunction with the developed optimization **Algorithms** 1 and 3. Subsequently, we provide detailed descriptions of low-rank matrices along with the associated evaluation metrics.

- Using the learned low-rank matrix, denoted as $\mathbf{X}^*$, for image inpainting involves utilizing the inherent structure and dependencies within the image data to fill in missing or corrupted regions. The low-rank matrix, obtained through the RMC method in Problem (2), captures the underlying relationships in the image. Then, the missing regions can be filled in by incorporating information from the remaining image data. This process is useful for computing the Peak Signal-to-Noise Ratio (PSNR), i.e.,

$$\mathrm{PSNR} = 10\log_{10}\Big(\frac{255^2}{\frac{1}{3mn}\sum_{i=1}^{3}\|\mathbf{X}_i^* - \mathbf{D}_i\|_F^2}\Big), \quad (26)$$

  where the matrix size is $m \times n$, and the inpainting image generated using the learned low-rank matrix $\mathbf{X}^*$ is compared with the original image $\mathbf{D}$, as computed and recovered in previous studies [52, 59].
- In the content of classification of RMR, the learned low-rank matrix obtained from Problem (3) serves as a compact and informative representation of input images. To utilize this matrix for classification, we compute class-wise error matrices for all testing samples and assign a final label to each sample $\mathbf{D}_i$ based on the criterion [42, 70]:

$$\mathbf{Label}(\mathbf{D}_i) = \mathrm{argmin}_j \frac{h(\mathcal{M}(\mathbf{D}_i - \mathbf{A}\mathbf{X}^*_{i,c_j}))}{\|\mathbf{X}^*_{i,c_j}\|_2}, \quad (27)$$

  where we calculate the difference between the testing sample matrix $\mathcal{M}(\mathbf{D}_i)$ and the reconstructed image matrix $\mathcal{M}(\mathbf{A}\mathbf{X}^*_{i,c_j})$. This difference is divided by the ℓ_2-norm of the $\mathbf{X}^*_{i,c_j}$ associated with the i-th testing sample for the j-th class (c_j). The label $\mathbf{Label}(\mathbf{D}_i)$ is determined as the index j that minimizes the expression in (27) across all classes, which ensures that the chosen label corresponds to the smallest value.
- For clustering purposes using the LRR method, low-rank matrices play a crucial role in capturing the underlying structure and similarity patterns in data. By leveraging the low-rank solution, clustering algorithms such as k-means or spectral clustering can effectively group similar images together. This facilitates the discovery of meaningful clusters or subgroups, which are desired outcomes. Building on the insights from previous subspace clustering methods that utilize the representation coefficient matrix $\mathbf{X}^*$ to construct a clustering graph, we employ the learned solution $\mathbf{X}^*$ obtained from solving Problem (4). This solution is then used to construct a graph with weights [11, 25], denoted as

$$\mathbf{W}^* = \frac{|\mathbf{X}^*| + |\mathbf{X}^*|}{2}, \quad (28)$$

  where to partition the data points into multiple clusters, we utilize the popular Normalized Cut algorithm, which is based on the condition (28). For evaluating the quality of the resulting performance, we compute the clustering accuracy (ACC) and Normalized Mutual Information (NMI) metrics. The mathematical formulas for calculating these metrics in detail can be found in [25, 71, 72]. These metrics provide quantitative measures to assess the effectiveness of the clustering algorithm in producing meaningful and reliable cluster assignments.

3 Theoretical Analysis for Convergence

Before delving into the theoretical results, it is crucial to offer a brief overview of key concepts that play a pivotal role in establishing local and global convergence guarantees for nonconvex optimization algorithms in generalized assumptions. These fundamental concepts encompass the subdifferential, critical point, distance, and the uniformized KŁ inequality, each of which is elucidated as follows:

Definition 1 *[56, 73] Let $\varrho(\mathbf{X}) : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. The limiting subdifferential of $\varrho(\mathbf{X})$ at a point $\mathbf{X} \in \mathbb{R}^n$, denoted by $\partial\varrho(\mathbf{X})$, is defined as the closure of the set*

$$\partial\varrho(\mathbf{X}) = \big\{\mathbf{u} \in \mathbb{R}^n : \exists\ \mathbf{X}_k \to \mathbf{X},\ \varrho(\mathbf{X}_k) \to \varrho(\mathbf{X}), \\ \text{and } \mathbf{u} \leftarrow \partial\hat{\varrho}(\mathbf{X}_k) \ni \mathbf{u}_k \text{ as } k \to +\infty\big\}, \quad (29)$$

where $\partial\hat{\varrho}(\mathbf{X})$ is the Frèchet subdifferential of $\varrho(\mathbf{X})$ at a given point $\mathbf{X} \in dom\varrho$. The set $\partial\hat{\varrho}(\mathbf{X})$ consists of all $\mathbf{u} \in \mathbb{R}^n$ that satisfy

$$\liminf_{\mathbf{y}\neq\mathbf{X},\mathbf{y}\to\mathbf{X}} \frac{\varrho(\mathbf{y}) - \varrho(\mathbf{X}) - \langle\mathbf{u}, \mathbf{y} - \mathbf{X}\rangle}{\|\mathbf{y} - \mathbf{X}\|} \geq 0. \quad (30)$$

Remark 2 *[56, 73] (**i**) According to the Fermat's rule in the nonsmooth content, if $\mathbf{X} \in \mathbb{R}^n$ is a local minimizer of $\varrho(\mathbf{X})$, then we have $\mathbf{0} \in \partial\varrho(\mathbf{X})$.*

*(**ii**) Consider a sequence $\{(\mathbf{X}_k, \mathbf{u}_k)\}$ that converges to $(\mathbf{X}, \mathbf{u})$ and $\varrho(\mathbf{X}_k) \to \varrho(\mathbf{X})$ as $k \to +\infty$. If $\mathbf{u}_k \in \partial\varrho(\mathbf{X}_k)$, then we have $\mathbf{u} \in \partial\varrho(\mathbf{X})$.*

*(**iii**) For a continuous differentiable function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$, we have $\partial(f + \varrho)(\mathbf{X}) = \nabla f(\mathbf{X}) + \partial\varrho(\mathbf{X})$.*

*(**iv**) Points whose subdifferential contains $\mathbf{0}$ are called (limiting-)critical points.*

Definition 2 *[56, 73] Consider a compact subset $\Gamma \subset \mathbb{R}^n$ and a point $\mathbf{X} \in \mathbb{R}^n$. The distance from $\mathbf{X}$ to Γ is defined as*

$$dist(\mathbf{X}, \Gamma) = inf\{\|\mathbf{X} - \mathbf{y}\| : \mathbf{y} \in \Gamma\}. \tag{31}$$

Here, $dist(\mathbf{X}, \Gamma) = +\infty$ if $\Gamma = \emptyset$, and $dist(\mathbf{X}, \Gamma) = 0$ if and only if $\mathbf{X} \in \Gamma$.

Definition 3 *(KŁ inequality) [56, 73] A function $\varrho(\cdot) : \mathbb{R}^n \to (-\infty, +\infty]$ is said to satisfy the KŁ property at $\bar{\mathbf{X}} \in dom\partial\varrho := \{\mathbf{X} \in \mathbb{R}^n : \partial\varrho(\mathbf{X}) \neq \emptyset\}$ if there exists $\eta \in (0, +\infty]$, a neighborhood $\mathbf{U}$ of $\bar{\mathbf{X}}$, and a function $\Theta \in \Xi_\eta$ such that for all $\mathbf{X} \in \mathbf{U} \cap \{\mathbf{X} \in \mathbb{R}^n : \varrho(\bar{\mathbf{X}}) < \varrho(\mathbf{X}) < \varrho(\bar{\mathbf{X}}) + \eta\}$, the following KŁ inequality holds:*

$$\Theta'(\varrho(\mathbf{X}) - \varrho(\bar{\mathbf{X}}))dist(\mathbf{0}, \partial\varrho(\mathbf{X})) > 1, \tag{32}$$

where Φ_η is a class of functions $\Theta : [0, \eta) \to \mathbb{R}^+$ satisfying the following conditions:

*(**i**) Θ is concave and C^1 on $(0, \eta)$;*

*(**ii**) Θ is continuous at 0 and $\phi(0) = 0$;*

*(**iii**) $\Theta'(\mathbf{X}) > 0$ for all $x \in (0, \eta)$.*

Proposition 3 *(Uniformized KŁ inequality) [56] Let $\varrho(\mathbf{X})$ be a proper and lower semicontinuous function. Assume that $\varrho(\mathbf{X})$ is constant on a compact set Γ and satisfies the KŁ property at each point of Γ. Then $\varrho(\mathbf{X})$ is called a KŁ function, and there exist $\epsilon > 0$, $\eta > 0$, and $\Theta \in \Phi_\eta$ such that for all $\bar{\mathbf{X}} \in \Gamma$ and all $\mathbf{X} \in \mathbb{R}^n$ in the following intersection:*

$$\begin{aligned}&\{\mathbf{X} : dist(\mathbf{X}, \Gamma) < \epsilon\}\\ &\cap \{\mathbf{X} : \varrho(\bar{\mathbf{X}}) < \varrho(\mathbf{X}) < \varrho(\bar{\mathbf{X}}) + \eta\},\end{aligned} \tag{33}$$

which shows that the inequality (32) still holds.

It should be noted that **Definition** 1 and **Remark** 2 are crucial for analyzing the objective function, which can be either differentiable or nondifferentiable. **Definition** 2 provides a computational formula for point-to-set distance. On the other hand, **Proposition** 3 contributes to the theoretical analysis of global convergence guarantees and can be seen as a modified version of **Definition** 3. They are associated with the widely used KŁ property, which generalizes the Łojasiewicz gradient inequality to nonconvex nonsmooth functions.

By building upon the preliminary concepts and theoretical findings discussed earlier [21, 47, 56], we can establish convergence guarantees to stationary points for generalized nonconvex functions that capture both low-rank and residual components. Our theoretical analysis focuses on finding stationary points of the objective function and utilizes the well-known KŁ property. This property simplifies the main arguments in the local and global convergence analysis of the variable sequences generated by **Algorithms** 1 and 3. The KŁ property, being closely related to the optimization algorithm, provides valuable insights into its behavior and ensures convergence towards desirable solutions.

Theorem 1 *Suppose that the assumptions (**A1**) and (**A2**) hold. Let a variable sequence $\{(\mathbf{X}_k, \mathbf{E}_k)\}$ be generated by **Algorithm** 1 for Problem (10). For positive values $\phi_{\mathbf{X}}$ and $\phi_{\mathbf{E}}$, then we have the following assertions, which can be made:*

*(**i**) (**Sufficient decrease property**) The objective function sequence $\{\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)\}$ of Problem (1) exhibits non-increasing behavior, expressed as*

$$\begin{aligned}&\Phi_{\lambda,\mu}(\mathbf{X}_{k+1}, \mathbf{E}_{k+1}) - \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)\\ &\leq -\frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2 - \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2.\end{aligned} \tag{34}$$

*(**ii**) (**Boundedness and convergence**) As $k \to +\infty$, the following limits hold:*

$$\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F \to 0, \quad \|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F \to 0. \tag{35}$$

***Proof** (**i**) By substituting $\mathbf{X} = \mathbf{X}_k$ and $\mathbf{E} = \mathbf{E}_k$ into the right side of inequalities (16) and (17), and considering the definition of $\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E})$ in Problem (1), we can sum both sides of the inequalities to obtain the following expression:*

$$\begin{aligned}&\Phi_{\lambda,\mu}(\mathbf{X}_{k+1}, \mathbf{E}_{k+1}) + \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2\\ &+ \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2 \leq \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k),\end{aligned} \tag{36}$$

which implies that the inequality (34) naturally holds. Then, it is easy to conclude the sufficient decrease property for the objective function sequence.

*(**ii**) The sequence $\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)$ is proved to be non-increasing based on (34) or (36). Moreover, since $\Phi_{\lambda,\mu}(\cdot, \cdot)$ is satisfied to be bounded from below, it converges to a real number denoted by $\Phi_{\lambda,\mu}(\mathbf{X}^*, \mathbf{E}^*) \hat{=} \Phi^*_{\lambda,\mu}$. Without loss of generality, assume that $\mathbf{X}^*$ and $\mathbf{E}^*$ represent the limiting points of the sequences $\{\mathbf{X}_k\}$ and $\{\mathbf{E}_k\}$.*

By summing the inequalities (34) or (36) over k from 1 to $+\infty$ and rearranging the terms on the left-hand side, we obtain the following expression

$$\begin{aligned}&\Phi^*_{\lambda,\mu} - \Phi_{\lambda,\mu}(\mathbf{X}_1, \mathbf{E}_1)\\ &= \sum_{k=1}^{\infty} \Phi_{\lambda,\mu}(\mathbf{X}_{k+1}, \mathbf{E}_{k+1}) - \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)\\ &\leq -\sum_{k=1}^{\infty} \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2 + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2. \quad (37)\end{aligned}$$

Hence, it follows from (37) that

$$\sum_{k=1}^{\infty} \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2 + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2 < +\infty, \quad (38)$$

which implies that (35) holds easily. □

Theorem 2 *(**Global subsequential convergence**) Suppose that the sum of both $h_\lambda(\mathbf{X})$ and $f(\mathbf{E})$ is coercive, i.e., $h_\lambda(\mathbf{X}) + f(\mathbf{E}) \to +\infty$ if and only if $\|\mathbf{X}\|_F \to +\infty$ and $\|\mathbf{E}\|_F \to +\infty$. The other assumptions are the same as in **Theorem** 1. Then, we can conclude that any cluster point $(\mathbf{X}^*, \mathbf{E}^*)$ of a sequence $\{(\mathbf{X}_k, \mathbf{E}_k)\}$ is a stationary point of Problem (1).*

Proof *For the coercive property, the boundedness of the sequence $\{(\mathbf{X}_k, \mathbf{E}_k)\}$ is a direct consequence, indicating the existence of a cluster point.*

Without loss of generality, let $(\mathbf{X}^*, \mathbf{E}^*)$ be a cluster point of the sequence $\{(\mathbf{X}_k, \mathbf{E}_k)\}$, and consider a convergent subsequence $\{(\mathbf{X}_{k_j}, \mathbf{E}_{k_j})\}$ such that

$$\lim_{j\to+\infty} (\mathbf{X}_{k_j}, \mathbf{E}_{k_j}) = (\mathbf{X}^*, \mathbf{E}^*), \quad (39)$$

*where we note that $\{(\mathbf{X}_{k_j+1}, \mathbf{E}_{k_j+1}, \mathbf{X}_{k_j}, \mathbf{E}_{k_j})\}$ satisfies the optimality conditions for $\Phi_{\lambda,\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}(\mathbf{X}, \mathbf{E}, \hat{\mathbf{X}}, \hat{\mathbf{E}})$ in subproblems (14) and (15). Therefore, we can conclude from **Remark** 2 that*

$$\begin{cases} \mathbf{0} \in \partial h_\lambda(\mathbf{X}_{k_j+1}) + \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_{k_j+1}, \mathbf{E}_{k_j}) \\ \qquad +\phi_{\mathbf{X}}(\mathbf{X}_{k_j+1} - \mathbf{X}_{k_j}); & (40) \\ \mathbf{0} \in \partial f(\mathbf{E}_{k_j+1}) + \nabla_{\mathbf{E}} g_\mu(\mathbf{D}; \mathbf{X}_{k_j+1}, \mathbf{E}_{k_j+1}) \\ \qquad +\phi_{\mathbf{E}}(\mathbf{E}_{k_j+1} - \mathbf{E}_{k_j}), & (41) \end{cases}$$

*where as $j \to +\infty$, we observe that $\mathbf{X}_{k_j+1} - \mathbf{X}_{k_j}$ and $\mathbf{E}_{k_j+1} - \mathbf{E}_{k_j}$ both tend to $\mathbf{0}$, as concluded from (**ii**) of **Theorem** 1. Additionally, the differentiability of $g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E})$ implies that*

$$\begin{cases} \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_{k_j+1}, \mathbf{E}_{k_j}) \to \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}^*, \mathbf{E}^*); & (42) \\ \nabla_{\mathbf{E}} g_\mu(\mathbf{D}; \mathbf{X}_{k_j+1}, \mathbf{E}_{k_j+1}) \to \nabla_{\mathbf{E}} g_\mu(\mathbf{D}; \mathbf{X}^*, \mathbf{E}^*), & (43) \end{cases}$$

*hold naturally from the **Remark** 2. In addition, following from the fact that $h_\lambda(\mathbf{X}_k)$ and $f(\mathbf{E}_k)$ are closed, proper, and lower semi-continuous nonconvex functions, then it follows from **Definition** 1 that we have*

$$h_\lambda(\mathbf{X}^*) \leq \liminf_{j\to+\infty} h_\lambda(\mathbf{X}_{k_j}), \quad (44)$$

$$f(\mathbf{E}^*) \leq \liminf_{j\to+\infty} f(\mathbf{E}_{k_j}). \quad (45)$$

By substituting $\mathbf{X} = \mathbf{X}^*$ and $\mathbf{E} = \mathbf{E}^*$ into the inequalities (16) and (17), with $k = k_j$, we can obtain the following expressions:

$$\begin{aligned}&h_\lambda(\mathbf{X}_{k_j+1}) + g_\mu(\mathbf{D}; \mathbf{X}_{k_j+1}, \mathbf{E}_{k_j}) + \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k_j+1} + \mathbf{X}_{k_j}\|_F^2\\ &\leq h_\lambda(\mathbf{X}^*) + g_\mu(\mathbf{D}; \mathbf{X}^*, \mathbf{E}_{k_j}) + \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}^* - \mathbf{X}_{k_j}\|_F^2, \quad (46)\\ &f(\mathbf{E}_{k_j+1}) + g_\mu(\mathbf{D}; \mathbf{X}_{k_j+1}, \mathbf{E}_{k_j+1}) + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k_j+1} - \mathbf{E}_{k_j}\|_F^2\\ &\leq f(\mathbf{E}^*) + g_\mu(\mathbf{D}; \mathbf{X}_{k_j+1}, \mathbf{E}^*) + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}^* - \mathbf{E}_{k_j}\|_F^2. \quad (47)\end{aligned}$$

Then, taking limit in both sides of above inequalities, we see from above-given conclusion (35) that

$$\limsup_{j\to+\infty} h_\lambda(\mathbf{X}_{k_j}) = \limsup_{j\to+\infty} h_\lambda(\mathbf{X}_{k_j+1}) \leq h_\lambda(\mathbf{X}^*), \quad (48)$$

$$\limsup_{j\to+\infty} f(\mathbf{E}_{k_j}) = \limsup_{j\to+\infty} f(\mathbf{E}_{k_j+1}) \leq f(\mathbf{E}^*). \quad (49)$$

By combining (44) and (45) with (48) and (49), it is easy to obtain

$$h_\lambda(\mathbf{X}^*) = \lim_{j\to+\infty} h_\lambda(\mathbf{X}_{k_j}); \quad (50)$$

$$f(\mathbf{E}^*) = \lim_{j\to+\infty} f(\mathbf{E}_{k_j}), \quad (51)$$

which further implies that as $j \to +\infty$, it is not hard to achieve that

$$\partial h_\lambda(\mathbf{X}_{k_j+1}) \to \partial h_\lambda(\mathbf{X}^*); \quad (52)$$

$$\partial f(\mathbf{E}_{k_j+1}) \to \partial f(\mathbf{E}^*), \quad (53)$$

*hold through **Remark** 2. Furthermore, borrowing from (40)-(43), (52), and (53), it is easy to get that*

$$\begin{cases} \mathbf{0} \in \partial h_\lambda(\mathbf{X}^*) + \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}^*, \mathbf{E}^*); & (54) \\ \mathbf{0} \in \partial f(\mathbf{E}^*) + \nabla_{\mathbf{E}} g_\mu(\mathbf{D}; \mathbf{X}^*, \mathbf{E}^*), & (55) \end{cases}$$

holds easily. This expression confirms that $(\mathbf{X}^*, \mathbf{E}^*)$ is a stationary point of $\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E})$, validating the convergence property of the generated subsequence. □

Theorem 3 *(**Global convergence of the entire sequence**) Let $\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E})$ be a KL function satisfying the assumptions in **Theorems** 1 and 2. Then, we have the assertion, i.e., If there exists a point $(\mathbf{X}^*, \mathbf{E}^*) \in \Gamma$ at which $\Phi_{\lambda,\mu}(\cdot, \cdot)$ attains its minimum, then the sequence $\{(\mathbf{X}_k, \mathbf{E}_k)\}$ generated by **Algorithm** 1 converges to the stationary point $(\mathbf{X}^*, \mathbf{E}^*)$ of Problem (1).*

Proof *Since the objective function sequence* $\{\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)\}$ *is non-increasing, as proven in* ***Theorem*** *1, and it is also bounded from below, we can conclude that*

$$\lim_{k\to+\infty} \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) = \Phi_{\lambda,\mu}(\mathbf{X}^*, \mathbf{E}^*) \hat{=} \Phi^*_{\lambda,\mu} \tag{56}$$

exists. In the following, we will consider two cases.

Case 1*: Suppose there exists an integer* $\hat{k} \geq 1$ *for which* $\Phi_{\lambda,\mu}(\mathbf{X}_{\hat{k}}, \mathbf{E}_{\hat{k}}) = \Phi^*_{\lambda,\mu}$. *The non-increasing property of the function sequence in Problem (1) implies that* $\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) = \Phi^*_{\lambda,\mu}$ *holds for all* $k \geq \hat{k}$. *Through (34), we have* $\mathbf{X}_{k+l} = \mathbf{X}_{\hat{k}}$ *and* $\mathbf{E}_{k+l} = \mathbf{E}_{\hat{k}}$ *for all* $l \geq 1$. *Thus, the sequences* $\{(\mathbf{X}_k, \mathbf{E}_k)\}$ *converge finitely. Furthermore, we can deduce that*

$$\sum_{k=\hat{k}}^{+\infty} \|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2 + \|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2 \leq +\infty. \tag{57}$$

This implies that the generated variable sequence converges globally in this case.

Case 2*: Suppose that there exists an integer* $k \geq 1$ *for which* $\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) > \Phi^*_{\lambda,\mu}$. *In this case, we divide the process of proofs into three steps and present a detailed explanation as follows:*

Step 1*: We will prove that* $\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E})$ *is constant on the set of cluster points, denoted as* Γ, *of the generated sequence* $\{(\mathbf{X}_k, \mathbf{E}_k)\}$ *and then apply the uniformized KŁ property provided in* ***Proposition*** *3.*

Let any $(\bar{\mathbf{X}}, \bar{\mathbf{E}}) \in \Gamma$ *be a cluster point of the sequence* $\{(\mathbf{X}_k, \mathbf{E}_k)\}$. *Since the sequence* $\{\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)\}$ *is non-increasing and bounded below (lower semicontinuous), it follows that* $\lim_{k\to+\infty} \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) = \Phi_{\lambda,\mu}(\bar{\mathbf{X}}, \bar{\mathbf{E}})$ *exists.*

Let us suppose for contradiction that $\Phi_{\lambda,\mu}(\bar{\mathbf{X}}, \bar{\mathbf{E}}) \neq \Phi^*_{\lambda,\mu}$, *where without loss of generality, we can assume* $\Phi_{\lambda,\mu}(\bar{\mathbf{X}}, \bar{\mathbf{E}}) > \Phi^*_{\lambda,\mu}$. *By exploiting the fact that* $\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)$ *is non-increasing again, there must exist an integer* $N \geq k$ *such that*

$$\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) \geq \Phi_{\lambda,\mu}(\mathbf{X}_N, \mathbf{E}_N) > \Phi_{\lambda,\mu}(\bar{\mathbf{X}}, \bar{\mathbf{E}}). \tag{58}$$

However, this contradicts the fact that $\lim_{k\to\infty} \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) = \Phi_{\lambda,\mu}(\bar{\mathbf{X}}, \bar{\mathbf{E}})$. *Thus, we get that* $\Phi_{\lambda,\mu}(\bar{\mathbf{X}}, \bar{\mathbf{E}}) = \Phi^*_{\lambda,\mu}$. *Considering that* $(\bar{\mathbf{X}}, \bar{\mathbf{E}}) \in \Gamma$ *is arbitrary, it is not hard to conclude that* $\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)$ *is constant on* Γ.

Exploiting the aforementioned fact and assuming that $\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E})$ *is a KŁ function, we can apply the uniformized KŁ property. As a result, there exist* $\varepsilon > 0$, $\eta > 0$, *and* $\Theta \in \Xi_\eta$ *such that for all* $(\mathbf{X}, \mathbf{E})$ *satisfying* $dist((\mathbf{X}, \mathbf{E}), \Gamma) < \varepsilon$ *and* $\Phi^*_{\lambda,\mu} < \Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E}) < \Phi^*_{\lambda,\mu} + \eta$, *then we can achieve*

$$\Theta'(\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E}) - \Phi^*_{\lambda,\mu}) dist(\boldsymbol{0}, \partial\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E})) \geq 1, \tag{59}$$

where the ***Definition*** *3 is used for (59). Moreover, due to the fact that* $\lim_{k\to+\infty} dist((\mathbf{X}_k, \mathbf{E}_k), \Gamma) = 0$ *according to the definition of* Γ, *and* $\lim_{k\to\infty} \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) = \Phi^*_{\lambda,\mu}$, *for such* $\varepsilon > 0$ *and* $\eta > 0$, *there must exist a* $\bar{k}$ *such that* $dist((\mathbf{X}_k, \mathbf{E}_k), \Gamma) < \varepsilon$ *and*

$$\Phi^*_{\lambda,\mu} < \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) < \Phi^*_{\lambda,\mu} + \eta \tag{60}$$

hold for all $k > \bar{k}$. *As a result, by setting* $(\mathbf{X}_k, \mathbf{E}_k)$ *to* $(\mathbf{X}, \mathbf{E})$, *we naturally satisfy (59), i.e.,*

$$\begin{aligned} &\Theta'(\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) - \Phi^*_{\lambda,\mu}) dist(\boldsymbol{0}, \partial\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)) \\ &\geq 1. \end{aligned} \tag{61}$$

This can be easily achieved from both ***Proposition*** *3 and* ***Definition*** *3.*

Step 2*: We aim to prove the boundedness for the distance from* $\boldsymbol{0}$ *to* $\partial\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E})$. *To achieve this, we consider the subdifferential with respect to* $\mathbf{X}$ *and* $\mathbf{E}$, *accordingly. By computing the corresponding expressions, it is easy from Problem (1) to obtain:*

$$\begin{aligned} &\partial_{\mathbf{X}}\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E}_k)|_{\mathbf{X}=\mathbf{X}_k} \\ &= \partial h_\lambda(\mathbf{X}_k) + \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_{k-1}) \\ &\quad + \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_k) - \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_{k-1}) \\ &\ni -\phi_{\mathbf{X}}(\mathbf{X}_k - \mathbf{X}_{k-1}) \\ &\quad + \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_k) - \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_{k-1}); \end{aligned} \tag{62}$$

$$\begin{aligned} &\partial_{\mathbf{E}}\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E})|_{\mathbf{E}=\mathbf{E}_k} \\ &= \partial f(\mathbf{E}_k) + \nabla_{\mathbf{E}} g_\mu(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_k) \\ &\ni -\phi_{\mathbf{E}}(\mathbf{E}_k - \mathbf{E}_{k-1}), \end{aligned} \tag{63}$$

where both (62) and (63) hold derived from the above-given (40) and (41), accordingly.

By using the decent Lemma 1 of [56] for the differentiable property of $g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E})$, *for instance,* $\frac{\mu}{2}\|\mathbf{X} + \mathbf{E} - \mathbf{D}\|_F^2$ *or* $\frac{\mu}{2}\|\mathbf{A}\mathbf{X} + \mathbf{E} - \mathbf{D}\|_F^2$, *with respect to* $\mathbf{E}$, *there exist a constant* $\bar{C}_\mu > 0$ *such that*

$$\begin{aligned} &\|\nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_k) - \nabla_{\mathbf{X}} g_\mu(\mathbf{D}; \mathbf{X}_k, \mathbf{E}_{k-1})\|_F \\ &\leq \bar{C}_\mu \|\mathbf{E}_k - \mathbf{E}_{k-1}\|_F. \end{aligned} \tag{64}$$

Thus, from the aforementioned relations in (62)-(64), there exists $C_{\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}} > 0$ *so that*

$$\begin{aligned} &dist(\boldsymbol{0}, \partial\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k)) \\ &\leq C_{\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}(\|\mathbf{X}_k - \mathbf{X}_{k-1}\|_F + \|\mathbf{E}_k - \mathbf{E}_{k-1}\|_F), \end{aligned} \tag{65}$$

where $C_{\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}$ *is related to the parameters* μ, $\phi_{\mathbf{X}}$, *and* $\phi_{\mathbf{E}}$, *and this inequality validates the desired boundedness of this distance.*

Step 3*: We, for notational simplicity, first define*

$$\begin{aligned} \triangle_{k,k+1} = &\Theta(\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) - \Phi^*_{\lambda,\mu}) \\ &- \Theta(\Phi_{\lambda,\mu}(\mathbf{X}_{k+1}, \mathbf{E}_{k+1}) - \Phi^*_{\lambda,\mu}), \end{aligned} \tag{66}$$

where since $\Phi_{\lambda,\mu}(\cdot,\cdot)$ is non-increasing and $\Theta(\cdot)$ is monotonic, it is easy to have $\triangle_{k,k+1} \geq 0$ for all $k \geq 1$. Then, we have for all $k \geq \bar{k}$ such that

$$\begin{aligned}
&C_{\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}(\|\mathbf{X}_k - \mathbf{X}_{k-1}\|_F + \|\mathbf{E}_k - \mathbf{E}_{k-1}\|_F)\triangle_{k,k+1} \\
&\geq dist(\boldsymbol{0}, \partial\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k))\triangle_{k,k+1} \\
&\geq dist(\boldsymbol{0}, \partial\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k))\Theta'(\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) - \Phi^*_{\lambda,\mu}) \\
&\quad \times (\Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) - \Phi_{\lambda,\mu}(\mathbf{X}_{k+1}, \mathbf{E}_{k+1})) \\
&\geq \Phi_{\lambda,\mu}(\mathbf{X}_k, \mathbf{E}_k) - \Phi_{\lambda,\mu}(\mathbf{X}_{k+1}, \mathbf{E}_{k+1}) \\
&\geq \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2 + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2 \\
&\geq \gamma\left(\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F + \|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F\right)^2, \qquad (67)
\end{aligned}$$

where $\gamma = \frac{min(\phi_{\mathbf{X}},\phi_{\mathbf{E}})}{4}$, the last inequality holds from the basic inequality, i.e., $(a+b)^2 \leq 2(a^2+b^2)$ for a, $b \geq 0$, and the forth inequality holds from (34, directly. The first to third inequalities hold from (59), (61), (65), (66), and the concavity of Θ.

Taking the square root of (67) and using the basic inequality, i.e., $2\sqrt{ab} \leq a+b$ for a, $b \geq 0$, we obtain for $\beta = \frac{C_{\mu,\phi_{\mathbf{X}},\phi_{\mathbf{E}}}}{\gamma}$ and rearrange the objective terms that

$$\begin{aligned}
&2\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F + 2\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F \\
&\leq (\|\mathbf{X}_k - \mathbf{X}_{k-1}\|_F + \|\mathbf{E}_k - \mathbf{E}_{k-1}\|_F) + \beta\triangle_{k,k+1}, \quad (68)
\end{aligned}$$

where let us prove that for any $k > l$, the subsequent inequality can be achieved by summing up both sides of (68) for $i = l+1, ..., k$, represented by

$$\begin{aligned}
&2\sum_{i=l+1}^{k}\|\mathbf{X}_{i+1} - \mathbf{X}_i\|_F + 2\sum_{i=l+1}^{k}\|\mathbf{E}_{i+1} - \mathbf{E}_i\|_F \\
&\leq \sum_{i=l+1}^{k}(\|\mathbf{X}_i - \mathbf{X}_{i-1}\|_F + \|\mathbf{E}_k - \mathbf{E}_{k-1}\|_F) + \beta\sum_{i=l+1}^{k}\triangle_{i,i+1} \\
&\leq \sum_{i=l+1}^{k}(\|\mathbf{X}_{i+1} - \mathbf{X}_i\|_F + \|\mathbf{E}_{i+1} - \mathbf{E}_i\|_F) + \beta\triangle_{l+1,k+1} \\
&\quad + \|\mathbf{X}_{l+1} - \mathbf{X}_l\|_F + \|\mathbf{E}_{l+1} - \mathbf{E}_l\|_F \\
&\quad - \|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F + \|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F, \qquad (69)
\end{aligned}$$

where the last inequality follows from the relations for computations of sum, and the fact that $\triangle_{s,v} = \triangle_{s,t} + \triangle_{t,v}$, for all the integers s, t, and r. Using $\Theta \geq 0$ and $\|\cdot\|_F \geq 0$, we thus have for any $k > l$ that (69) shows

$$\begin{aligned}
&\sum_{i=l+1}^{k}\|\mathbf{X}_{i+1} - \mathbf{X}_i\|_F + \sum_{i=l+1}^{k}\|\mathbf{E}_{i+1} - \mathbf{E}_i\|_F \\
&\leq \beta\Theta(\Phi_{\lambda,\mu}(\mathbf{X}_{l+1}, \mathbf{E}_{l+1}) - \Phi^*_{\lambda,\mu}) \\
&\quad + \|\mathbf{X}_{l+1} - \mathbf{X}_l\|_F + \|\mathbf{E}_{l+1} - \mathbf{E}_l\|_F. \qquad (70)
\end{aligned}$$

This inequality implies that the sequence $\{(\mathbf{X}_k, \mathbf{E}_k)\}$ has finite length as $j \to +\infty$, that is to say,

$$\sum_{i=1}^{+\infty}\|\mathbf{X}_{i+1} - \mathbf{X}_i\|_F + \sum_{i=1}^{+\infty}\|\mathbf{E}_{i+1} - \mathbf{E}_i\|_F < +\infty. \qquad (71)$$

Importantly, for $v > t > s$, we have

$$\begin{aligned}
&\|\mathbf{X}_v - \mathbf{X}_s\|_F + \|\mathbf{E}_v - \mathbf{E}_s\|_F \\
&= \|\sum_{i=s}^{v-1}(\mathbf{X}_{i+1} - \mathbf{X}_i)\|_F + \|\sum_{i=s}^{v-1}(\mathbf{E}_{i+1} - \mathbf{E}_i)\|_F \\
&\leq \sum_{i=s}^{v-1}\|\mathbf{X}_{i+1} - \mathbf{X}_i\|_F + \sum_{i=s}^{v-1}\|\mathbf{E}_{i+1} - \mathbf{E}_i\|_F. \qquad (72)
\end{aligned}$$

Then, combining (72) with (71) implies that $\{\mathbf{X}_k\}$ is a Cauchy sequence due to the fact that $\sum_{i=s+1}^{+\infty}\|\mathbf{X}_{i+1} - \mathbf{X}_i\|_F$ converges to 0 as $s \to +\infty$. Similarly, we conclude that $\{\mathbf{E}_k\}$ is also a Cauchy sequence. Thus, it is clear that $\{\mathbf{X}_k, \mathbf{E}_k\}$ is a convergent sequence. □

Remark 3 *The proofs of **Theorems** 1–3 are valid for nonconvex PBCD, where the solutions of (14) and (15) are usually analytic. While the usage of RSVD and linearization techniques to obtain low-rank solutions is not explicitly mentioned, it is necessary to modify the inequalities (16) and (17) to account for the specific cases considered in (21) and **Proposition** 2.*

*Similar processing techniques for the proofs have been provided in previous works, including [45, 47, 74]. However, these works do not elaborate on the application of faster SVD strategies and linearization techniques to achieve low-rank solutions with proximal operators and thus motivate us give the modifications required in the above proof steps. By incorporating these techniques, the proofs can be adapted modified to show the theoretical analysis to **Theorems** 1–3.*

Theorem 4 *Consider a variable sequence $\{(\mathbf{X}_k, \mathbf{E}_k)\}$ generated by **Algorithm** 1 for Problem (10). Under the conditions $\phi_{\mathbf{X}} > 0$ and $\phi_{\mathbf{E}} > 0$ with the continuation strategy, i.e., $\mu_k = \min(\alpha\mu_{k-1}, \mu_{max})$ for $\alpha > 1$, we can establish the following inequality for the objective function sequence $\{\Phi_{\lambda,\mu_k}(\mathbf{X}_k, \mathbf{E}_k)\}$ of Problem (1):*

$$\begin{aligned}
&\Phi_{\lambda,\mu_{k+1}}(\mathbf{X}_{k+1}, \mathbf{E}_{k+1}) - \alpha\Phi_{\lambda,\mu_k}(\mathbf{X}_k, \mathbf{E}_k) \\
&\leq -\frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2 - \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2. \qquad (73)
\end{aligned}$$

Proof *By considering the definition of $\Phi_{\lambda,\mu}(\mathbf{X}, \mathbf{E})$ in Problem (1) and summing both sides of (16) and*

(17) with $\mathbf{X} = \mathbf{X}_k$ *and* $\mathbf{E} = \mathbf{E}_k$*, respectively, and then setting* $\mu = \mu_{k+1}$*, we obtain the expression:*

$$\Phi_{\lambda,\mu_{k+1}}(\mathbf{X}_{k+1}, \mathbf{E}_{k+1}) + \frac{\phi_{\mathbf{X}}}{2}\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2 + \frac{\phi_{\mathbf{E}}}{2}\|\mathbf{E}_{k+1} - \mathbf{E}_k\|_F^2 \leq \Phi_{\lambda,\mu_{k+1}}(\mathbf{X}_k, \mathbf{E}_k). \tag{74}$$

From the representation of $\Phi_{\lambda,\mu_{k+1}}(\mathbf{X}_k, \mathbf{E}_k)$ *defined in Problem (1), together with the non-negative property of the involved terms, we conclude from* $\mu_{k+1} \leq \alpha\mu_k$ *that for all* $\alpha > 1$*, it is easy to get*

$$\Phi_{\lambda,\mu_{k+1}}(\mathbf{X}_k, \mathbf{E}_k) \leq \alpha\Phi_{\lambda,\mu_k}(\mathbf{X}_k, \mathbf{E}_k). \tag{75}$$

This inequality holds easily. Thus, a direct consequence of (73) is achieved from both (74) and (75). □

Remark 4 *The inequality (73) indeed provides a generalized extension that encompasses various choices of* α*. When* $\alpha = 1$*, it corresponds to the result obtained in* ***(i)*** *of* ***Theorems*** *1 without using the continuation strategy. As the value of* α *increases, the difference on the left-hand side of the inequality for " ≤ " becomes smaller, indicating faster convergence.*

Followed by ***Theorem*** *4, we can naturally provide theoretical analysis and guarantees for local and global convergence by incorporating the revisions to* ***Theorems*** *1–3 for nonconvex PBCD with the continuation strategy in a similar way. Therefore, for the sake of brevity and space constraints, we have omitted the detailed proofs and revisions in this content.*

4 Extensions of Nonconvex PBCD

In this section, our main focus is to extend and study the proposed nonconvex PBCD algorithm, allowing us to handle problems with multiple variables. This extension enables efficient updates of each variable within a cycle of iteration procedures. The extended formulation of the multi-variable minimization problem can be represented as follows:

$$\begin{aligned}&\min_{(\{\mathbf{X}_i\}),(\{\hat{\mathbf{X}}_i\})} \Phi_{(\{\lambda_i\}),\mu,(\{\phi_i\})}(\{\mathbf{X}_i\}, \{\hat{\mathbf{X}}_i\})\\ &= \sum_{i=1}^{I} \rho_{\lambda_i}(\mathbf{X}_i) + g_\mu(\mathbf{D}; (\{\mathbf{X}_i\}))\\ &\quad + \sum_{i=1}^{I} \frac{\phi_i}{2}\|\mathbf{X}_i - \hat{\mathbf{X}}_i\|_F^2,\end{aligned} \tag{76}$$

Here, I represents the total number of variables in $(\{\mathbf{X}_i\})$, which denotes the set of multiple variables $(\mathbf{X}_1, \ldots, \mathbf{X}_I)$. $(\{\lambda_i\})$ represents the sequences of regularization parameters $(\lambda_1, \lambda_2, \ldots, \lambda_I)$, while both $(\{\hat{\mathbf{X}}_i\})$ and $(\{\phi_i\})$ have their corresponding definitions as given above.

Algorithm 4 Optimization for Problem (76)

Input: $\mathbf{D}$, $\mathcal{X}^k$, $\lambda_i > 0$, $\mu > 0$, $1 \leq i \leq I$, $\eta_{\mathbf{X}_i} > l_{\mu,\mathbf{X}_i}$, $\rho > 1$, $k = 0$.
Output: $\mathcal{X}^* \leftarrow \mathcal{X}^{k+1}$.

While not converged do
  for $i = 1 : I$
    Update $\mathbf{X}_i^{k+1}$ by solving (82) with two specific cases as shown in (83) and (84), accordingly.
  end
until convergence

It is evident that Problem (1) can be considered as a special case of Problem (76) with $I = 2$ and $\phi_1 = \phi_2 = 0$, by setting $\rho_{\lambda_1}(\mathbf{X}_1) = h_\lambda(\mathbf{X})$ with $\lambda_1 = \lambda$, $\rho_{\lambda_2}(\mathbf{X}_2) = f(\mathbf{E})$ with $\lambda_2 = 1$, and $g_\mu(\mathbf{D}; \mathbf{X}_1, \mathbf{X}_2) = g_\mu(\mathbf{D}; \mathbf{X}, \mathbf{E})$. With these settings and by requiring $\hat{\mathbf{X}}_1 = \hat{\mathbf{X}}$, $\hat{\mathbf{X}}_2 = \hat{\mathbf{E}}$, $\phi_1 = \phi_{\mathbf{X}}$, and $\phi_2 = \phi_{\mathbf{E}}$, Problem (10) can also be viewed as a specific instance of Problem (76). Hence, each $\rho_i(\mathbf{X}_i)$ can be utilized to quantify the low-rank matrix or residual matrix with the proper choices and explanations. Then, we assume that $g_\mu(\mathbf{D}; (\{\mathbf{X}_i\}))$ possesses the property defined below.

Definition 4 *The function* $g_\mu(\mathbf{D}; \{\mathbf{X}_i\})$ *is a* C^1 *function with respect to each involved variable* $\mathbf{X}_i$*, and its gradient* $\nabla g_\mu(\mathbf{D}; \{\mathbf{X}_i\})$ *is Lipschitz continuous when all* $\mathbf{X}_j$*'s are fixed for* $j \neq i$ $(1 \leq j \leq I)$*. Specifically, there exist the Lipschitz constants* $l_{\mu,\mathbf{X}_i} \geq 0$ *such that*

$$\begin{aligned}&\|\nabla g_\mu(\mathbf{D}; \{\mathbf{X}_i\}) - \nabla g_\mu(\mathbf{D}; \{\widetilde{\mathbf{X}}_i\})\|_F\\ &\leq \sum_{i=1}^{I} l_{\mu,\mathbf{X}_i}\|\mathbf{X}_i - \widetilde{\mathbf{X}}_i\|_F,\end{aligned} \tag{77}$$

which plays a key role in the theoretical convergence analysis, as exemplified in (64) for this extension.

This definition highlights the nontrivial extension of the proposed algorithms to handle Problem (76). For the sake of notational simplicity, we introduce the following definitions:

$$\begin{cases} \mathcal{X}^k = (\mathbf{X}_1^k, \mathbf{X}_2^k, \ldots\ldots, \mathbf{X}_I^k); & (78)\\ \mathcal{X}_0^k(i) = (\mathbf{X}_1^{k+1}, \ldots\ldots, \mathbf{X}_{i-1}^{k+1}, \mathbf{X}_i, \mathbf{X}_{i+1}^k \ldots\ldots, \mathbf{X}_I^k); & (79)\\ \mathcal{X}_0^k(i_+) = (\mathbf{X}_1^{k+1}, \ldots\ldots, \mathbf{X}_{i-1}^{k+1}, \mathbf{X}_i^k, \mathbf{X}_{i+1}^k \ldots\ldots, \mathbf{X}_I^k); & (80)\\ \mathcal{X}_0^k(i_-) = (\mathbf{X}_1^{k+1}, \ldots\ldots, \mathbf{X}_{i-1}^{k+1}, \mathbf{X}_{i+1}^k \ldots\ldots, \mathbf{X}_I^k); & (81)\end{cases}$$

From (78)-(81), we can outline the optimization procedure of the algorithms for multiple separable variables. Starting with any initial point $\mathcal{X}^0$, the multi-variable solution can generate the sequence value at the $(k+1)$-step, denoted by $\mathcal{X}^{k+1}$ in (78), via the successively updating rule for each involved variable for all

$1 \leq i \leq I$, as follows:

$$\mathbf{X}_i^{k+1} \in \operatorname{argmin}_{\mathbf{X}_i} \rho_{\lambda_i}(\mathbf{X}_i) + g_\mu(\mathbf{D}; \mathcal{X}_0^k(i)) + \frac{\phi_i}{2}\|\mathbf{X}_i - \mathbf{X}_i^k\|_F^2, \tag{82}$$

where the closed-form solution relies on various $\rho_{\lambda_i}(\mathbf{X}_i)$ and the specific structure of $g_\mu(\mathbf{D}; \mathcal{X}_0^k(i))$, as discussed in the formulations (18), (21), and (23), respectively.

In the subsequent analysis, we focus on considering two cases of the subproblem (82) aiming to solve Problem (76), effectively. These cases can be regarded as extensions of (18), (21), and (23) in the optimization of Problem (10). Subsequently, we provide the detailed revisions for the updated variables:

- First, we extend (18) and (23) to the following equation, as follows

$$\mathbf{X}_i^{k+1} = \operatorname{argmin}_{\mathbf{X}_i} \frac{1}{\phi_i + \mu}\rho_{\lambda_i}(\mathbf{X}_i) + \frac{1}{2}\left\|\mathbf{X}_i - \hat{g}_{\mu,\phi_i}(\mathbf{D}; \mathcal{X}_0^k(i))\right\|_F^2, \tag{83}$$

  This minimization problem is achieved by setting $\hat{g}_{\mu,\phi_i}(\mathbf{D}; \mathcal{X}_0^k(i_+)) = \frac{1}{\phi_i+\mu}[\phi_i \mathbf{X}_i^k + \mu(\mathbf{D} - \mathcal{A}(\mathcal{X}_0^k(i_-)))]$, where $\mathcal{A}(\mathcal{X}_0^k(i_-))$ is represented by the linearization combination of the variables in (81) that are unrelated to the desired $\mathbf{X}_i$.
- Next, we extend (21) to the following equation:

$$\mathbf{X}_i^{k+1} = \operatorname{argmin}_{\mathbf{X}_i} \frac{1}{\phi_i + l_{\mu,\mathbf{X}_i}}\rho_{\lambda_i}(\mathbf{X}_i) + \frac{1}{2}\left\|\mathbf{X}_i - \bar{g}_{\mu,l_{\mu,\mathbf{X}_i}}(\mathbf{D}; \mathcal{X}_0^k(i))\right\|_F^2, \tag{84}$$

  Here, the minimization problem is achieved by setting $\bar{g}_{\mu,l_{\mu,\mathbf{X}_i}}(\mathbf{D}; \mathcal{X}_0^k(i_+)) = \frac{1}{\phi_i+l_{\mu,\mathbf{X}_i}}[\phi_i \mathbf{X}_i^k + l_{\mu,\mathbf{X}_i}(\mathbf{D} - \mathcal{A}(\mathcal{X}_0^k(i_-)))]$. The value of $l_{\mu,\mathbf{X}_i}$ is obtained using (77). It should be special noted that the resulting solutions of (83) and (84) can be easily obtained using the generalized proximal operators described in **Proposition** 2 or **Proposition** 1.

Remark 5 *The proposed nonconvex PBCD algorithm for solving Problems (10) and (76) differs from existing methods such as PALM [56] and SPJIM [47] in several aspects. While PALM and SPJIM update the variables in an alternating order or jointly optimize them, the PBCD algorithm minimizes the unconstrained problems with respect to each variable belonging to $\{\mathbf{X}_i\}$ independently in a cyclic iteration. Then, combined this strategy with the employment of the RSVD and continuity strategies, enhances the efficiency of the proposed algorithms. The closed-form solutions for each variable contribute to its computational simplicity in solving low-rank matrix learning problems. These advantages of the proposed algorithms are thus further highlighted.*

5 Numerical Experiments

In this section, we present the experimental validation of our proposed methods on a diverse set of original data to evaluate their superiority in three image low-level and high-level vision tasks. To establish a comprehensive comparison, we benchmark our methods against several existing algorithms, which are referenced using concise abbreviated notations commonly used in the literature. The databases used in our experiments comprise a collection of raw images, each accompanied by a brief description, as outlined below:

- For the Image Inpainting task, our primary objective is to apply the revised RMC (ReRMC) approach to color images that consist of three channels: R (red), G (green), and B (blue), with each channel being treatable as a matrix. We capture a test image of dimensions 600×650 using a mobile phone, which serves as the non-strictly low-rank reference, representing the original image. To investigate the strictly low-rank case, we perform truncation of the singular values of the original image and retain only the top 10% largest values. Subsequently, we introduce various missing pixel patterns, such as random (30% ratio), text, and curves, to simulate different inpainting scenarios. In this process, we treat each channel of the image as an independent low-rank matrix and further perform inpainting on each channel separately.
- For the Image Classification task, our primary focus is on utilizing the revised RMR (ReRMR) approach for both face images. To conduct our experiments, we utilize the following datasets:
  - AR[2]: This database comprises 126 face images of subjects, each with a size of 60×43, showcasing various facial expressions, illuminations, and occlusions. For our experiments, we utilize the images of the first 100 individuals for training, with 7 samples per subject, and testing, with 6 samples per subject occluded by sunglasses. This leads to the creation of data matrices of size 2580×700 and 2580×600, respectively.
  - ExtYaleB[3]: This database consists of 2,414 frontal-face images distributed among 38 subjects. Each subject is represented by 64 images captured under varying illuminations. All experimental images are resized to a uniform size of 96×84 pixels. For the training data, we utilize the first 7 images for each subject, resulting in a size of 8064×266. For the testing data of each subject, we employ

[2] https://www2.ece.ohio-state.edu/ aleix/ARdatabase.html
[3] https://paperswithcode.com/dataset/extended-yale-b-1

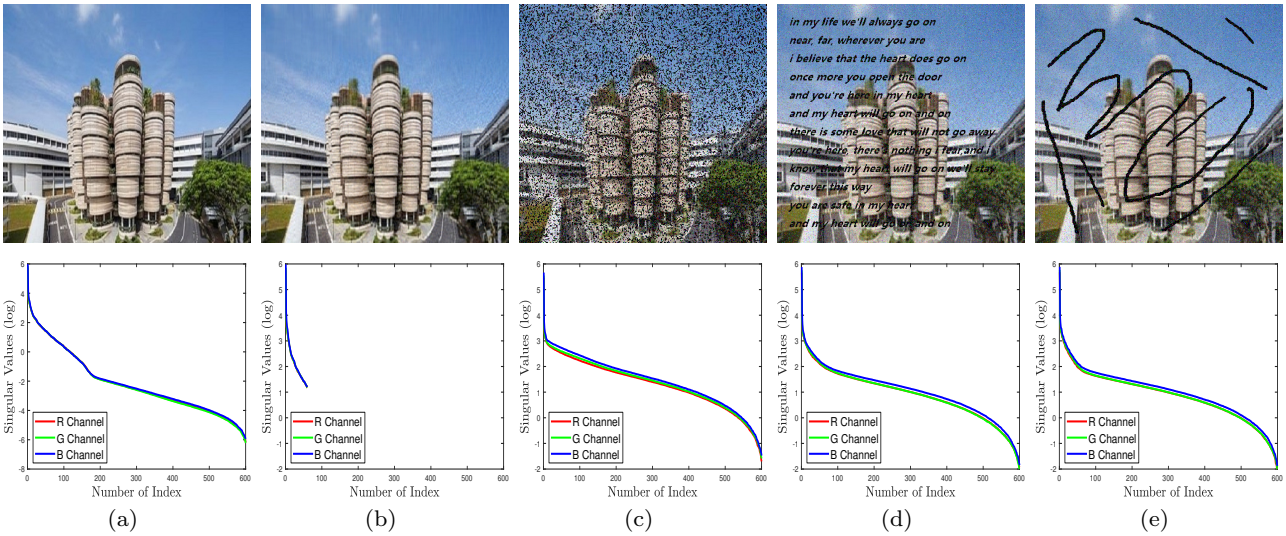


Fig. 4 Illustration of raw-image based photography, showcasing: (a) the original image, (b) the processed image, and (c)-(e) three different sampling masks with 10% of pixels corrupted with sparse noise on the top row. The corresponding sorted singular values for each channel are also plotted in curves on the bottom row.

19 images with different illuminations, leading to a size of 8064×722, and introduce "banbon" occlusions to an additional 12 images, resulting in a size of 8064×456.

- For the Image Clustering task, our main focus is on utilizing the revised LRR (ReLRR) approach for face, digital, and object images. To validate the effectiveness and efficiency of our proposed method, we consider the following datasets:
  - GT [4]: This dataset comprises 750 grayscale face images of 50 subjects, with each subject having 15 images of size 40×30. The images show some variations, resulting in a clustering data matrix with a size of 1200×750.
  - USPS[5]: This is a handwritten digit database containing 9,298 images. For clustering, we selected the first 100 images of each digit. Each image in USPS has a resolution of 16×16 pixels, resulting in a 256-dimensional vector. Thus, a clustering data matrix of size 256×1000 was constructed.
  - FLAVIA[6]: This dataset is a leaf recognition system comprising 1907 images from 32 different plant species. The data matrix size is 1907×1200 pixels, with each image resized to 30×40 pixels. For clustering, we selected the first 20 classes with 40 samples per subject, resulting in a clustering data matrix with a size of 1200×800.
  - COIL20[7]: This dataset includes 1,440 images of 20 objects. Each object provides 72 images captured from various angles. The original images are resized to grayscale images of 32×32 pixels. For clustering, we select the first 30 images per object, resulting in a clustering data matrix with a size of 1024×600.

In the following subsections, we first provide partial images of the databases used in our experiments as a prelude to our evaluations. The experiments were conducted using MATLAB R2021b on a 64-bit PC with an Intel(R) Core(TM) i7-7700 CPU @ 3.6GHz and 8.0GB RAM. To implement the relevant methods, we utilized their released codes and carefully fine-tuned the parameters while keeping the given termination conditions fixed with the same stopping thresholding value to ensure the best possible results. The quantitative evaluations involve the use of numerical metrics to assess the performance of the methods, encompassing both traditional and recently published works. By combining these evaluations, our goal is to offer a comprehensive and insightful analysis of the performance of the image processing and analysis methods optimized using **Algorithm** 3 with double acceleration strategies. Note that the underlying reason stems from the theoretical analysis of computational complexity, which is essential for efficiently learning low-rank matrices. This foundational insight allows us to conduct a more insightful analysis of the acceleration performance of RSVD in comparison to full SVD. The proposed meth-

[4] https://academictorrents.com/details/0848b2c9b40e49041eff85ac4a2da71ae13a3e4f

[5] https://paperswithcode.com/dataset/usps

[6] https://www.researchgate.net/figure/Leaves-in-Flavia-dataset_fig7_224954031

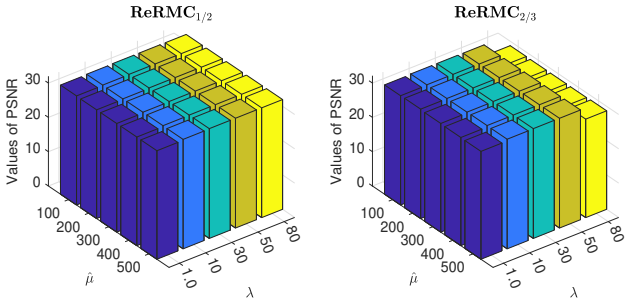
[7] https://git-disl.github.io/GTDLBench/datasets/coil20/

Table 2 Evaluation metrics for various methods applied to color image inpainting using three sampling masks and corruption noises.

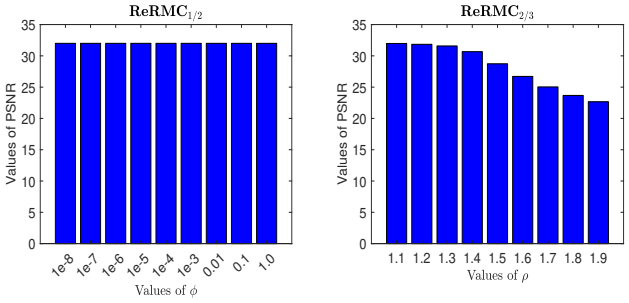
//	Random		Text		Curve	
	PSNR	TIME	PSNR	TIME	PSNR	TIME
APG [16]	21.780	8s	22.349	8s	22.384	9s
NNM [18]	18.611	36s	18.430	39s	18.372	31s
IRNN [53]	22.566	14s	22.944	15s	22.935	14s
GPG [17]	22.675	13s	23.111	14s	23.084	14s
FNM [40]	31.469	3s	31.600	6s	31.540	5s
BiNM [40]	31.700	4s	31.770	4s	31.742	5s
UNBF [50]	28.554	10s	28.545	9s	28.438	9s
PSVT [51]	22.773	12s	23.375	17s	23.355	21s
WNNM [52]	16.703	111s	17.585	130s	17.564	131s
DNNR [59]	22.433	11s	23.123	12s	23.025	12s
FaNCL [74]	21.825	14s	23.023	29s	23.079	29s
HQASD [75]	23.728	4s	24.589	4s	24.288	6s
HOAT [76]	31.497	336s	31.523	372s	31.685	380s
HOMT [76]	**32.000**	135s	**31.999**	128s	**31.995**	306s
ReRMC$_{1/2}$	31.999	28s	31.991	17s	31.940	19s
ReRMC$_{2/3}$	31.996	28s	31.975	17s	31.869	17s

ods, which are revised variants of Problems (2)-(4), are denoted as ReRMCp, ReRMRp, and ReLRR$_p$, where $p \in \{1/2, 2/3\}$. These variants are designed to achieve effectiveness and efficiency in handling the low-rank matrix learning problem across the image low-level and high-level application tasks.

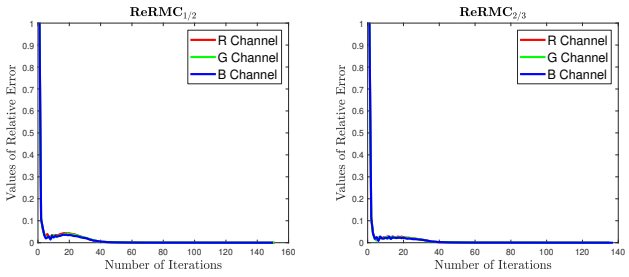
It is crucial to perform a thorough analysis that highlights the differences and advantages of our proposed method in comparison to existing approaches across various application tasks. This endeavor is in line with our commitment to delivering a comprehensive evaluation, enhancing the discussion across diverse dimensions. These aspects primarily include: 1) Real-world applications and databases (as introduced above), 2) Different methods for various vision tasks, 3) Both numerical and visual comparisons, and 4) Results from the ablation analysis. To save space in the subsequent experimental comparisons, we will use symbol names to substitute the full names of the comparison methods among the various application tasks. If interested in these comparison methods, please refer to the corresponding references. Additionally, our proposed methods exhibit superior performance and computational efficiency, attributes attributed to nonconvexity and the application of dual acceleration techniques.



(a) Performance influences of (λ, μ_0) on random mask



(b) Performance Influences of ϕ and ρ on the text mask

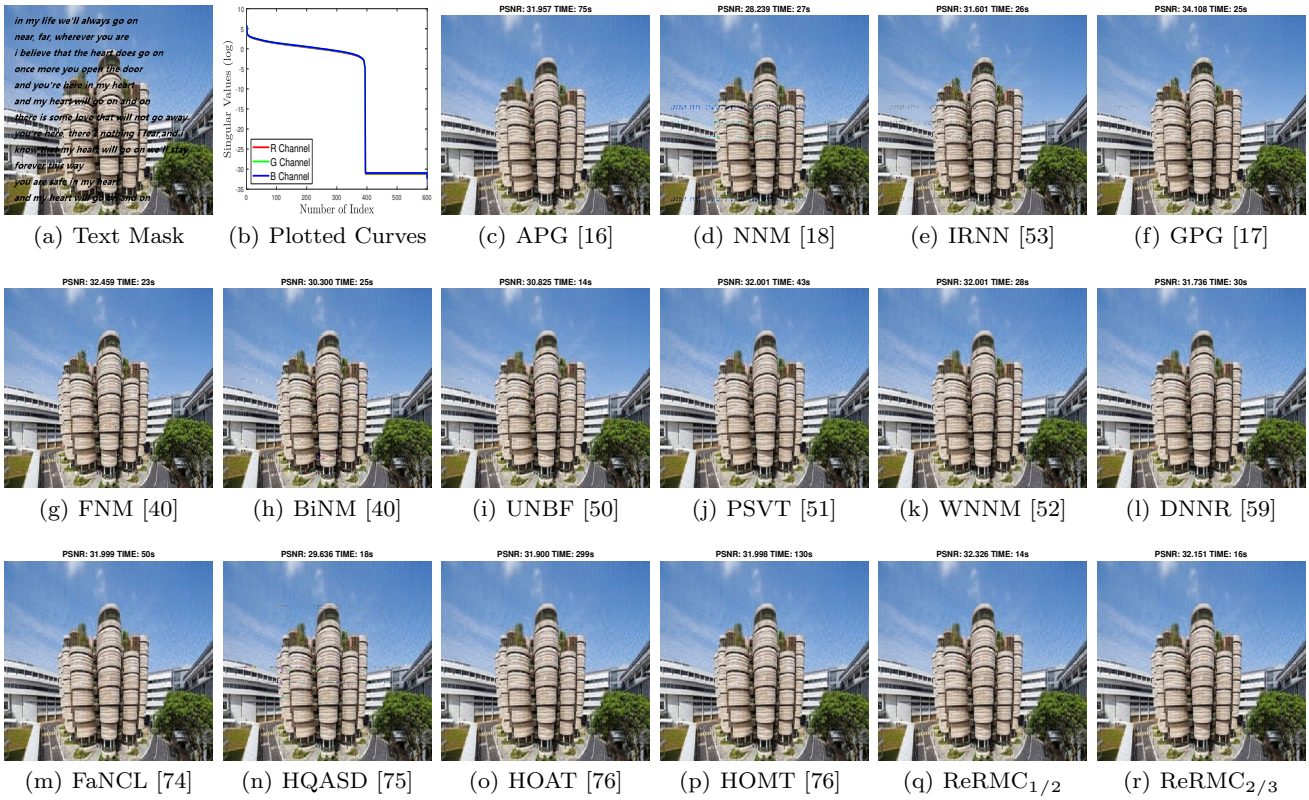


(c) Two plotted convergence curves on the curve mask

Fig. 5 Visual comparisons of the proposed inpainting methods, i.e., ReRMC$_{1/2}$ and ReRMC$_{2/3}$, from different perspectives in various experimental settings, including (a) with random mask, (b) with text mask, and (c) with curve mask.

5.1 Experiments on the Image Inpainting

In this subsection, we present **Fig**. 4, which showcases the original image along with the corresponding strictly low-rank images generated for three different missing masks. The objective of this subsection is to investigate the effectiveness and efficiency of various inpainting methods, as listed in **Table** 2, for completing the missing entries in images (c)-(e) with 10% sparse corruption noises. The comparison includes both convex and nonconvex matrix completion methods solved using first-order optimization algorithms. Notably, APG [16] and NNM [18] achieve lower PSNR values compared to other methods, primarily due to the biased estimator used for the nuclear norm in the rank function. On the other hand, nonconvex inpainting methods such as IRNN [53], GPG [17], DNNR [59], PSVT [51],



(a) Text Mask (b) Plotted Curves (c) APG [16] (d) NNM [18] (e) IRNN [53] (f) GPG [17]

(g) FNM [40] (h) BiNM [40] (i) UNBF [50] (j) PSVT [51] (k) WNNM [52] (l) DNNR [59]

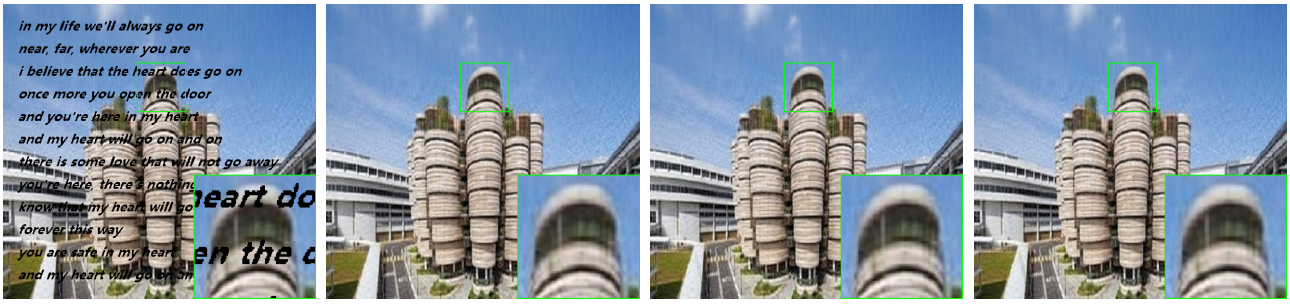
(m) FaNCL [74] (n) HQASD [75] (o) HOAT [76] (p) HOMT [76] (q) ReRMC$_{1/2}$ (r) ReRMC$_{2/3}$

Fig. 6 Illustrations for quantitative and qualitative evaluations on the original inpainting image only with missing entries are shown. Additionally, (a) displays the corresponding image with plotted curves of singular values in (b). The inpainting results obtained by various methods are depicted in (c)-(p), while our proposed two methods are presented in (q) and (r).

HOAT, and HOMT [76], and HQASD [75] exhibit better evaluation performance in most cases. Additionally, the adoption of factorization and continuity strategies helps reduce the computational time in methods such as UNBF [50], FNM and BiNM [40], FaNCL [74], IRNN [53], and GPG [17]. However, some methods such as NNM [18], WNNM [52], HOAT, and HOMT [76] have higher computation times, as reflected in **Table** 2. Notably, our proposed ReRMC method, which combines a nonconvex formulation and two-fold acceleration strategies, demonstrates superiority in terms of higher PSNR values and lower computation time (TIME) in seconds under most cases. These results solidly confirm the advantages of the proposed methods to some degree.

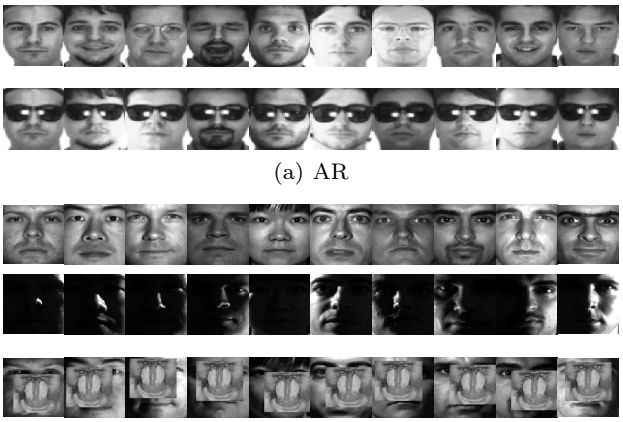
To further analyze both ReRMC$_{1/2}$ and ReRMC$_{2/3}$, we conducted experiments on three different inpainting tasks and examined the impact of various parameters, including random initialization variables and the convergence property of the relative error. The visual comparisons of these analyses are presented in **Fig**. 5 (a)-(c). In detail, we first analyzed the effect of the function parameters, such as λ and $\hat{\mu}$, on the PSNR values for the random mask scenario, as depicted in (a). The PSNR values are provided for different choices of the model parameter ϕ by setting it equal to $\phi_{\mathbf{X}} = \phi_{\mathbf{E}}$ and the algorithm parameter, i.e., step size ρ, for the case of the text mask, as shown in (b). Furthermore, in (c), we plotted the relative error values for the R, G, and B channels in the case of the curve mask. These visual results provide meaningful insights into the distributions of numerical results, the non-increasing property of the relative errors, and the evaluation efficacy and efficiency of the involved parameters for setting $\phi = 1e-5$ and $\rho = 1.1$. This analysis effectively validates the effectiveness of the proposed methods.

In **Fig**. 6 (c)-(r), we present the inpainting performance and recovery images of the compared algorithms, but this time without the corruption of sparse noises, focusing solely on text missing entries. Interestingly, each algorithm achieves significantly higher PSNR values when recovering the inpainting image. This discrepancy can be attributed to the noise measurements used in the objective formulation of the inpainting model. Additionally, with the absence of sparse noise corruption, the inpainting tasks become relatively easier. It can be observed that most of methods perform better in this case. Meanwhile, the involved clustering methods have shown good performance in this task. As a re-



(a) Test Mask (b) Original Image (c) ReRMC$_{1/2}$ (d) ReRMC$_{2/3}$

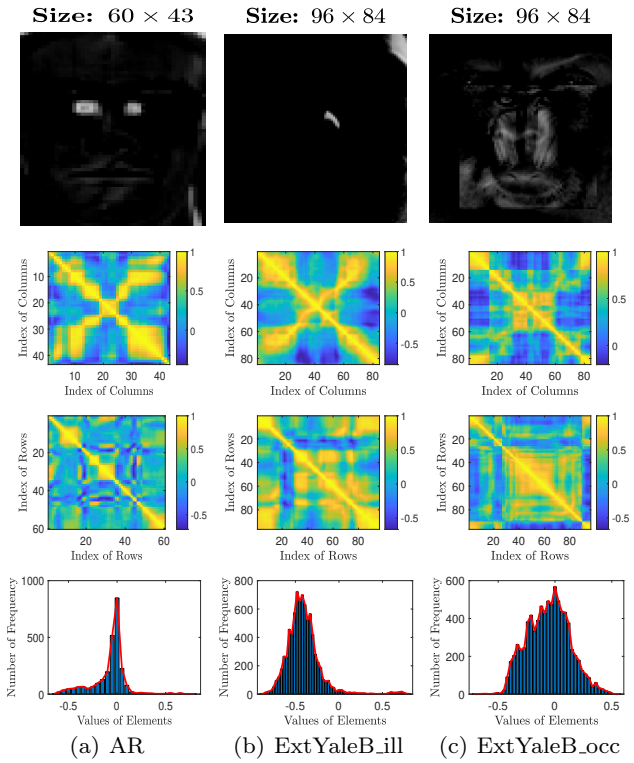
Fig. 7 Incorporating visual comparisons for presenting close-ups of local regions over the original inpainting image, both with and without missing entries, alongside the recovered images using proposed ReRMC$_{1/2}$ and ReRMC$_{2/3}$, respectively.



(a) AR

(b) ExYaleB

Fig. 8 Illustrations of partial training and testing face images selected from two experimental databases.



(a) AR (b) ExtYaleB_ill (c) ExtYaleB_occ

Fig. 9 Visual results of the error images with heatmaps of columns and rows, as well as the histograms, are presented for three different types of corruption noises.

sult, the algorithms showcase higher PSNR values and reduced timing costs under these specific experimental settings with appropriate noise measurements.

To further enhance the clarity of our image inpainting experiments, we have incorporated detailed visual comparisons that focus on close-ups of local regions. In Fig. 7, we meticulously present the visual comparison results, starting with (a) and (b), which showcase the original natural image both with and without missing text entries. Subsequently, (c) and (d) guide you through the visual revisions, accompanied by corresponding descriptions specifically tailored for ReRMC$_{1/2}$ and ReRMC$_{2/3}$, respectively. By delving into these visual comparisons and zooming in on local regions, our main intention is to provide a clear understanding of the proposed method's outcomes, effectively highlighting its inpainting ability in the low-level task of recovering the missing entries.

5.2 Experiments on the Image Classification

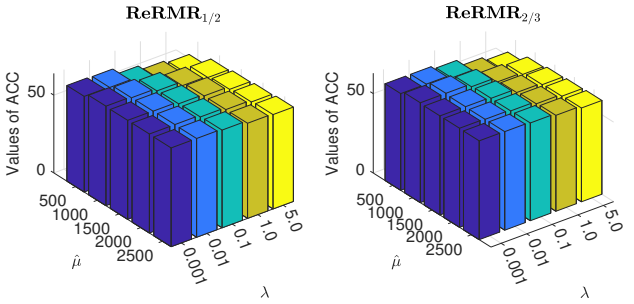
In this subsection, we present partial face and object images in **Fig**. 8 (a)-(c). The training samples appear clear, while the testing ones are intentionally corrupted with occlusions, illuminations, or other types of corruptions. Furthermore, we provide visual analysis from multiple viewpoints for the residual images to enhance the usage of correlations in **Fig**. 9 (a)-(c). Note that the residual images in the first row are obtained by subtracting the first training image from the first testing

Table 3 The evaluation values for various methods applied to image classification using three real-world databases.

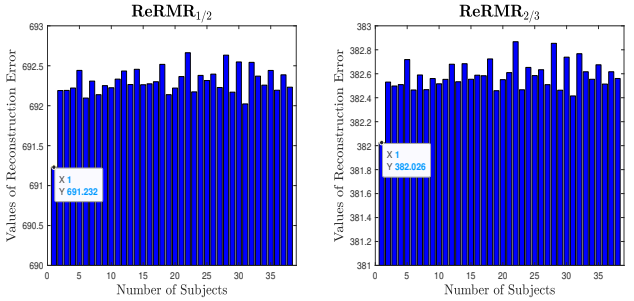
//	**AR**		**ExtYaleB_ill**		**ExtYaleB_occ**	
	ACC	TIME	ACC	TIME	ACC	TIME
SRC [19]	49.500	4s	27.285	5s	53.290	3s
CRC [29]	53.167	2s	49.031	3s	51.097	2s
NMR [22]	73.000	226s	48.338	440s	94.518	298s
FNMR [22]	72.667	158s	47.230	255s	94.518	171s
ALPR [77]	41.667	176s	14.543	623s	28.070	470s
CESR [78]	**78.833**	74s	22.161	74s	39.693	57s
DLSR [79]	40.000	24s	28.033	490s	47.149	525s
ULR_{*} [70]	47.149	7s	49.308	10s	45.614	7s
ULR_{**} [70]	50.833	13s	49.308	12s	47.149	6s
$\text{GIST}_{1/2}$ [21]	50.500	5s	27.285	5s	53.290	4s
$\text{GIST}_{2/3}$ [21]	50.000	5s	27.285	6s	53.509	3s
$\text{ReRMR}_{1/2}$	64.000	23s	**90.028**	50s	**97.807**	27s
$\text{ReRMR}_{2/3}$	62.833	22s	83.657	46s	95.833	25s

image. These visual examples aim to show the importance for the removal of corruptions from the testing images, validating the capability of preserving image fidelity and effectively handling various types of corruptions. Besides this, this visual evidence can also confirm the effectiveness and importance of considering the relations among columns, rows, and elements.

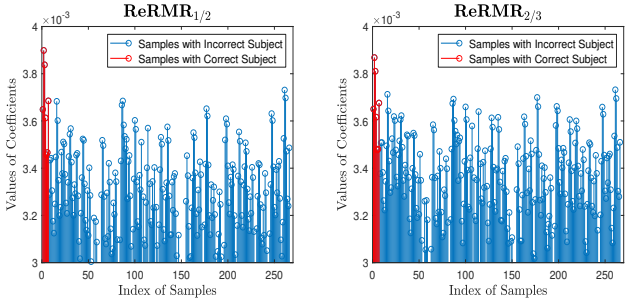
The quantitative results, summarized in **Table** 3, include the proposed methods and ten comparison techniques across three face databases. These methods encompass both convex and nonconvex approaches, as well as vector and matrix-level linear regressions. Due to the presence of occlusions and illuminations in the corrupted images, the residual matrix exhibits low-rank or approximated low-rank structures. It makes sense that NMR and FNMM [22] achieve higher classification accuracy compared to SRC [19], CRC [29], ULR_{*} and ULR_{**} [70], and CESR [78]. The use of unbiased estimators in nonconvex GIST [21] also contributes to its strong performance, surpassing SRC. Incorporating prior information into the regression models, such as ALPR [77], DLSR [79], NMR, and FNMM [22], leads to superior performance as well. Additionally, CRC, ULR, GIST, and FNMR demonstrate higher computational efficiency due to the differentiable nature of the objective function, the specific optimization procedure, and the use of acceleration strategies. Overall, we observe that the proposed methods exhibit advantages in terms



(a) Coefficients vector on the AR database



(b) Reconstruction error on the ExtYaleB_ill database



(c) Coefficients matrix on the ExtYaleB_occ database

Fig. 10 Visual illustrations of (a) the influences of $(\lambda, \hat{\mu})$, (b) reconstruction error values, and (c) coefficient values for the two proposed methods across three databases.

of classification accuracy and computation time over the rank-relaxed regression methods.

In addition to the quantitative results, we further analyze the influences of parameters $(\lambda, \hat{\mu})$, the reconstruction errors, and the distributions of representation coefficients of the proposed methods on the AR, ExtYaleB_ill, and ExtYaleB_occ databases, which contain illumination and occlusion corruptions. In **Fig**. 10 (a), we depict the influences of parameters $(\lambda, \hat{\mu})$ on the performance. The reconstruction errors for the same subjects as the testing samples display lower values, while errors for different subjects exhibit higher values, as shown in **Fig**. 10 (b). Furthermore, the coefficient values for the same subjects exhibit larger magnitudes, while values for different subjects tend to be lower, as illustrated in **Fig**. 10 (c). These visual results

(a) GT (b) USPS (c) FLAVIA (d) COIL20

Fig. 11 Illustrations of partial face, digital, and object images selected from four publicly real-world databases.

Table 4 Comparisons of performance evaluations and timing consumptions for various unsupervised methods applied to image clustering on four real-world databases.

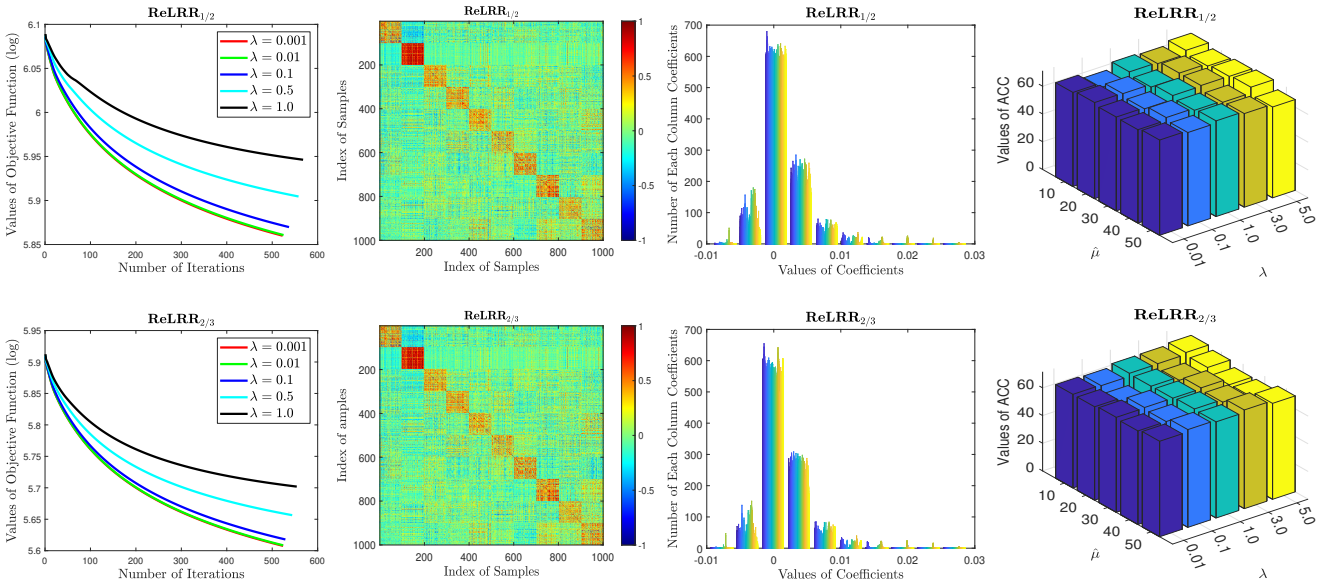
//	**GT**			**USPS**			**FLAVIA**			**COIL20**		
	ACC	NMI	TIME	ACC	NMI	TIME	ACC	NMI	TIME	ACC	NMI	TIME
LRR [11]	47.600	63.939	28s	66.200	62.030	7s	64.750	74.314	48s	68.167	80.139	25s
SSC [26]	44.933	53.996	7s	44.400	39.124	7s	**73.125**	77.221	13s	74.000	80.621	7s
SGL [80]	45.200	62.084	117s	85.700	77.875	207s	69.125	73.743	77s	70.500	80.555	93s
IRLS [68]	46.000	62.445	143s	67.500	66.006	135s	65.125	71.318	110s	66.167	75.861	107s
LSR1 [81]	49.333	64.284	<1s	78.900	66.746	<1s	68.000	71.899	<1s	68.833	77.095	<1s
LSR2 [81]	50.267	64.196	<1s	79.000	66.219	<1s	69.500	72.621	<1s	69.667	78.251	<1s
LRSSC [27]	**53.867**	**66.516**	18s	67.900	67.381	50s	68.625	75.199	25s	81.000	90.472	15s
SSRSC [82]	49.733	67.241	58s	**93.200**	**87.016**	58s	72.875	**78.174**	59s	**92.833**	**97.686**	32s
$\text{FULRR}_{1/2}$ [71]	44.933	59.679	46s	73.900	62.718	9s	63.750	73.204	68s	65.333	78.167	35s
$\text{FULRR}_{2/3}$ [71]	47.200	60.116	41s	74.500	63.746	9s	66.750	71.800	56s	63.667	74.883	29s
$\text{WSLog}_{1/2}$ [83]	48.133	61.662	33s	75.800	61.789	8s	66.375	72.704	50s	62.500	70.869	17s
$\text{WSLog}_{2/3}$ [83]	46.400	62.204	37s	73.000	59.548	8s	66.000	71.840	50s	64.167	75.401	22s
$\text{ReLRR}_{1/2}$	50.800	63.954	63s	78.400	73.983	9s	70.875	75.165	62s	71.167	79.189	43s
$\text{ReLRR}_{2/3}$	52.933	64.675	65s	77.200	71.425	7s	70.875	75.109	57s	70.833	79.504	41s

indicate the coherent representation of subjects in the coefficient matrix, aligning with the inherent grouping of data samples, and ensure the preservation of effective performance under proper parameter choices.

5.3 Experiments on the Image Clustering

In this subsection, we present experimental results for subspace clustering using images obtained from four commonly used databases, as depicted in **Fig**. 11 (a)-(d). We compare our proposed methods with several existing approaches, including standard methods like SSC [26], LSR [81], SGL [80], and LRR [11], as well as their extensions such as IRLS [68], SSRSC [82], and LRSSC [27]. Additionally, we consider SGL and LSR, which involve differentiable clustering model formulations, and WSLog [83] and FULRR [71], which are primarily related to matrix factorization using Schatten-p norm. These diverse approaches yield varying clustering performance and timing consumption. Then, we have some observations and explanations as follows:

It follows from **Table** 4 that the extended methods consistently outperform the standard ones in most cases. Each clustering method, coupled with its corresponding optimization algorithm, shows its specific applicability. These findings underscore the effectiveness and efficiency of our proposed methods, achieved by carefully selecting appropriate parameters for $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0\}$ and $\hat{\mu} \in \{10, 20, 30, 40, 50\}$. The evidence is supported by higher values of ACC and NMI, as well as lower timing costs. While the results show the superiority of the proposed methods compared to other involved methods, including SSC, LRR, IRLS, LSR, WSLog, and FULRR, we acknowledge that the proposed two methods have comparable timing costs and indeed exhibit a degree of advan-



(a) GT (b) USPS (c) FLAVIA (d) COIL20

Fig. 12 Visual comparison results of the proposed methods include (a) convergence curves, (b) block-diagonal structures, (c) coefficient distributions, and (d) parameter sensitivity across four different databases.

tage. Additionally, introducing prior information can be helpful for improving the clustering performance, as verified in SSRSC and LRSSC. This further explain that they have the achieved better performance.

To provide visual comparisons, **Fig**. 12 showcases informative results for our proposed two methods. In (a), we present plotted curves for the values of objective function over the number of iterations, validating their convergence behavior on the GT database. In (b), we depict the block-diagonal structures of the coefficient matrix with normalization, highlighting the relationship with the number of subjects in the USPS database. Additionally, in (c), we display the statistical distribution of the coefficient matrix, offering insights into the clustering samples' representation in the FLAVIA database. Lastly, in (d), we analyze the influences of the parameter choices of $(\lambda, \hat{\mu})$ by conducting experiments on the COIL20 database. These visual results enhance the understanding of the performance and behavior, confirming the feasibility and stability.

5.4 Further Discussion and Analysis

In this subsection, we have chosen specific experimental settings for each application task: inpainting for the curve mask, classification for the ExtYaleB_occ database, and clustering for the COIL20 database. Our primary focus is to test the proposed two methods of interest for $p = 1/2$ and $2/3$ and investigate the impact with and without the RSVD acceleration module. The comparisons are conducted using the following two setups:

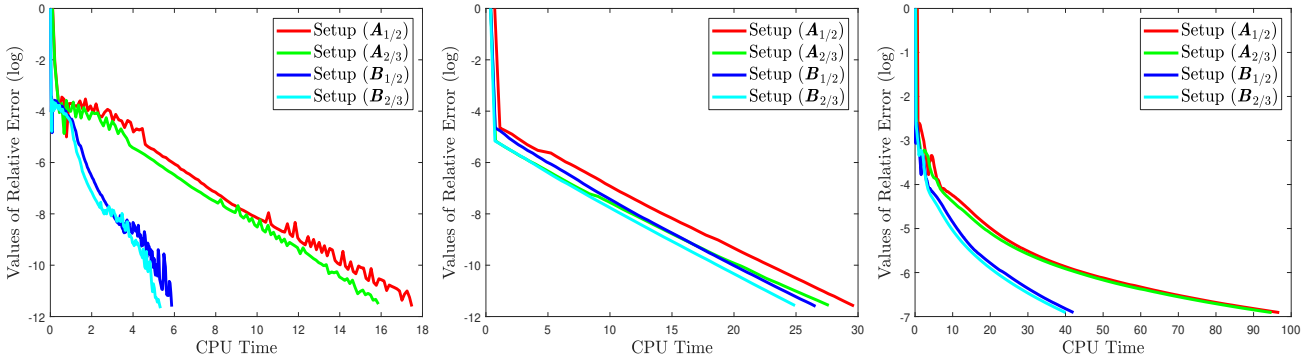
- (***A***) **Algorithm** 1 using full SVD with the continuation strategy;
- (***B***) **Algorithm** 1 using RSVD with the continuation strategy;

To analyze the acceleration techniques, it is crucial to consider the time complexity of full SVD, specifically $O(mn^2)$ for an $m \times n$ matrix with $n \leq m$, making it computationally demanding for large matrices. In contrast, RSVD primarily accrues costs associated with matrix multiplication, exhibiting a complexity of $O(mnr)$, where r represents the target rank. Additionally, the continuation strategy, as verified in [45], can reduce the number of iterations. Through the ablation analysis, we have made the following observations:

- **Table** 5 presents quantitative results that emphasize the effectiveness of the acceleration modules in reducing timing costs on experimental data. It is important to note that the RSVD is employed to decrease complexity at each iteration for inpainting, classification, and clustering tasks. Remarkably, these changes in the computational approach do not significantly impact the evaluation performance and the total number of iterations. This validation underscores the efficiency of incorporating the RSVD into the proposed nonconvex PBCD algorithm, making it a sensible and effective choice.

Table 5 Comparisons and ablation analysis of evaluation accuracy, timing computation, and iteration numbers for the proposed methods with and without acceleration RSVD strategy on selected three image vision tasks.

//	Inpainting (**Curve Mask**)			Classification (**ExtYaleB_occ**)			Clustering (**COIL20**)			
	PSNR	TIME	ITER	ACC	TIME	ITER	ACC	NMI	TIME	ITER
Setup ($\boldsymbol{A}_{1/2}$)	31.928	53s	(135, 133, 133)	**97.807**	29s	70	69.167	78.175	98s	554
Setup ($\boldsymbol{A}_{2/3}$)	31.828	49s	(122, 121, 120)	96.053	28s	66	69.167	78.175	96s	542
Setup ($\boldsymbol{B}_{1/2}$)	**31.940**	19s	(153, 150, 150)	**97.807**	27s	70	**71.167**	79.189	43s	422
Setup ($\boldsymbol{B}_{2/3}$)	31.869	17s	(135, 135, 136)	95.833	25s	66	70.833	**79.504**	41s	411



(a) Inpainting (**Curve Mask**) (b) Classification (**ExtYaleB_occ**) (c) Clustering (**COIL20**)

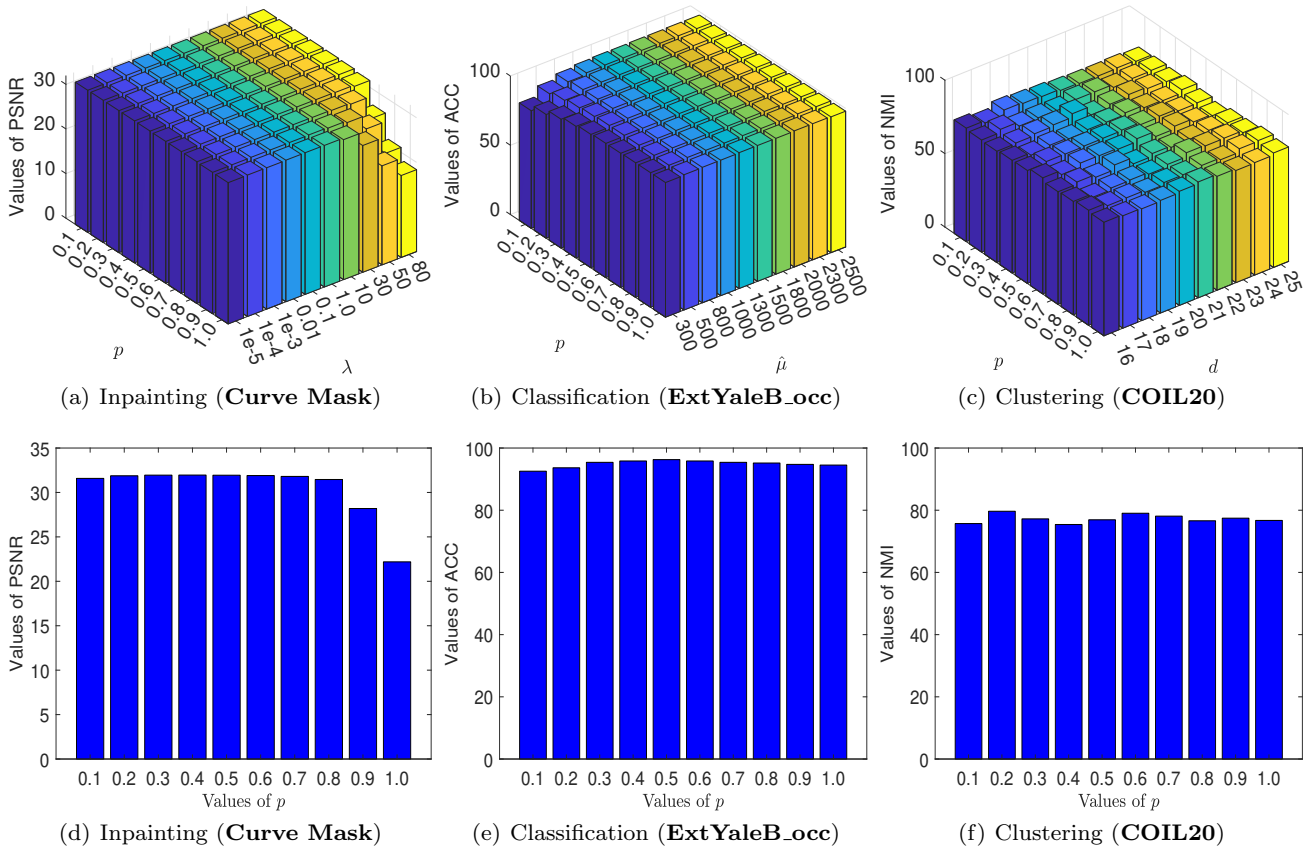
Fig. 13 Plotted curves values of relative errors over CPU time for the two setups on three different application tasks.

– **Fig** 13 provides visual results illustrating how the convergence curves with and without the RSVD influence computational efficiency. By comparing the setups in ($\boldsymbol{A}$) and ($\boldsymbol{B}$), which not only achieve comparable performance but also result in faster convergence speed, the acceleration strategies lead to reduced computational time, as demonstrated in (a)-(c). It should be mentioning that the single channel is considered, and the plotted curve exhibits a floating trend but keeps decreasing in (a). This, in turn, achieves higher computational efficiency and further validates the effectiveness of the RSVD strategy.

Considering that in the preceding experiments, we constrained the parameter p to specific values, i.e., $1/2$ and $2/3$, guided by the availability of closed-form solutions for relevant proximal operators in (group) sparse coding and low-rank matrix recovery problems. Subsequently, we present additional experiments focused on ablation analysis for the hyper-parameters and discuss the impact of parameter p as follows:

– To address concerns about the extensive number of hyper-parameters, we conducted experiments involving ablation analysis, as illustrated in **Fig. 14**. Panels (a)-(c) focus on examining the influence and necessity of specific hyper-parameter pairs, such as (λ, p), $(\hat{\mu}, p)$, and (d, p), on the overall model and algorithmic performance. This comprehensive approach involves systematically isolating and modifying individual hyper-parameters while keeping others constant. By doing so, researchers can gain insights into the relative importance and contributions of each parameter to the model's effectiveness across three various application tasks.
– Acknowledging the importance of discerning the impact of varied p values, we conducted supplementary experiments spanning a spectrum of p values, as illustrated in **Fig. 14**. (d)-(f). Then, we concisely present our findings, underscoring the thorough investigations carried out by altering the parameter p within the range $\{0.1, 0.2, 0.3, \ldots, 0.9, 1.0\}$ across three application tasks. This investigation aims to highlight the varied impacts on PSNR, ACC, and NMI values linked to different p-values, while maintaining fixed values for λ at 50, 0.01, and 0.5 respectively. This demonstration underscores the relative robustness of diverse settings.

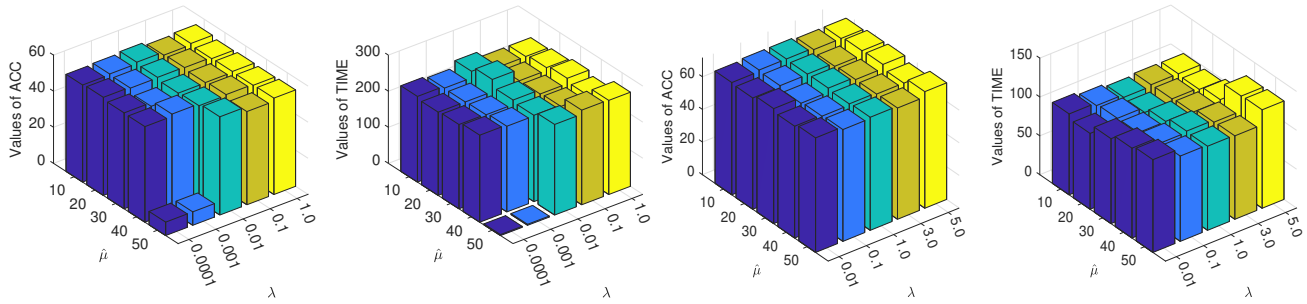
Through a comprehensive analysis and the derived experimental results, our goal is to elucidate the significance and sensitivity of the accelerations and hyper-parameters within the given set. Subsequently, we extend our investigation by exploring how the proposed method performs in large-scale data scenarios. More-



(a) Inpainting (**Curve Mask**) (b) Classification (**ExtYaleB_occ**) (c) Clustering (**COIL20**)

(d) Inpainting (**Curve Mask**) (e) Classification (**ExtYaleB_occ**) (f) Clustering (**COIL20**)

Fig. 14 Comparisons of performance across different parameter choices, including λ, $\hat{\mu}$, d, and p values, within the experimental settings of the three previously mentioned application tasks.



(a) FLAVIA (1200×1907) (b) COIL20 (1024×1440)

Fig. 15 Comparison of clustering accuracy and computational time values across various parameter settings for the two large-scale datasets: (a) FLAVIA with $\text{ReLRR}_{1/2}$ and (b) COIL20 with $\text{ReLRR}_{2/3}$.

over, we systematically present instances of possible failure cases, offering readers opportunities for in-depth analysis. These aspects are introduced and discussed to foster a deeper understanding of the method's performance and limitations among the readers.

– Addressing concerns regarding the proposed method's applicability in large-scale data settings, we conducted experiments to evaluate its robustness for clustering accuracy and computational complexity, as illustrated in **Fig. 15**. (a) and (b). Assessing the performance of the method with different parameter choices and experimental settings on a larger scale is crucial to validate its effectiveness and generalizability in real-world scenarios with extensive datasets. Notably, the dimensions of FLAVIA and COIL20 datasets are 1200×1907 and 1024×1440, respectively. Conducting experiments on complete, large-scale datasets instead of subsets enables researchers

to explore potential scalability issues, evaluate the method's computational efficiency, and collect insights into its adaptability to diverse and more complex datasets. This comprehensive evaluation focuses on enhancing the overall credibility of the research findings, emphasizing the method's robustness in handling large-scale data scenarios.

– In the domain of low-rank matrix recovery, which encompasses tasks like matrix completion, image classification, and subspace clustering, it is crucial to systematically analyze cases where the recovery process fails. Specifically, in non-strictly low-rank matrix learning scenarios, we explore the challenges associated with learning low-rank matrices. Traditional algorithms face difficulties in these situations due to the absence of a strictly low-rank structure, which is particularly evident in image low-level and high-level vision tasks. In addition, these methods are inherently limited in effectively capturing intricate structures within non-strictly low-rank matrices. These instances highlight the pressing need for adaptive approaches capable of accommodating diverse matrix structures. Such adaptability is essential for ensuring robust performance across various real-world applications, especially in situations where strict low-rank assumptions cannot be consistently maintained. Therefore, a thorough examination of the effectiveness of low-rank constraints is essential, ensuring that it considers the presence of low-rank structure in the data and involves appropriate preprocessing steps.

6 Conclusion and Future Work

In this work, our focus is on introducing a unified problem framework designed to effectively tackle common challenges encountered in real-world images, such as noise, variations, and missing data. By capitalizing on the inherent low-rank structure of the data and integrating specific residual measurements, we present robust and efficient solutions to the optimization problem associated with low-rank matrix learning. Our chosen approach employs the nonconvex PBCD algorithm, ensuring both local and global convergence through iterative updates of variable blocks using closed-form solutions. To enhance efficiency, we incorporate the RSVD technique and employ a continuous strategy for adaptive model parameter updates. These strategies serve to reduce computational complexity and offer control over the desired rank of the learned matrix, all while preserving result quality. Additionally, our framework extends the nonconvex PBCD algorithm to handle problems involving multiple variables, supported by a series of thorough analyses. Experimental evaluations, conducted across various image low-level and high-level vision tasks, including inpainting, clustering, and classification, showcase the effectiveness and efficiency of our approach across diverse datum.

In our future research directions, we envision two primary directions for expanding upon the current study. Firstly, our goal is to extend the efficient optimization algorithms developed in this work to address tensor recovery problems, similar to those explored in [72, 84]. This extension holds the potential to advance tensor-based data analysis, paving the way for tackling complex data structures in innovative ways. Secondly, we plan to delve into the integration of additional information into our unified frameworks. This involves incorporating discriminative information or latent variables, as demonstrated in previous studies like [80, 85, 86]. Solely relying on low-rank matrix recovery, without leveraging supplementary information, might lead to suboptimal outcomes, especially when dealing with intricate and diverse image content. Therefore, by harnessing this additional information, our aim is to bolster the performance and robustness of our approach, ultimately enhancing accuracy and reliability across various image low-level and high-level vision tasks.

Acknowledgment

The authors would like to thank the editors and the anonymous reviewers for their critical and constructive comments. This work was supported in part by the Ministry of Education, Republic of Singapore, through its Start-Up Grant and Academic Research Fund Tier 1 under Grant RG61/22; in part by the National Natural Science Fund (NSF) of China under Grant 61906067, Grant 62176124, Grant 61973162, and Grant 12371510; in part by the China Postdoctoral Science Foundation under Grant 2019M651415 and Grant 2020T130191; in part by the Fundamental Research Funds for the Central Universities under Grant 30918014108, Grant 30920032202, and Grant 30921013114; in part by the NSF of Jiangsu Province under Grant BZ2021013; in part by the NSF for Distinguished Young Scholar of Jiangsu Province under Grant BK20220080; and in part by the "111" Program under Grant B13022.

References

1. R. Dian, S. Li, L. Fang, Learning a low tensor-train rank representation for hyperspectral image super-resolution, IEEE Trans. Neural. Netw. Learn. Syst. 30 (9) (2019) 2672–2683.

2. R. Dian, S. Li, Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization, IEEE Trans. Image. Process. 28 (10) (2019) 5135–5146.
3. S. Zhao, J. Wu, L. Fei, B. Zhang, P. Zhao, Double-cohesion learning based multiview and discriminant palmprint recognition, Inf. Fusion 83 (2022) 96–109.
4. H. Zhang, J. Yang, J. Xie, J. Qian, B. Zhang, Weighted sparse coding regularized nonconvex matrix regression for robust face recognition, Inf. Sci. 394-395 (2017) 1–17.
5. R. Dian, S. Li, L. Fang, Q. Wei, Multispectral and hyperspectral image fusion with spatial-spectral sparse representation, Inf. Fusion 49 (2019) 262–270.
6. R. Dian, A. Guo, S. Li, Zero-shot hyperspectral sharpening, IEEE Trans. Pattern. Anal. Mach. Intell. 45 (10) (2023) 12650–12666.
7. B. Wen, S. Ravishankar, Y. Bresler, Structured overcomplete sparsifying transform learning with convergence guarantees and applications, Inter. J. Comput. Vis. 114 (2-3) (2015) 137–167.
8. H. Zhang, J. Yang, J. Qian, G. Gao, X. Lan, Z. Zha, B. Wen, Efficient image classification via structured low-rank matrix factorization regression, IEEE Trans. Inf. Forensics Security 19 (2023) 1496–1509.
9. T. Bouwmans, A. Sobral, S. Javed, S. Jung, E. Zahzah, Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset, Comput. Sci. Rev. 23 (2017) 1–71.
10. H. Yin, S. Li, L. Fang, Simultaneous image fusion and super-resolution using sparse representation, Inf. Fusion 14 (3) (2013) 229–240.
11. G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern. Anal. Mach. Intell. 35 (1) (2012) 171–184.
12. Q. Wang, S. Li, H. Qin, A. Hao, Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis, Inf. Fusion 26 (2015) 103–121.
13. Q. Zhang, Y. Liu, R. S. Blum, J. Han, D. Tao, Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review, Inf. Fusion 40 (2018) 57–75.
14. T. Oh, Y. Matsushita, Y. Tai, I. Kweon, Fast randomized singular value thresholding for low-rank optimization, IEEE Trans. Pattern. Anal. Mach. Intell. 40 (2) (2017) 376–391.
15. H. Zhang, J. Yang, J. Qian, W. Luo, Nonconvex relaxation based matrix regression for face recognition with structural noise and mixed noise, NeuroComputing 269 (2017) 188–198.
16. K. Toh, S. Yun, An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems, Pac. J. Optim. 6 (615-640) (2010) 15.
17. C. Lu, C. Zhu, C. Xu, S. Yan, Z. Lin, Generalized singular value thresholding, in: Proc. Assoc. Adv. Artif. Intell (AAAI), 2015, pp. 1805–1811.
18. Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, arXiv preprint arXiv:1009.5055 (2010).
19. J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern. Anal. Mach. Intell. 31 (2) (2009) 210–227.
20. W. Zuo, D. Meng, L. Zhang, X. Feng, D. Zhang, A generalized iterated shrinkage algorithm for non-convex sparse coding, in: Proc. IEEE Conf. Comput. Vis. Pattern. Recogn (CVPR), 2013, pp. 217–224.
21. H. Zhang, F. Qian, F. Shang, W. Du, J. Qian, J. Yang, Global convergence guarantees of (A)GIST for a family of nonconvex sparse learning problems, IEEE Trans. Cybern. 52 (5) (2022) 3276–3288.
22. J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, Y. Xu, Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes, IEEE Trans. Pattern. Anal. Mach. Intell. 39 (1) (2017) 156–171.
23. J. Qian, W. K. Wong, H. Zhang, J. Xie, J. Yang, Joint optimal transport with convex regularization for robust image classification, IEEE Trans. Cybern. 52 (3) (2020) 1553–1564.
24. M. Abavisani, V. M. Patel, Multimodal sparse and low-rank subspace clustering, Inf. Fusion 39 (2018) 168–177.
25. H. Zhang, J. Yang, F. Shang, C. Gong, Z. Zhang, LRR for subspace segmentation via tractable Schatten-p norm minimization and factorization, IEEE Trans. Cybern. 49 (5) (2019) 1722–1734.
26. E. Elhamifar, R. Vidal, Sparse subspace clustering: Algorithm, theory, and applications, IEEE Trans. Pattern. Anal. Mach. Intell. 35 (11) (2013) 2765–2781.
27. M. Brbić, I. Kopriva, ℓ_0-motivated low-rank sparse subspace clustering, IEEE Trans. Cybern. 50 (4) (2018) 1711–1725.

28. Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.
29. L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition?, in: Proc. IEEE Conf. Comput. Vis. (ICCV), 2012, pp. 471–478.
30. Y. Chen, J. Yang, L. Luo, H. Zhang, J. Qian, Y. Tai, J. Zhang, Adaptive noise dictionary construction via irrpca for face recognition, Pattern Recognition 59 (2016) 26–41.
31. W. Jiang, J. Liu, H. Qi, Q. Dai, Robust subspace segmentation via nonconvex low rank representation, Inf. Sci. 340 (2016) 144–158.
32. 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. Circ. Syst. Vid. 34 (2) (2024) 828–838.
33. J. Li, Z. Li, G. Lu, Y. Xu, B. Zhang, D. Zhang, Asymmetric gaussian process multi-view learning for visual classification, Inf. Fusion 65 (2021) 108–118.
34. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends. Mach. Learn. 3 (1) (2011) 1–122.
35. H. Zhang, F. Qian, P. Shi, W. Du, Y. Tang, J. Qian, C. Gong, J. Yang, Generalized nonconvex nonsmooth low-rank matrix recovery framework with feasible algorithm designs and convergence analysis, IEEE Trans. Neural. Netw. Learn. Syst. 34 (9) (2023) 5342–5353.
36. Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: Proc. Adv. Neural. Inf. Process. Syst (NIPS), 2011, pp. 612–620.
37. C. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2) (2010) 894–942.
38. J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (456) (2001) 1348–1360.
39. F. Nie, H. Wang, H. Huang, C. Ding, Joint Schatten-p norm and ℓ_p-norm robust matrix completion for missing value recovery, Knowl. Infor. Syst. 42 (3) (2015) 525–544.
40. F. Shang, J. Cheng, Y. Liu, Z. Luo, Z. Lin, Bilinear factor matrix norm minimization for robust PCA: algorithms and applications, IEEE Trans. Pattern. Anal. Mach. Intell. 40 (9) (2018) 2066–2080.
41. Y. Chen, A. Jalali, S. Sanghavi, C. Caramanis, Low-rank matrix recovery from errors and erasures, IEEE Trans. Infor. Theo. 59 (7) (2013) 4324–4337.
42. L. Luo, J. Yang, J. Qian, Y. Tai, G. Lu, Robust image regression based on the extended matrix variate power exponential distribution of dependent noise, IEEE Trans. Neural. Netw. Learn. Syst. 28 (9) (2017) 2168–2182.
43. J. Chen, J. Yang, L. Luo, J. Qian, W. Xu, Matrix variate distribution-induced sparse representation for robust image classification, IEEE Trans. Neural. Netw. Learn. Syst. 26 (10) (2015) 2291–2300.
44. H. Zhang, S. Li, J. Qiu, Y. Tang, J. Wen, Z. Zha, B. Wen, Efficient and effective nonconvex low-rank subspace clustering via SVT-free operators, IEEE Trans. Circ. Syst. Vid. 33 (12) (2023) 7515–7529.
45. S. Zhang, H. Qian, X. Gong, An alternating proximal splitting method with global convergence for nonconvex structured sparsity optimization, in: Proc. Assoc. Adv. Artif. Intell (AAAI), 2016, pp. 2330–2336.
46. H. Zhang, J. Zhao, B. Zhang, C. Gong, J. Qian, J. Yang, Unified framework for faster clustering via joint Schatten p-norm factorization with optimal mean, IEEE Trans. Neural. Netw. Learn. Syst. 35 (3) (2024) 3012–3026.
47. H. Zhang, J. Qian, J. Gao, J. Yang, C. Xu, Scalable proximal jacobian iteration method with global convergence analysis for nonconvex unconstrained composite optimizations, IEEE Trans. Neural. Netw. Learn. Syst. 30 (9) (2019) 2825–2839.
48. B. Recht, M. Fazel, P. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM. Rev. 52 (3) (2010) 471–501.
49. V. Larsson, C. Olsson, Convex low rank approximation, Inter. J. Comput. Vis. 120 (2016) 194–214.
50. R. Cabral, F. De la Torre, J. P. Costeira, A. Bernardino, Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition, in: Proc. IEEE Intern. Conf. Comput. Vis. (ICCV), 2013, pp. 2488–2495.
51. T. Oh, Y. Tai, J. C. Bazin, H. Kim, I. S. Kweon, Partial sum minimization of singular values in robust pca: Algorithm and applications, IEEE Trans. Pattern. Anal. Mach. Intell. 38 (4) (2016) 744–758.
52. S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, L. Zhang, Weighted nuclear norm minimization and its applications to low level vision, Inter. J. Comput. Vis. 121 (2) (2017) 183–208.
53. C. Lu, J. Tang, S. Yan, Z. Lin, Nonconvex nonsmooth low-rank minimization via iteratively reweighted nuclear norm, IEEE Trans. Image. Pro-

cess. 25 (2) (2016) 829–839.
54. E. G. Sánchez Manzano, M. A. Gomez Villegas, J. Marín Diazaraque, A matrix variate generalization of the power exponential family of distributions, Commun. Stat. Theor. M. 31 (12) (2002) 2167–2182.
55. T. Sun, D. Li, General nonconvex total variation and low-rank regularizations: Model, algorithm and applications, Pattern. Recogn. 130 (2022) 108692.
56. J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (1-2) (2014) 459–494.
57. 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 (1-2) (2016) 57–79.
58. S. Setzer, Operator splittings, bregman methods and frame shrinkage in image processing, Inter. J. Comput. Vis. 92 (2011) 265–280.
59. H. Zhang, C. Gong, J. Qian, B. Zhang, C. Xu, J. Yang, Efficient recovery of low-rank matrix via double nonconvex nonsmooth rank minimization, IEEE Trans. Neural. Netw. Learn. Syst. 30 (10) (2019) 2916–2925.
60. C. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statis. 38 (2) (2010) 894–942.
61. T. Sun, H. Jiang, L. Cheng, Convergence of proximal iteratively reweighted nuclear norm algorithm for image processing, IEEE Trans. Image. Process. 26 (12) (2017) 5632–5644.
62. Z. Lin, C. Xu, H. Zha, Robust matrix factorization by majorization minimization, IEEE Trans. Pattern. Anal. Mach. Intell. 40 (1) (2018) 208–220.
63. Y. Sun, P. Babu, D. P. Palomar, Majorization-minimization algorithms in signal processing, communications, and machine learning, IEEE Trans. Signal. Process. 65 (3) (2017) 794–816.
64. J. Mairal, Incremental majorization-minimization optimization with application to large-scale machine learning, SIAM J. Optim. 25 (2) (2015) 829–855.
65. Z. Hu, F. Nie, R. Wang, X. Li, Low rank regularization: A review, Neural. Netw. 136 (2021) 218–232.
66. N. Halko, P. Martinsson, J. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, SIAM. Rev. 53 (2) (2011) 217–288.
67. Z. Lin, Some software packages for partial SVD computation, arXiv preprint arXiv:1108.1548 (2011).
68. C. Lu, Z. Lin, S. Yan, Smoothed low-rank and sparse matrix recovery by iteratively reweighted least squares minimization, IEEE Trans. Image. Process. 24 (2) (2015) 646–654.
69. P. Yu, T. K. Pong, Iteratively reweighted ℓ_1 algorithms with extrapolation, Comput. Optim. Appl. 73 (2) (2019) 353–386.
70. H. Zhang, F. Qian, B. Zhang, W. Du, J. Qian, J. Yang, Incorporating linear regression problems into an adaptive framework with feasible optimizations, IEEE Trans. Multi. 25 (2023) 4041–4051.
71. Q. Shen, Y. Liang, S. Yi, J. Zhao, Fast universal low rank representation, IEEE Trans. Circ. Syst. Vid. 32 (3) (2022) 1262–1272.
72. Y. Chen, X. Xiao, C. Peng, G. Lu, Y. Zhou, Low-rank tensor graph learning for multi-view subspace clustering, IEEE Trans. Circ. Syst. Vid. 32 (1) (2022) 92–104.
73. R. T. Rockafellar, R. J. Wets, Variational analysis, Vol. 317, Springer Science & Business Media, 2009.
74. Q. Yao, J. T. Kwok, T. Wang, T. Liu, Large-scale low-rank matrix learning with nonconvex regularizers, IEEE Trans. Pattern. Anal. Mach. Intell. 41 (11) (2019) 2628–2643.
75. Y. He, F. Wang, Y. Li, J. Qin, B. Chen, Robust matrix completion via maximum correntropy criterion and half-quadratic optimization, IEEE Trans. Signal. Proc. 68 (2020) 181–195.
76. Z. Wang, X. Li, H. So, Robust matrix completion based on factorization and truncated-quadratic loss function, IEEE Trans. Circ. Syst. Vid. 33 (4) (2023) 1521–1534.
77. J. Wen, Z. Zhong, Z. Zhang, L. Fei, Z. Lai, R. Chen, Adaptive locality preserving regression, IEEE Trans. Circ. Syst. Vid. 30 (1) (2020) 75–88.
78. R. He, W. Zheng, B. Hu, Maximum correntropy criterion for robust face recognition, IEEE Trans. Pattern. Anal. Mach. Intell. 33 (8) (2010) 1561–1576.
79. J. Wen, Y. Xu, Z. Li, Z. Ma, Y. Xu, Inter-class sparsity based discriminative least square regression, Neural Networks 102 (2018) 36–47.
80. Z. Kang, Z. Lin, X. Zhu, W. Xu, Structured graph learning for scalable subspace clustering: From single view to multiview, IEEE Trans. Cybern. 52 (9) (2022) 8976–8986.
81. C. Lu, H. Min, Z. Zhao, L. Zhu, D. Huang, S. Yan, Robust and efficient subspace segmentation via least squares regression, in: Proc. Europ. Conf. Comput. Vis. (ECCV), 2012, pp. 347–360.
82. J. Xu, M. Yu, L. Shao, W. Zuo, D. Meng, L. Zhang, D. Zhang, Scaled simplex representation for subspace clustering, IEEE Trans. Cybern. 51 (3)

(2019) 1493–1505.
83. Q. Shen, Y. Chen, Y. Liang, S. Yi, W. Liu, Weighted Schatten p-norm minimization with logarithmic constraint for subspace clustering, Signal. Process. 198 (2022) 108568.
84. Y. Xie, D. Tao, W. Zhang, Y. Liu, L. Zhang, Y. Qu, On unifying multi-view self-representations for clustering by tensor multi-rank minimization, Inter. J. Comput. Vis. 126 (2018) 1157–1179.
85. B. Wang, H. Niu, J. Zeng, G. Bai, S. Lin, Y. Wang, Latent representation learning model for multi-band images fusion via low-rank and sparse embedding, IEEE Trans. Multi. 23 (2021) 3137–3152.
86. C. Gong, H. Zhang, J. Yang, D. Tao, Learning with inadequate and incorrect supervision, in: Proc. IEEE Inter. Conf. Data Mining (ICDM), IEEE, 2017, pp. 889–894.